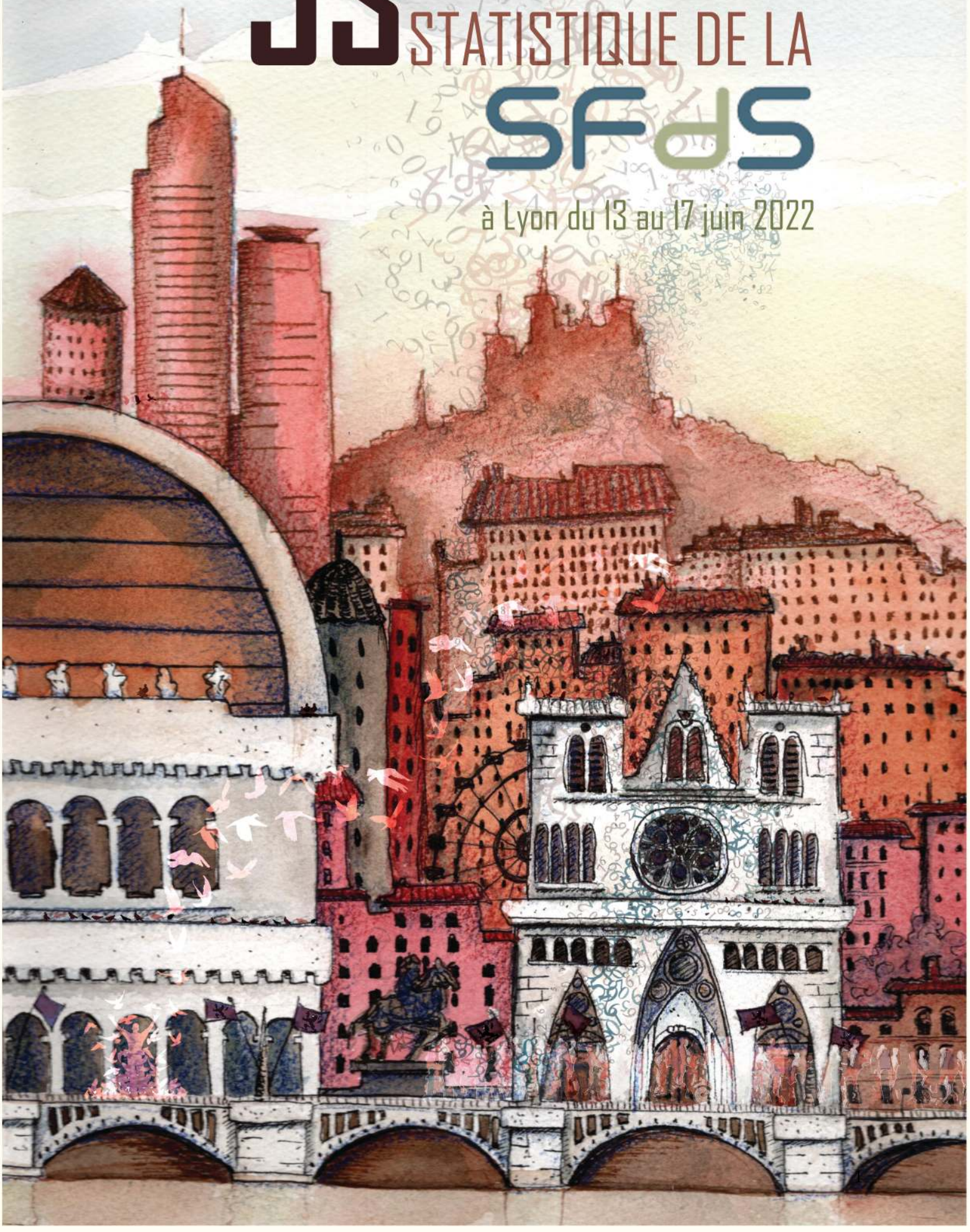
53es JOURNÉES DE STATISTIQUE DE LA SFdS

à Lyon du 13 au 17 juin 2022



53es Journées de statistique de la SFdS

Lyon, 13 au 17 juin 2022

C'est avec un immense plaisir que nous vous accueillons pour les 53èmes journées de statistique, congrès annuel de la société française de statistique (SFdS), qui se dérouleront du 13 au 17 juin 2022 à l'université Claude Bernard Lyon 1. Après plusieurs années de chamboulements majeurs des vies sociales et scientifiques, cet événement revêt une importance particulière, celle du retour à un rassemblement d'une communauté humaine impatiente de pouvoir interagir directement avec des collègues partageant un intérêt commun pour la "chose" statistique. Cette édition 2022 sera rythmée par les exposés scientifiques patiemment sélectionnés par le comité scientifique, mais aussi par des événements sociaux rassemblant notre communauté scientifique (apéritif, assemblée générale, activités sociales et soirée du mercredi). Nous souhaitons donc remercier chaleureusement l'ensemble des comités scientifique et d'organisation, dont le travail a permis l'élaboration de cette semaine de conférences et débats qui nous permettront de "faire société", à nouveau.

Anne-Laure Fougères, Mathilde Mougeot, Franck Picard

Image de couverture réalisée par Laure Savetier

Comité scientifique

Pascal Ardilly	Insee, Lyon
Nicolas Bousquet	EDF
Julie Delon	CNRS Université de Paris
Alain Célisse	Université Paris 1 Panthéon-Sorbonne
Alice Cleynen	CNRS, Montpellier
Nicolas Champagnat	INRIA, U Lorraine
Arthur Charpentier	UQAM
Amandine Marrel	CEA, cadarache
Mathilde Mougeot	ENSIIE, ENS Paris-Saclay, **présidente**
Madalina Olteanu	Université Paris Dauphine
Nelly Pustelnik	CNRS, Lyon
Nicolas Verzelen	INRA, Montpellier

Comité d'organisation

Jean-Baptiste Aubin
Caroline Bayart
Jérémie Becker
Christophette Blanchet
Yohann De Castro
Stéphane Chrétien
Gabriela Ciuperca
Jairo Cugliari
Anne-Béatrice Dufour
Thibault Espinasse
Anne-Laure Fougères (co-présidente)
Irène Gannaz
Aurélien Garivier
Rémi Gribonval
Sonia Guérin-Hamdi
Céline Helbert (co-trésorière)
Laurent Jacob
Julien Jacques, Clélia Lopez
Clément Marteau
Esterina Masiello (co-trésorière)
Cécile Mercadier
Franck Picard (co-président)
Antoine Rolland
Pierre Ribereau
Didier Rullière
Mathieu Sart
Vivian Viallon

Table des matières

13 juin 2022 **12**
PLENIERE : Prix Laplace – Marc Hallin 16
SFdS - Royal Statistical Society 17
Tips for Applied Medical Statisticians from Florence Nightingale - Data Visualisation, Catey Bunce 18
Scientific Reform and Visual Data Science: Retiring the EDA/CDA dichotomy, Roger Beecham 22
Data visualisation of adverse events in randomised clinical trials, Victoria Cornelius 27
Data visualisation of adverse events in randomised clinical trials, Victoria Cornelius 28
Data visualisation of adverse events in randomised clinical trials, Victoria Cornelius 29
Data visualisation of adverse events in randomised clinical trials, Victoria Cornelius 30
Data visualisation of adverse events in randomised clinical trials, Victoria Cornelius 31
Data visualisation of adverse events in randomised clinical trials, Victoria Cornelius 32
Statistique spatiale 38
Fully nonseparable Gneiting covariance functions for multivariate space-time data, Denis Allard 39
Investigating spatial scan statistics for multivariate functional data, Camille Frevent [et al.] 45
ANNULÉ : Spatial k-nearest neighbor regression estimate for random fields, Ahmad Younso 52
Estimation de processus de mélanges spatio-temporels pour détecter des changements dynamiques de populations, Solange Pruilh [et al.] 58
Modélisation spatiale des accidents routiers de Besançon (France) à l'aide de processus de Cox log-Gaussien, Cécile Spychala [et al.] 64
Statistique bayésienne 69
Sélection de variables bayésienne en grande dimension dans les modèles non-linéaires à effets mixtes utilisant l'algorithme SAEM, Marion Naveau [et al.] 70
Mixture of experts posterior surrogates for approximate Bayesian computation, Florence Forbes [et al.] 76
Bayesian nonparametric mixtures inconsistency for the number of clusters, Louise Alamichel [et al.] 82

ANNULÉ : Pedestrian traffic mapping over time at the scale of a city, Mounia Zaouche [et al.] . . . 88
Bayesian functional linear regression estimation. Extension to scalar and categorical covariates, Feriel Bouhadjera [et al.] . . . 94
Applications : société & industrie . . . 100
Modélisation de la Substitution entre Articles pour Optimiser leur Réapprovisionnement, Axel Potier [et al.] . . . 101
Méthode de détection automatique des pannes sur avion : outil d'aide à la décision, Corentin Duprey [et al.] . . . 107
Number of zero velocity points: a critical parameter for handwriting model estimation towards dysgraphia diagnosis assistance, Yunjiao Lu [et al.] . . . 108
Prévision probabiliste d'ensemble en environnement, Eric Parent [et al.] . . . 114
Apprentissage non supervisé . . . 120
Clustering et détection d'anomalies dans les données fonctionnelles, Messan Martial Amovin-Assagba [et al.] . . . 121
Clustering spectral adaptatif, Aurélie Fischer [et al.] . . . 127
Sparse and group-sparse clustering for mixed data. An illustration of the vimpclust package, Alex Mourer [et al.] . . . 132
Cartes auto-organisatrices pour l'exploration des données partiellement observées et l'imputation des données manquantes, Sara Rejeb [et al.] . . . 140
Détection d'atypiques pour données de composition avec la méthode ICS, Christine Thomas-Agnan . . . 146
Extrêmes . . . 152
On the estimation of extreme quantiles with neural networks, Michael Allouche [et al.] . . . 153
Théorème de Donsker pour des mesures empiriques locales sur des régions aléatoires, Benjamin Bobbia [et al.] . . . 159
Non-parametric estimator of a multivariatemadogram for missing-data and extreme valueframework, Alexis Boulin . . . 165
Méthode de plug-in pour l'estimateur non-paramétrique du coefficient de dépendance de queues, Matthieu Garcin . . . 171
Estimation fonctionnelle de L^1-expectiles multivariés extrêmes à queue lourde, Cambyse Pakzad [et al.] . . . 177
Statistique mathématique . . . 183
Full Model Estimation for Non-Parametric Multivariate Finite Mixture Models, Marie Du Roy De Chaumaray [et al.] . . . 184
Modèles linéaires généralisés multivariés à composantes supervisées et facteurs latents, avec partitionnement thématique des variables explicatives, Julien Gibaud [et al.] . . . 190
Calibrations de pénalités pour la régression linéaire gaussienne en grande dimension, Perrine Lacroix . . . 196
Beyond L1: Faster and Better Sparse Models with skglm, Mathurin Massias [et al.] . . . 202
ISDE : Estimation de Densité Multidimensionnelle sous Hypothèse de Structure D'indépendance, Louis Pujol . . . 208
Statistique mathématique 2 . . . 214

Comparaison d'arbres CART par sous-échantillonnage sans remplacement, Felix Cheysson . . . 215
Online multiple-testing with super-uniformity reward, Iqraa Meah [et al.] 221
ANNULÉ : Test de linéarité dans un modèle de régression non paramétrique, Zaher Mohdeb [et al.] . . . 222
Multi-sample comparison of multivariate copulas and clustering, Yves Ismaël Ngounou Bakam [et al.] . . . 227
Consistance de l'intervalle de confiance plug-in en régression, Ahmed Zaoui [et al.] . . . 228
SFdS - Biométrie . . . 234
Modélisation conjointe des relations temporelles entre marqueurs longitudinaux multivariés de progression de la Maladie d'Alzheimer et événements cliniques, Anaïs Rouanet . . . 235
Evaluation par la modélisation des stratégies de rappel vaccinal contre la Covid-19 dans une population partiellement vaccinée, Clement Massonaud [et al.] . . . 239
Quantifying horizontal pleiotropy in human genetic variation using GWAS summary statistics, Marie Verbanck . . . 245
SFdS - MALIA - SSFAM . . . 246
Analyse statistique de la topologie de Mapper pour des filtres stochastiques, Mathieu Carrière [et al.] . . . 247
Graph Neural Networks on Large Random Graphs, Samuel Vaiter . . . 253
Apprentissage pour l'amélioration de surfaces définies par des nuages de points, Julie Digne . . . 254
PLENIERE : Aurélien Bellet . . . 255
Graphes – Réseaux . . . 256
Détection des structures communes dans une collection de réseaux, Pierre Barbillon . . . 257
Clustering networks with textual edges by combining the Embedded Topic Model and the Stochastic Block model. Variational inference and derivation of a model selection criterion., Rémi Boutin [et al.] . . . 263
Modèles génératifs de graphes : application à la connectivité cérébrale, Clément Mantoux [et al.] . . . 269
Analyse statistique de graphes, via des processus de diffusion de la chaleur., Etienne Lasalle . . . 275
Convergence en loi des U-statistiques sur une matrice exchangeable ligne-colonne, Tâm Le Minh . . . 281
Séries temporelles . . . 287
Modèle de régression et prévision de séries temporelles pour la demande en électricité à Djibouti, Mohamed Abdillahi Isman [et al.] . . . 288
Estimation Adaptative des Variances dans un Modèle Espace-État, Joseph De Vilmarest [et al.] . . . 295
ANNULÉ : On multi-change-points tests based on multi-samples U-Statistics for weakly dependent observations, Echarif El Harfaoui [et al.] . . . 300
On estimating in generalized integer-valued GARCHX models with structural breaks, Mohamed Djemaa Sadoun [et al.] . . . 306

Tight Risk Bound For High Dimensional Time Series Completion, Amélie Rosier [et al.] . . . 312
Statistique mathématique – grande dimension . . . 318
Sparse PLS with group lasso: inside the dual.spls package, Louna Alsouki [et al.] . . . 319
Sélection de modèles linéaires emboîtés par règle d'arrêt prématuré, Samy Clementz [et al.] . . . 325
Régression non paramétrique par projection sur des bases à support non compact, Florian Dussap . . . 331
Phénomène de Stein pour l'estimation multiple de moyennes en grande dimension, Jean-Baptiste Fermanian . . . 337
Estimation of high dimensional gamma convolutions through random projections, Oskar Laverny . . . 343
Applications : biologie et santé . . . 348
Méthodes pour l'inférence post-clustering appliquées à l'expression génique, Benjamin Hivert [et al.] . . . 349
Prédiction génomique en grande dimension : densité de marqueurs nécessaire pour une prédiction fiable en sélection génomique, Charles-Elie Rabier [et al.] . . . 355
Identification of prognostic and predictive biomarkers in high-dimensional data, Wencan Zhu [et al.] . . . 361
Impact d'une copule non-Gaussienne dans l'estimation REML du modèle génétique animal bivarié pour des populations sous sélection, Tom Rohmer [et al.] . . . 367
Évaluation de l'impact de scénarios de consommation de tabac sur la charge future de l'infarctus du myocarde en France jusqu'en 2035 : une approche basée sur un modèle illness-death, Johann Kuhn [et al.] . . . 373
Planification – Plans d'expérience . . . 382
Simulation d'événements rares par échantillonnage préférentiel adaptatif pour des processus de Markov déterministes par morceaux, Guillaume Chennetier [et al.] . . . 383
Estimation de courbes de fragilité sismique par planification séquentielle d'expériences., Clément Gauchy [et al.] . . . 389
Design of Experiment for Bayesian transferred model, Loïc Iapteff [et al.] 395
Données expérimentales à bas coût pour la prévision de la distance parcourue par une avalanche de pierres, Frederique Leblanc . . . 401
Plans d'expériences pour mélanges avec une distribution de Dirichlet, Astrid Jourdan . . . 407
Prix ENSAI - Enora Alaoui, Pierre Barbe, Félix Lucas et Victoria Mas . . . 413

14 juin 2022 **414**
PLENIERE : Francoise Berthoud 417
SFdS - Environnement 418
Comptage de macrodéchets en zone riparienne par apprentissage profond et modèle d'états, Mathis Chagneux 419
Bivariate Log-Gaussian Cox Process Model for Presence-only Data, Florian Lasgorceux [et al.] 428
When strong annotators lead to bad model choices, Maximilien Servajean [et al.] 434
SFdS - Enseignement 440
Analyse des tendances de la qualité de l'air, Manon Andre 441
ANNULÉ : Hybridation Numérique : la réussite des étudiants à EPITECH, Nefeli Paparisteidi [et al.] 447
Scientific Reform and Visual Data Science: Retiring the EDA/CDA dichotomy, Roger Beecham 453
Comprendre l'échec des étudiants en vue d'un enseignement différencié, Sylvie Viguier-Pla [et al.] 458
Transposition didactique et enseignement de la statistique, Antoine Rolland 464
Application : santé & société 469
Application des réseaux de neurones aux données longitudinales de santé: cas du suivi de la fonction rénale aux patients transplantés rénaux, Clément Benoist [et al.] 470
Data visualisation of adverse events in randomised clinical trials, Victoria Cornelius 476
Impact de la variabilité de la pression artérielle sur le risque d'AVC, Léonie Courcoul [et al.] 482
Utilisation des matrices emplois expositions pour les troubles musculosquelettiques. Application aux douleurs sévères du genou, Guillaume Deltreil [et al.] 488
Optimal Individualized Treatment Regime of Plasma Transfusion for Severe Trauma, Pan Zhao [et al.] 494
Données Fonctionnelles 499
Estimation adaptative dans le modèle linéaire fonctionnel à sortie fonctionnelle, Gaëlle Chagny [et al.] 500
Procédure de classification plug-in pour des données fonctionnelles, Eddy Ella Mintsa 506
Fusion tardive en analyse de données fonctionnelles élastique, Julien Ah-Pine [et al.] 507
Prédiction de la Qualité d'Expérience dans les Réseaux Mobiles : Cas de la VoIP, Jean Steve Tamo Tchomgui [et al.] 513
Sélection d'intervalles pour des prédicteurs fonctionnels à partir de forêts aléatoires, Rémi Servien [et al.] 519
Statistique mathématique 525
Multivariate Poisson Lognormal model: optimisation, inference and application to high dimensional data, Bastien Batardière [et al.] 526
Les approximations à noyau à base de processus ponctuels déterminantaux, Ayoub Belhadji 532

Un multi-critère pour contrôler la multicolinéarité dans les modèles linéaires de régression multiple, Christian Derquenne . . . 538
High-dimensional Two-sample Rank Statistics as Scalar Performance Criteria, Myrto Limnios [et al.] . . . 544
Clustering de modalités en régression logistique, Slimane Makhlouf [et al.] 550
PLENIERE : Clémentine Prieur . . . 551
PLENIERE ANNULÉE : Chloé-Agathe Azencott . . . 552
PLENIERE ANNULEE : Nicolas Papadakis . . . 553
SFdS - Fiabilité . . . 554
Degradation Model Selection Using Depth Functions, Mitra Fouladirad [et al.] . . . 555
On misspecification of a Gamma with an inverse Gaussian process and its effects in terms of prognostic and related decision processes, Nicola Esposito [et al.] . . . 557
Sequential Interval Reliablity for semi-Markov systems, Vlad Stefan Barbu [et al.] . . . 559
Building wind loading estimation under uncertain inflow parameters by means of lattice Boltzmann simulation., Jérôme Jacob [et al.] . . . 565
Méta-modélisation multi-fidélité par processus Gaussien en utilisant une base d'ondelettes, Baptiste Kerleguer [et al.] . . . 566
SFdS - Jeunes Statisticiens . . . 569
Agrégation experts . . . 570
Modélisation de la charge et de l'occupation des véhicules électriques, Yvenn Amara-Ouali [et al.] . . . 571
Functional mixture-of-experts for classification, Nhat Thien Pham [et al.] 577
Ensemble Methods for Data Analytics, Camila Fernandez [et al.] . . . 583
Consensual aggregation on randomly projected high-dimensional features of predictions for regression, Sothea Has . . . 588
Prédictions Conformelles Adaptatives pour les Séries Temporelles, Margaux Zaffran [et al.] . . . 594
Model selection by penalization in mixture of experts models with a non-asymptotic approach, Trungtin Nguyen [et al.] . . . 600
Apprentissage non supervisé . . . 606
Clustering Longitudinal Ordinal Data via Finite Mixture of Matrix-Variate Latent Gaussians, Francesco Amato [et al.] . . . 607
Process Mining : une nouvelle approche de l'évaluation des clusters à l'aide d'un Quantitative Model Checking, Emmanuelle Claeys [et al.] . . . 613
Model-Based Clustering by Greedy Maximization of the Integrated Classification Likelihood, Nicolas Jouvin [et al.] . . . 619
A Dynamic Latent Block Model for Co-clustering of Zero-Inflated Count Data Streams, Giulia Marchello [et al.] . . . 625
False clustering rate control in mixture models, Ariane Marandon [et al.] . 632
Transport optimal . . . 639
Transport optimal pour l'appariement automatique de données métabolomiques non-ciblées, Marie Breeur [et al.] . . . 640
GAN Estimation of Lipschitz Optimal Transport, Lucas De Lara [et al.] . 646

JDCOT : an Algorithm for Transfer Learning in Incomparable Domains using Optimal Transport, Marion Jeamart [et al.] . . . 652
La Distance de Sliced-Wasserstein pour l'Apprentissage à Grande Échelle, Kimia Nadjahi . . . 658

15 juin 2022 **664**
SFdS - Società Italiana di Statistica . . . 667

Analysing spatio-temporal patterns in the association of confirmed deaths by COVID-19 and intensive care admissions, Andrea Bucci [et al.] . . . 668
Sieve bootstrap based on extreme learning machines in time series analysis, Cira Perna [et al.] . . . 674
Advances in clustering asymmetric proximity data, Donatella Vicari . . . 676
SFdS - ENBIS . . . 678

Scalable additive stacking models for electricity demand forecasting, Euan Enticott . . . 679
Metamodeling and sensitivity analysis for models with spatial output. Application to coastal flooding models., Tran Vi-Vi Élodie Perrin [et al.] . . . 680
Textual Data for Electricity Demand Forecasting, David Obst [et al.] . . . 686
On forecasting bounded variables with latent stochastic bounds, Amandine Pierrot [et al.] . . . 687
SFdS - Sport . . . 693

L'origine de la déficience visuelle est-elle un facteur de performance ? Analyse des para-nageurs et para-athlètes de niveau international, Julien Schipman [et al.] . . . 694
Relative age effect among French swimmers, Audrey Difernand . . . 701
Parametric and non-parametric modeling by a Makovian multistate model of the evolution of performance in French alpine skiing, Quentin De Larochelambert [et al.] . . . 707
Reparlons de notes en patinage, Léo Gerville-Réache . . . 713
Modélisation de la probabilité de réussite d'une passe au football, Sébastien Déjean [et al.] . . . 718
Estimation of race scenarios in standard triathlon according to level density, Thibaut Ledanois . . . 724
Statistique mathématique . . . 731

Estimateurs non paramétriques à noyau général birnbaum-saunders des densités multivariées à support non négatifs, Smail Adjabi . . . 732
Integration of bounded monotone functions: Revisiting the nonsequential case, with a focus on unbiased Monte Carlo (randomized) methods, Subhasish Basak [et al.] . . . 738
Subsampling under constraint, Florian Combes [et al.] . . . 744
Recursive conditional distribution estimation using kernel method and Bernstein polynomials, Salima Helali [et al.] . . . 750
Explainability as statistical inference, Hugo Senetaire [et al.] . . . 751
La Géométrie pour l'Unicité, la Parcimonie et le Regroupement des Estimateurs Pénalisés, Patrick Tardivel [et al.] . . . 759
PLENIERE : Li-Chun Zhang . . . 765

PLENIERE : Emilie Kaufmann . . . 766

16 juin 2022 **767**

PLENIERE : Samory Kpotufe . . . 771

PLENIERE : Pierre Chainais . . . 772

SFdS - Statistique mathématique . . . 773

Profondeur de Tukey: Ensembles de niveau empiriques et théoriques, Victor-Emmanuel Brunel . . . 774

Statistical depth functions: Characterization and reconstruction problems, Stanislav Nagy . . . 780

SFdS - BFA . . . 786

Multi-normex for evaluating the distribution of aggregated heavy tailed risks, Marie Kratz [et al.] . . . 787

Lundberg-Aumann-Serrano dangerousness risk index, Etienne Marceau . 792

ANNULÉ : Three-Step Risk Inference In Insurance Ratemaking, Liang Peng 794

Optimal pooling and distributed inference for the tail index and extreme quantiles, Gilles Stupfler [et al.] . . . 798

Clustering parcimonieux pour extrêmes multivariés, Olivier Wintenberger [et al.] . . . 804

Applications . . . 810

Votes par évaluation avec des fonctions de profondeur, Jean-Baptiste Aubin [et al.] . . . 811

Impact of the energy transition on long-term factor productivity, Guillaume Flament . . . 818

Pl@ntNet-300K: une base de données d'images de plantes avec une ambiguïté entre classes importante et une distribution à longue traîne, Camille Garcin [et al.] . . . 824

Variance estimators for weighted and stratified linear dose-response function estimators using generalized propensity score, Valerie Gares [et al.] . 830

Topological period counting method for reparametrized functions, Wojciech Reise . . . 836

Extrêmes & Prediction . . . 842

Extremal cluster inference for heavy-tailed time series, Gloria Buritica [et al.] . . . 843

Étude de la dépendance spatiale dans les extrêmes : un point de vue géostatistique, Marine Demangeot [et al.] . . . 849

Maximum Mean Discrepancy invariant par translation et convergence en loi, Thibault Modeste [et al.] . . . 855

Vitesse de Convergence Minimax pour la Régression Distributionnelle basée sur le Continuous Ranked Probability Score, Romain Pic [et al.] . . . 860

Non-Asymptotic Analysis of Stochastic Approximation Algorithms for Streaming Data, Nicklas Werge . . . 866

Survie . . . 872

Machine learning for survival data prediction: Application of the super learner on pseudo-observations, Ariane Cwiling [et al.] . . . 873

Comparaisons de méthodes pour données de survie en grande dimension sur de petits échantillons : optimisation des hyperparamètres et validation., Audrey Lavenu [et al.] . . . 879

Modèle de Cox avec des données hétérogènes, Eliz Peyraud . . . 885

Cox regression with linked data, Thanh Huan Vo [et al.] 891
SFdS - AMIES 897
Lights: a generalized joint model for high-dimensional multivariate longitudinal data and censored durations, Van Tuan Nguyen 898
L'agence pour les mathématiques en interaction avec l'entreprise et la société (AMIES), Véronique Maume-Deschamps 899
Prédiction des niveaux de risque pollinique à partir de données historiques multi-sources, Esso-Ridah Bleza [et al.] 900
Deep Calibration of Interest Rates Models, Djibril Sarr 906
ML - Graphes - Réseaux 912
The deep latent position model for node partitioning in graphs, Dingge Liang [et al.] 913
Etude de la Pénalisation GraphNet en Analyse de Données Multi-blocs, Annabelle Beaudoin [et al.] 921
Don't fear the unlabelled: safe deep semi-supervised learning via simple debiasing, Hugo Schmutz [et al.] 927
Generalised Mutual Information Maximisation for Deep Clustering, Louis Ohl [et al.] 933
Méta-modèles – Analyse de sensibilité 939
Heterogeneous Treatment Effects Estimation: When Machine Learning meets multiple treatments regime, Naoufal Acharki [et al.] 940
Risk bounds for Sobol'-based Global Sensitivity Analysis Indices when using metamodels., Clément Bénesse 941
Test d'indépendance basé sur les indices HSIC-ANOVA d'ordre total, Gabriel Sarazin [et al.] 944
Sélection de variables lors de la généralisation d'un effet causal, Bénédicte Colnet [et al.] 950
Calage conditionnel bayésien d'un modèle numérique, Oumar Baldé [et al.] 956
ML & Regret 962
Multitask Online Mirror Descent, Pierre Laforgue [et al.] 963
Identification de Meilleur Bras à Budget Fixé, Antoine Barrier [et al.] . . 969
On Best Policy Identification with Instance-specific sample complexity, Aymen Al Marjani 975
Bayesian Actor for Deep Reinforcement Learning, Léo Grill 980
ANNULÉ : Apprentissage robuste pour la résolution d'EDPs, Applications aux marchés de l'énergie, Carl Remlinger 986
Statistique mathématique 992
Estimation confidentielle de quantiles en présence d'atomes, Clément Lalanne [et al.] 993
Généralisation et estimation de la Q-distribution gaussienne, Oumaima Ben Mrad [et al.] 999
Projection de mesures de probabilité sous contraintes de quantile par distance de Wasserstein et approximation monotone polynomiale, Marouane Il Idrissi [et al.] 1005
Sparse regression to anticipate the declarations of natural disaster in France, Thi Thanh Yen Nguyen [et al.] 1011

Estimation adaptative optimale de fonctions moyennes et covariances irrégulières, Steven Golovkine [et al.] . . . 1017
Réduction d'une fonction aléatoire cyclostationnaire définie sur R, Sylvie Viguier-Pla [et al.] . . . 1023
PLENIERE : Frédéric Chazal . . . 1029
PLENIERE : Gersende Fort . . . 1030
Séries Temporelles . . . 1031
Une nouvelle méthode pour les propriétés asymptotiques de modèles à seuil autoexcités, Guy Mélard [et al.] . . . 1032
Barycentres de séries temporelles : une nouvelle approche basée sur la méthode de la signature, Raphael Mignot [et al.] . . . 1038
Réduction de dimension pour les séries temporelles avec des auto-encodeurs variationnels, William Todo [et al.] . . . 1044
ML et Extrêmes . . . 1050
In-store traffic flow forecasting using high dimensional time series methods, Siham Alaoui Belghiti [et al.] . . . 1051
Estimation de paramètres pour un modèle de propagation de la fièvre typhoïde à Mayotte, Ibrahim Bouzalmat [et al.] . . . 1057
Crowdsourcing label noise simulation on image classification tasks, Tanguy Lefort [et al.] . . . 1063
Evaluating the Impact of Parenthood on Career Progression in STEM jobs, across Gender, Mira Rahal [et al.] . . . 1069
Profondeur à noyau multivariée et mesure de risque extrême associée, Sara Armaut . . . 1075
Kalman Recursions Aggregated Online, Eric Adjakossa . . . 1081
Méthodes Génératives - ML - Deep . . . 1088
Etude du comportement de l'algorithme DbAS pour l'optimisation de propriétés de séquences nucléiques, Teddy Ardouin [et al.] . . . 1089
Limite d'échelle pour les réseaux de neurones résiduels, Pierre Marion . . . 1095
Can Push-forward Generative Models Fit Multimodal Distributions?, Antoine Salmona [et al.] . . . 1101
Generating new crystal structures with statistical methods, Arsen Sultanov [et al.] . . . 1107
Comment imposer des pré-activations gaussiennes dans un réseau de neurones ?, Pierre Wolinski [et al.] . . . 1111
Processus - Statistique mathématique . . . 1117
Vitesses de contraction du posterior pour les processus gaussiens profonds contraints en classification et estimation de densité, Francois Bachoc . . . 1118
Sélection de variables en régression SIR par seuillage doux ou seuillage dur de la matrice d'intérêt, Hadrien Lorenzo [et al.] . . . 1124
Liens entre les décompositions de Hoeffding–Sobol et de Möbius, Cécile Mercadier . . . 1129
Mélange de modèles d'analyse factorielle longitudinale pour l'analyse de données longitudinales multivariées, Amine Ounajim [et al.] . . . 1133
Estimation adaptative dans un modèle de régression-convolution en base d'Hermite, Ousmane Sacko . . . 1139

A local version of R-hat for MCMC convergence diagnostic, Théo Moins [et al.] . . . 1145
Statistique mathématique – Valeurs Manquantes . . . 1151
Vitesse de convergence pour une prediction linéaire en présence de données manquantes, Alexis Ayme [et al.] . . . 1152
Imputation multiple dans le modèle linéaire fonctionnel avec une covariable partiellement observée et des valeurs manquantes dans la réponse, Christophe Crambes [et al.] . . . 1158
Usefulness of surrogate outcomes on the regression model and the PaO2 prediction, Firas Ibrahim [et al.] . . . 1164
Apprentissage semi-supervisé avec réponses manquantes informatives, Aude Sportisse [et al.] . . . 1169
Lights: a generalized joint model for high-dimensional multivariate longitudinal data and censored durations, Herbert Susmann . . . 1174
Rencontre Jeunes statisticiens . . . 1181
Café de la Statistique . . . 1182

Liste des auteurs **1184**

Sponsors **1190**

13 juin 2022

Programme

PLENIERE : Prix Laplace – Marc Hallin **16**

SFdS - Royal Statistical Society **17**

Tips for Applied Medical Statisticians from Florence Nightingale - Data Visualisation, Catey Bunce 18

Scientific Reform and Visual Data Science: Retiring the EDA/CDA dichotomy, Roger Beecham 22

Data visualisation of adverse events in randomised clinical trials, Victoria Cornelius 27

Data visualisation of adverse events in randomised clinical trials, Victoria Cornelius 28

Data visualisation of adverse events in randomised clinical trials, Victoria Cornelius 29

Data visualisation of adverse events in randomised clinical trials, Victoria Cornelius 30

Data visualisation of adverse events in randomised clinical trials, Victoria Cornelius 31

Data visualisation of adverse events in randomised clinical trials, Victoria Cornelius 32

Statistique spatiale **38**

Fully nonseparable Gneiting covariance functions for multivariate space-time data, Denis Allard 39

Investigating spatial scan statistics for multivariate functional data, Camille Frevent [et al.] 45

ANNULÉ : Spatial k-nearest neighbor regression estimate for random fields, Ahmad Younso 52

Estimation de processus de mélanges spatio-temporels pour détecter des changements dynamiques de populations, Solange Pruilh [et al.] 58

Modélisation spatiale des accidents routiers de Besançon (France) à l'aide de processus de Cox log-Gaussien, Cécile Spychala [et al.] 64

Statistique bayésienne **69**

Sélection de variables bayésienne en grande dimension dans les modèles non-linéaires à effets mixtes utilisant l'algorithme SAEM, Marion Naveau [et al.] . . . 70

Mixture of experts posterior surrogates for approximate Bayesian computation, Florence Forbes [et al.] 76

Bayesian nonparametric mixtures inconsistency for the number of clusters, Louise Alamichel [et al.] 82

ANNULÉ : Pedestrian traffic mapping over time at the scale of a city, Mounia Zaouche [et al.] 88

Bayesian functional linear regression estimation. Extension to scalar and categorical covariates, Feriel Bouhadjera [et al.] 94

Applications : société & industrie **100**

Modélisation de la Substitution entre Articles pour Optimiser leur Réapprovisionnement, Axel Potier [et al.] 101

Méthode de détection automatique des pannes sur avion : outil d'aide à la décision, Corentin Duprey [et al.] 107

Number of zero velocity points: a critical parameter for handwriting model estimation towards dysgraphia diagnosis assistance, Yunjiao Lu [et al.] 108

Prévision probabiliste d'ensemble en environnement, Eric Parent [et al.] 114

Apprentissage non supervisé **120**

Clustering et détection d'anomalies dans les données fonctionnelles, Messan Martial Amovin-Assagba [et al.] 121

Clustering spectral adaptatif, Aurélie Fischer [et al.] 127

Sparse and group-sparse clustering for mixed data. An illustration of the vimpclust package, Alex Mourer [et al.] 132

Cartes auto-organisatrices pour l'exploration des données partiellement observées et l'imputation des données manquantes, Sara Rejeb [et al.] 140

Détection d'atypiques pour données de composition avec la méthode ICS, Christine Thomas-Agnan 146

Extrêmes **152**

On the estimation of extreme quantiles with neural networks, Michael Allouche [et al.] 153

Théorème de Donsker pour des mesures empiriques locales sur des régions aléatoires, Benjamin Bobbia [et al.] 159

Non-parametric estimator of a multivariatemadogram for missing-data and extreme valueframework, Alexis Boulin 165

Méthode de plug-in pour l'estimateur non-paramétrique du coefficient de dépendance de queues, Matthieu Garcin 171

Estimation fonctionnelle de L^1-expectiles multivariés extrêmes à queue lourde, Cambyse Pakzad [et al.] 177

Statistique mathématique **183**

Full Model Estimation for Non-Parametric Multivariate Finite Mixture Models, Marie Du Roy De Chaumaray [et al.] 184

Modèles linéaires généralisés multivariés à composantes supervisées et facteurs latents, avec partitionnement thématique des variables explicatives, Julien Gibaud [et al.] 190

Calibrations de pénalités pour la régression linéaire gaussienne en grande dimension, Perrine Lacroix 196

Beyond L1: Faster and Better Sparse Models with skglm, Mathurin Massias [et al.] 202

ISDE : Estimation de Densité Multidimensionnelle sous Hypothèse de Structure D'indépendance, Louis Pujol 208

Statistique mathématique 2 . . . . . **214**

Comparaison d'arbres CART par sous-échantillonnage sans remplacement, Felix Cheysson 215

Online multiple-testing with super-uniformity reward, Iqraa Meah [et al.] 221

ANNULÉ : Test de linéarité dans un modèle de régression non paramétrique, Zaher Mohdeb [et al.] 222

Multi-sample comparison of multivariate copulas and clustering, Yves Ismaël Ngounou Bakam [et al.] 227

Consistance de l'intervalle de confiance plug-in en régression, Ahmed Zaoui [et al.] 228
SFdS - Biométrie **234**
Modélisation conjointe des relations temporelles entre marqueurs longitudinaux multivariés de progression de la Maladie d'Alzheimer et événements cliniques, Anaïs Rouanet 235
Evaluation par la modélisation des stratégies de rappel vaccinal contre la Covid-19 dans une population partiellement vaccinée, Clement Massonaud [et al.] 239
Quantifying horizontal pleiotropy in human genetic variation using GWAS summary statistics, Marie Verbanck 245
SFdS - MALIA - SSFAM **246**
Analyse statistique de la topologie de Mapper pour des filtres stochastiques, Mathieu Carrière [et al.] 247
Graph Neural Networks on Large Random Graphs, Samuel Vaiter 253
Apprentissage pour l'amélioration de surfaces définies par des nuages de points, Julie Digne 254
PLENIERE : Aurélien Bellet **255**
Graphes – Réseaux **256**
Détection des structures communes dans une collection de réseaux, Pierre Barbillon257
Clustering networks with textual edges by combining the Embedded Topic Model and the Stochastic Block model. Variational inference and derivation of a model selection criterion., Rémi Boutin [et al.] 263
Modèles génératifs de graphes : application à la connectivité cérébrale, Clément Mantoux [et al.] 269
Analyse statistique de graphes, via des processus de diffusion de la chaleur., Etienne Lasalle 275
Convergence en loi des U-statistiques sur une matrice exchangeable ligne-colonne, Tâm Le Minh 281
Séries temporelles **287**
Modèle de régression et prévision de séries temporelles pour la demande en électricité à Djibouti, Mohamed Abdillahi Isman [et al.] 288
Estimation Adaptative des Variances dans un Modèle Espace-État, Joseph De Vilmarest [et al.] 295
ANNULÉ : On multi-change-points tests based on multi-samples U-Statistics for weakly dependent observations, Echarif El Harfaoui [et al.] 300
On estimating in generalized integer-valued GARCHX models with structural breaks, Mohamed Djemaa Sadoun [et al.] 306
Tight Risk Bound For High Dimensional Time Series Completion, Amélie Rosier [et al.] 312
Statistique mathématique – grande dimension **318**
Sparse PLS with group lasso: inside the dual.spls package, Louna Alsouki [et al.] 319
Sélection de modèles linéaires emboîtés par règle d'arrêt prématuré, Samy Clementz [et al.] 325
Régression non paramétrique par projection sur des bases à support non compact, Florian Dussap 331
Phénomène de Stein pour l'estimation multiple de moyennes en grande dimension, Jean-Baptiste Fermanian 337

Estimation of high dimensional gamma convolutions through random projections, Oskar Laverny . . . 343
Applications : biologie et santé . . . **348**
Méthodes pour l'inférence post-clustering appliquées à l'expression génique, Benjamin Hivert [et al.] . . . 349
Prédiction génomique en grande dimension : densité de marqueurs nécessaire pour une prédiction fiable en sélection génomique, Charles-Elie Rabier [et al.] . . . 355
Identification of prognostic and predictive biomarkers in high-dimensional data, Wencan Zhu [et al.] . . . 361
Impact d'une copule non-Gaussienne dans l'estimation REML du modèle génétique animal bivarié pour des populations sous sélection, Tom Rohmer [et al.] . . . 367
Évaluation de l'impact de scénarios de consommation de tabac sur la charge future de l'infarctus du myocarde en France jusqu'en 2035 : une approche basée sur un modèle illness-death, Johann Kuhn [et al.] . . . 373
Planification – Plans d'expérience . . . **382**
Simulation d'événements rares par échantillonnage préférentiel adaptatif pour des processus de Markov déterministes par morceaux, Guillaume Chennetier [et al.] . . . 383
Estimation de courbes de fragilité sismique par planification séquentielle d'expériences., Clément Gauchy [et al.] . . . 389
Design of Experiment for Bayesian transferred model, Loïc Iapteff [et al.] . . . 395
Données expérimentales à bas coût pour la prévision de la distance parcourue par une avalanche de pierres, Frederique Leblanc . . . 401
Plans d'expériences pour mélanges avec une distribution de Dirichlet, Astrid Jourdan 407
Prix ENSAI - Enora Alaoui, Pierre Barbe, Félix Lucas et Victoria Mas . . . **413**

PLENIERE : Prix Laplace – Marc Hallin

Quantiles, Profondeur et Transports de Mesures

Marc Hallin

Univ. libre de Bruxelles

Le concept univarié de fonction quantile—l'inverse d'une fonction de répartition—joue un rôle fondamental en Statistique et en Analyse des Données aussi bien qu'en Calcul des Probabilités. En dimension supérieure à un, malheureusement, l'inversion de la fonction de répartition traditionnelle ne mène à aucun des résultats qui font des versions empiriques de ces concepts (quantiles empiriques, rangs) des outils statistiques de première importance. La raison fondamentale en est que, contrairement à la droite, l'espace réel en dimension d>1 n'est pas ordonné de façon canonique. Les concepts de profondeur (deont l'exemple le plus connu est la profondeur au sens de Tukey) ont été introduits pour pallier ce problème. Les concepts de profondeur, malheureusement, ne mènent pas à des notions jouissant des propriétés attendues d'une notion de quantile. En particulier, la probabilité d'une région délimitée par un contour de profondeur donnée n'est pas indépendante de la loi sous-jacente, ce qui contredit à l'essence même d'une région quantile. Fondée sur des idées de transports de mesures, un concept de fonction de répartition multivariée nouveau, dit "center-outward," donne lieu à une notion de contours quantiles multivariés présentant toutes les propriétés de la notion univariée. Sa version empirique, de même, conduit à des rangs et des signes multivariés et des tests qui étendent au contexte multivarié la théorie des tests de rangs univariés et des R-estimateurs associés au nom de Hajek. Les contours quantiles correspondants peuvent être interprétés comme une version "transformation-retransformation" des contours de profondeur traditionnels.

SFdS - Royal Statistical Society

Tips for Applied Medical Statisticians from Florence Nightingale – Data Visualisation

Catey Bunce [1,2,3]

1. Royal Marsden NHS Foundation Trust
2. Consultant in Applied Medical Statistics, Moorfields Eye Hospital
3. Co-lead of the NIHR Statistics Group

Email: Catey.bunce@rmh.nhs.uk

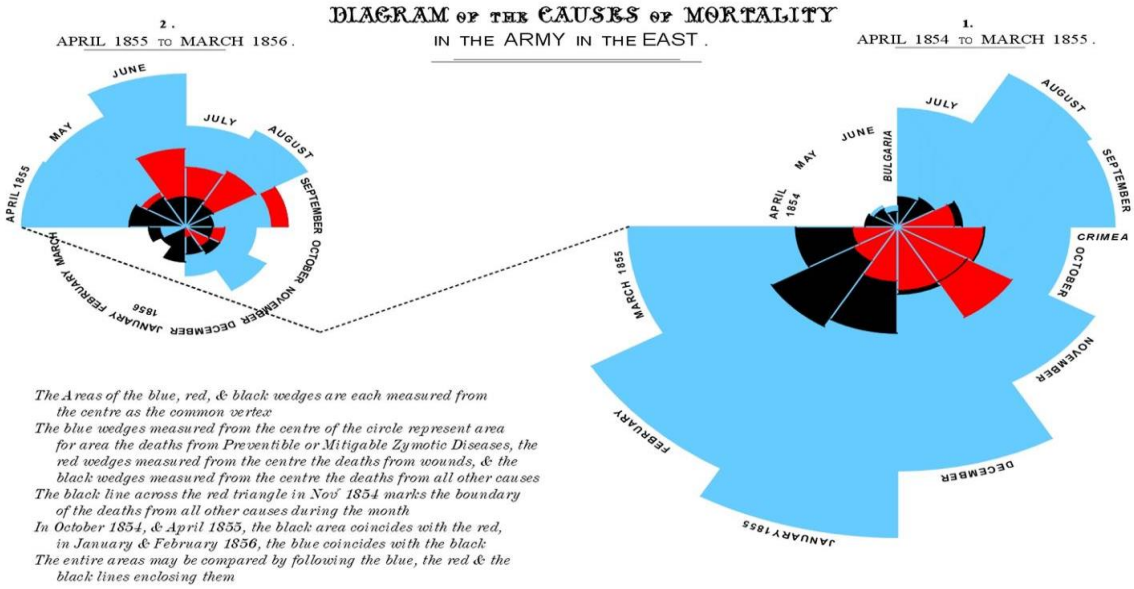
Resume

Catey Bunce is an applied medical statistician working at the Royal Marsden NHS Foundation Trust. She holds an honorary associate professorship at the London School of Hygiene Tropical Medicine and is an honorary consultant in applied medical statistics at Moorfields Eye Hospital. She is co lead of the National Institute for Health Research Medical Statistics group and an ambassador for the Royal Statistical Society championing the message that better data = better research = better healthcare.

Key words: Statistical Communication, medical statistics

Abstract: 12th May 2020 was the bicentenary of the birth of Florence Nightingale. Whilst many are aware of Florence Nightingale as a pioneer of modern nursing, less may be aware that she is an inspirational female leader of applied medical statistics (Cohen 1984). Florence Nightingale recommended by Farr was the first female member of the Royal Statistical Society (Spiegelhalter, 1999). Whilst there have been criticisms of Nightingale's work (Lezzoni 1996 a, b) this has been forcefully rebutted by Vandenbroucke and Vandenbroucke-Grauls (1996) and Vandenbroucke (2003). McDonald (2014) identified that much of the criticism levied against Nightingale was entirely lacking in hard evidence and Beyersmann and Schrade (2015) argue that Nightingale and Farr suggested forms for reporting hospital mortality that were conceptually more complete than many competing risk analyses today. I believe that Florence Nightingale was inspirational in her ability to communicate statistical concepts to those without statistical training and she was perhaps ahead of her peers in using data visualisation in this. Whilst statisticians may see patterns when presented with tables of data, people from other professions or from other academic disciplines may need different images for them to comprehend the same message. Florence Nightingale used her iconic polar chart or Rose diagram to make the importance of improved ventilation and sanitation in military hospitals leap off the page. It was a powerful before and after visualisation which cleverly directed readers towards one interpretation of the data and not another. (Harford 2020) It persuaded senior doctors and politicians of the need for better sanitation and as result death rates fell.

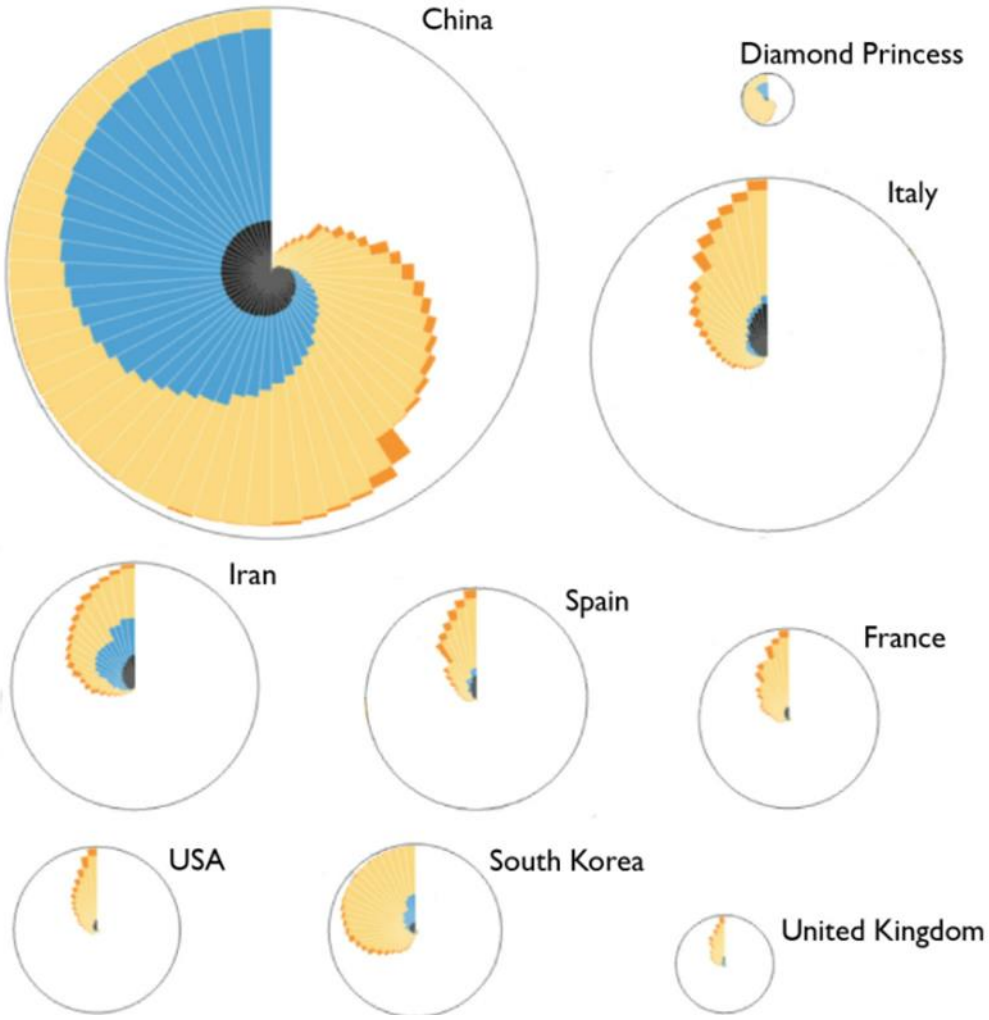
Figure 1: Nightingale's polar chart



There is today evidence of poor statistics within medical research (Smith 2014). There is evidence also that people can suffer harm because of statistical misunderstandings (Bunce 2019). It is important to note that even if the medical research has been conducted and reported correctly, its findings may be miscommunicated. Better communication between statistician and non-statistician has been encouraged for many years (Greenfield 1993). Data visualisation is a vital tool for communication and is one that statisticians need to engage with in order that the underlying statistical messages are accurately conveyed.

Ironically, 2020, the year celebrating the bicentenary of Florence Nightingale's birth, was the year that the COVID-19 pandemic struck the UK and other countries. Tens of millions of lives were potentially at risk and so were billions of people's livelihoods. (Harford 2020). Politicians needed to make vital decisions and to make these decisions quickly using data-informed forecasts. 2020 was arguably the year that data visualisation came into its element and a year that very much highlighted the value of statistical literacy and proper risk communication. (Palmeiro-Silva 2021). Whilst some researchers used conventional methods – for example, polar charts by countries, others realised that the rate of change in information meant that static images did not adequately convey key messages.

Figure 2: Polar charts of COVID-19 by country



https://www.esri.com/arcgis-blog/products/arcgis-pro/mapping/mapping-coronavirus-coxcombs

Animations and interactive tools were developed which provided moving stories illustrating to all how the pandemic was evolving and responding to changes in policy. Whilst Nightingale and Farr had access to a single static dataset which they understood well in terms of data definitions and data quality – statisticians and data scientists in 2020 had multiple data sources. Data definitions were at times unclear and missing data was not always highlighted. Transparency in relation to the robustness of the data was at times lacking and this lack of openness was not conducive to maintaining public confidence. (Royal Statistical Society 2020)

My talk will simply reflect upon the use of data visualisation over time and the challenges facing applied medical statisticians in providing messages that are clear to non-statisticians but do not compromise on statistical integrity to communicate simple and powerful messages. (Blastland 2020)

Bibliography

Beyersmann J and Schrade C (2017) Florence Nightingale, William Farr and competing risks J R Statist Soc A 180 Part 1 pp 285-293

Blastland M, Freeman ALJ, van der Linden S, Marteau TM, Spiegelhalter D. (2020) Five rules for evidence communication. Nature. 2020 Nov;587(7834):362-364.

Bunce C, Stratton IM, Elders A, Czanner G, Dore C, Freemantle N; Ophthalmic Statistics Group (2019) Ophthalmic Statistics Note 13: Method Agreement Studies in Ophthalmology – please don't carry on correlating … Br J Ophthalmol 2019 Sep 103(9) 1201-1293

Cohen IB (1984) Florence Nightingale Scient Am 250 128-137

Greenfield A (1993) Communicating Statistics J R Stat Soc (A) 156 Part 2 pp 287-297

Harford T (2020) How to Make the World Add Up The Bridge Street Press

McDonald L (2014) Florence Nightingale, Statistics and the Crimean War J R Statistic Soc A 177 569-586

Palmeiro-Silva YK, Weinstein-Oppenheimer C, Henríquez-Roldán CF, Bangdiwala S (2021) Statistical literacy and risk communication for COVID-19 vaccination: a scoping review Rev Panam Salud Publica 45 https://doi.org/10.26633/RPSP.2021.108

Royal Statistical Society (2020) Royal Statistical Society Submission to the Joint Health and Social Care Committee and Science Technology Committee Inquiry: 'Coronovirus: lessons learnt' 4.12.2020 https://rss.org.uk/policy-campaigns/search-policy-documents/covid-19/ date accessed 25/2/2022

Smith R (2014) Medical Research – still a scandal The BMJ 31 Jan 2014 https://blogs.bmj.com/bmj/2014/01/31/richard-smith-medical-research-still-a-scandal/ date accessed 25/2/2022

Spiegelhalter D J (1999) Surgical audit: statistical lessons from Nightingale and Codman JR Statist Soc A 162 45-48

Vandenbroucke J (2003) Continuing controversies over 'risk and rates' more than a century after William Farr's 'On Prognosis' Socl Prev Med 48 216-218

Vandenbroucke JP and Vandenbroucke-Grauls C (1996) In defence of Farr and Nightingale Ann Intern Med 125 1014

Scientific Reform and Visual Data Science: Retiring the EDA/CDA dichotomy

Roger Beecham[1]

[1]*School of Geography, University of Leeds, UK. e-mail: r.j.beecham@leeds.ac.uk*

Abstract

Concerns around the replicability of published scientific findings has prompted much introspection into the way in which scientific knowledge is produced. To address issues of data fishing, searching exhaustively for discriminating patterns in a dataset, picking and then publishing those that are statistically significant, an argument is made that research findings should only be claimed through pre-registered confirmatory data analyses. Pre-registration studies are, though, somewhat inimical to the more informal research environments typical of modern applied data analysis ('Data Science'). In this talk I enumerate some of these challenges and demonstrate, through an analysis of road crash data in the UK, how nascent visualization techniques can be used to navigate and inject statistical rigour into contemporary data analysis environments.

Keywords. Exploratory data analysis; confirmatory data analysis; data visualization; replicability crisis; graphical inference, uncertainty visualization.

1 Scientific Reform and Data Science

A key tenet of Scientific Reform is that research findings are claimed through out-of-sample hypothesis tests, pre-registered in advance (Open Science Collaboration, 2015; Amrhein et al., 2019; Devezer et al., 2021; Szollosi and Donkin, 2021). As an organising framework for this, data analysis is separated into two discrete phases: Exploratory Data Analysis (EDA) and Confirmatory Data Analysis (CDA). EDA makes heavy use of graphical methods in order to discover high-level properties and structure in a dataset and to help formulate plans for further analysis. Once this scoping activity has been completed, a set of hypotheses are recorded (pre-registered) and an entirely new dataset is collected. A formal data analysis, CDA, is then conducted to evaluate whether the new data are consistent with the pre-registered hypotheses. Only at this point can empirical findings can be claimed and submitted for publication.

This $EDA \rightarrow preregistration \rightarrow CDA$ workflow is best suited to mature research settings where the theories and methods are well-developed and where prior studies and

data exist that can be used to judge and evaluate observed effect sizes (McIntosh, 2017). In modern applied data analysis settings, where administrative and passively-collected data are variously combined and repurposed (Singleton and Arribas-Bel, 2021), these certainties are difficult to achieve: the potential analysis space is broad and there is no stand-out approach to formulating research questions, selecting datasets and statistical techniques. Additionally, too strict an adherence to separating EDA (informal pattern-finding) from CDA (statistically-grounded enquiry) may not always lead to strong theory and knowledge development. Attempts to impose pre-specification designs onto research settings that rely on secondary or 'found data' may result in knowledge statements that are overly specific or that do not express sufficient detail or richness in outcomes (Szollosi and Donkin, 2021).

2 Visualization and knowledge development

In a recent discussion in the Harvard Data Science Review, Hullman and Gelman (2021b) make the case for a larger role, or a realignment, of 'exploratory' approaches in knowledge generation. They argue that model development should be intrinsic to EDA. Rather than simply displaying descriptive summaries of observed data, an EDA should support inferential thinking by encouraging analysts to consider, and subsequently model for, the processes that might have generated that structure. A key aspect of exploratory analysis is then enabling detailed comparisons against those reference distributions. If the ambition of a CDA is to accept/reject pre-specified hypotheses, then EDA is concerned with '*the particularities of the discrepancies between model and data*' (Hullman and Gelman, 2021b) — or locating and characterising *where* the data depart from the reference distribution.

Data visualization is instrumental to this sort of activity. Hullman and Gelman (2021b) and others (Hullman and Gelman, 2021a; Heer, 2021; Cook et al., 2021) envision data graphics supporting iterative comparison to models of increasing complexity and sophistication. This may happen either explicitly, for example in graphical inference where plots are compared to others generated under simulated data (Buja et al., 2009; Wickham et al., 2010; Beecham et al., 2017; Morris et al., 2019), or indirectly by the way in which the plot is composed and read.

3 Visualization and spurious discovery

A charge against elevating the status of exploratory analysis in this way is that visual approaches are vulnerable to over-interpretation and false discovery. Data graphics emphasise data patterns, but the complexities around those patterns may be overlooked (Hullman and Gelman, 2021b). It is worth mentioning two counter-arguments here.

First is that Hullman and Gelman (2021a) invoke the idea of graphical inference as model check (Gelman, 2004) – where observed data are compared graphically to reference

data replicated under a model. They signpost to recent work on uncertainty visualization and a growing repertoire of empirically tested techniques for representing these reference distributions, deliberately designed to protect against spurious discovery. These techniques have gained widespread use in scientific communication (Data Journalism) and for which there are software libraries to support easy implementation (cf. Kay, 2021a,b).

Second is more fundamental to Hullman and Gelman (2021b)'s central thesis. Rather than constraining graphical inference as model check to narrow null hypothesis tests, as in graphical line-up tests (Buja et al., 2009), Hullman and Gelman (2021a) locate graphical model checks within a Bayesian framework. Their re-thinking of graphical inference accepts flexibility around the features in the observed data being tested – there may be several irregularly-defined estimators – and that expectation will be contingent on analysts' prior knowledge and experience. This is particularly compelling for data analysis settings that rely on 'found data', as conventional statistical procedures may not transfer well to the observations being claimed and therefore requiring scrutiny.

4 SFDS 2022

Motivating this extended abstract was a real data analysis scenario: a project where pedestrian casualties in STATS19 road crash data, the canonical road safety dataset for the UK, were analysed in order identify and justify, with statistical evidence, investment decisions. It was possible to derive many interesting geographic patterns in crash rates from detailed information on crash context and the vehicles and individuals involved. These were nevertheless patterns which could be reported with limited levels of confidence. Uncertainty around datasets, the selection of context variables and of appropriate statistical techniques and reporting mechanisms, meant it was difficult to make crisp statements around priority areas using the sorts of research designs required by the Scientific Reform movement (Devezer et al., 2021). At SFDS 2022 I will outline some of these difficulties, and demonstrate how the model-based visual techniques advocated by Hullman and Gelman (2021b) can be practically implemented to inject greater rigour into exploratory visual analysis activities that remain widespread, and necessary, to modern data analysis.

References

Amrhein, V., D. Trafimow, and S. Greenland
2019. Inferential statistics as descriptive statistics: There is no replication crisis if we don't expect replication. *The American Statistician*, 73(sup1):262–270.

Beecham, R., J. Dykes, W. Meulemans, A. Slingsby, C. Turkay, and J. Wood

2017. Map line-ups: effects of spatial structure on graphical inference. *IEEE Transactions on Visualization & Computer Graphics*, 23(1):391–400.

Buja, A., D. Cook, H. Hofmann, M. Lawrence, E.-K. Lee, D. F. Swayne, and H. Wickham
2009. Statistical inference for exploratory data analysis and model diagnostics. *Royal Society Philosophical Transactions A*, 367(1906):4361–4383.

Cook, D., N. Reid, and E. Tanaka
2021. The foundation is available for thinking about data visualization inferentially. *Harvard Data Science Review*.

Devezer, B., D. J. Navarro, J. Vandekerckhove, and E. O. Buzbas
2021. The case for formal methodology in scientific reform. *R. Soc. open sci*, 8:200805.

Gelman, A.
2004. Exploratory data analysis for complex models. *Journal of Computational and Graphical Statistics*, 13(4):755–779.

Heer, J.
2021. Exploratory analysis and its malcontents. *Harvard Data Science Review*. https://hdsr.mitpress.mit.edu/pub/vszs87oj.

Hullman, J. and A. Gelman
2021a. Challenges in incorporating exploratory data analysis into statistical workflow. *Harvard Data Science Review*. https://hdsr.mitpress.mit.edu/pub/2ym7zm34.

Hullman, J. and A. Gelman
2021b. Designing for interactive exploratory data analysis requires theories of graphical inference. *Harvard Data Science Review*. https://hdsr.mitpress.mit.edu/pub/w075glo6.

Kay, M.
2021a. *ggdist: Visualizations of Distributions and Uncertainty*. R package version 3.0.1.

Kay, M.
2021b. tidybayes: Tidy data and geoms for Bayesian models. R package version 3.0.0.

McIntosh, R. D.
2017. Exploratory reports: A new article type for Cortex. *Cortex*, 96:A1–A4.

Morris, T. P., I. R. White, and M. J. Crowther
2019. Using simulation studies to evaluate statistical methods. 38(11):2074–2102.

Open Science Collaboration
2015. Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716.

Singleton, A. and D. Arribas-Bel
2021. Geographic Data Science. *Geographical Analysis*, 53:61–75.

Szollosi, A. and C. Donkin
2021. Arrested theory development: The misguided distinction between exploratory and confirmatory research. *Perspectives on Psychological Science*, 16(4):717–724.

Wickham, H., D. Cook, H. Hofmann, and A. Buja
2010. Graphical inference for infovis. *IEEE Transactions on Visualization and Computer Graphics (Proc. InfoVis '10)*, 16(6):973–979.

DATA VISUALISATION OF ADVERSE EVENTS IN RANDOMISED CLINICAL TRIALS

Dr Victoria Cornelius[1], Dr Suzie Cro[1], Dr Leila Janani[1] & Dr Rachel Phillips[2]

[1]Imperial Clinical Trials Unit, School of Public Health, Imperial College London

[2]Pragmatic Clinical Trials Unit, Queen Mary University of London

Résumé:

Victoria Cornelius is a Reader in Medical Statistics and Clinical Trials at Imperial College London and Director Designate at Imperial Clinical Trials Unit (ICTU). She leads the Child and Digital Health portfolios and established the ICTU Trial Methodology Group. The Group undertakes research to improve the analysis in clinical trials, including the analysis of adverse events. She also has undertaken research into the design of trials for small populations and rare disease. Her work has primarily been applied in the evaluation of drug interventions and across a range of areas including severe pediatric asthma & allergy, mental health, and critical care. She has been involved and led national collaborative statistical projects for the analysis of laboratory studies and early phase clinical trials research to support methods translation.

Keywords: Adverse events, Harms, Randomised Controlled Trials, Data Visualisation

Background

Methods to analyse efficacy outcomes in randomised controlled trials (RCTs) are well established. As the process of randomisation provides protection against confounding, unadjusted hypothesis tests are frequently used to compare outcomes between arms. [Phillips et al 2019] However, these simple approaches cannot take account of important and frequently occurring features such as repeated measurements and clustering or the benefit for adjustment. With substantial improvements in statistical software it's now trivial to undertake advanced statistical modelling and present this data well. Despite this progress, the analysis and presentation of adverse events (AEs) in trial publications has seen very little progress.

AE data is particularly difficult to analyse due to its multi-faceted nature. One of these features is the large number of AE outcomes that get recorded in a trial. This can be due to the fact that AEs are collected from several sources including spontaneous reports from participants, clinical observations, as well as clinical and biological screening. How to analyse, summarise and present this large volume of data is problematic. In 2016 a collaboration of industry experts and journal editors undertook a consensus and published recommendations to improve AE reporting in trial publications. [Lineberry et al] The recommendations included: containing a summary of clinically important events; reporting important minimum numeric information; and a proposal to encourage the use of graphical displays to visualise AE data.

There is a lack of guidance on what and how to visually display complex AE data. We previously undertook a methodology review to identify statistical methods specifically developed to analyze AE data. [Phillips et al 2020] This current paper examines two visual analysis methods identified to be suitable for any adverse events collected during a trial, and one new approach proposed by the

authors for pre-specified harm outcomes that is particularly valuable for multi-arm studies We explore their value using data from a COVID-19 treatment trial and COVID-19 vaccination trial.

Methods

We first examined two plots identified from a previous methodological review [Phillips et al 2020]. These were the **Volcano plot** [Zink et.al. 2013] and **Dot plot** [Amit et.al. 2008]. Both are suitable for any adverse events collected during the clinical trial (emerging events or pre-specified) and are suitable to display binary, count and time-to-event data. We reanalyzed data reported in two trials. The first was a Covid-19 treatment trial, which was a multi-centre, placebo-controlled trial of remdesivir in adults admitted to hospital with severe COVID-19 infections in Wuhan, China. We extracted aggregated data from the published report to produce the visualisations and compared the interpretation to the original trial report.

We also developed a new approach to visualize data for a COVID-19 vaccine trial we were working on that had a number of pre-specified 'reactogenicity' harm events of interest and multiple arms that we wanted to compare. A visulisation was required to enable us to compare a 'fingerprint' of these harms. In order to facilitate easier between arm comparison and display a summary measure of burden we based our visulisation on a radial graph. The trial compared 7 different vaccines as a 3rd booster vaccine (following 2 original doses) to a control arm. Three of the vaccines were at full and half dose. [Munro et al. 2021] Reactogenicity events are important to understand and compare between different vaccine options as they have economic burden in terms of days off work and school. This information is important for governments to use to inform their policy decisions for population level vaccination deployment.

We developed statistical software in Stata to implement the volcano and dot plot [AEVOLCANO and AEDOT], and the radial graph can be constructed in R using ggradar package.

Results

The **volcano plot** [Zink et.al. 2013] displays the between-arm statistic on the x-axis against the log transformed p-value from a between-arm statistical test of the researcher's choice on the y-axis. An example can be seen in Figure 1. In this plot the absolute risk difference is plotted on the x-axis and the log transformed p-value from the Fisher's exact test is plotted on the y-axis. The summary statistic plotted can be changed to a relative measure such as the risk ratio, odds ratio, incidence rate ratio etc. The corresponding x-axis would then display the log transformed p-value based on the chosen statistical test. To adjust for issues of multiple testing Zink et al. suggest incorporation of a multiplicity adjustment such as the false discovery rate.

Compared to the standard practice of a frequency table, the volcano plot provides a way to immediately identify outlying events. In the absence of any disproportionality in harm between arms we would expect to see a symmetrical curve, with a random scatter of events. The strength of the volcano plot is its provision of an overarching and immediate representation of the AE profile that allows the user to draw attention to a particular signal of interest, but it does not allow a detailed review of all AEs. For example, we can see in Figure 1 fairly symmetrical scatter for most AEs, with only two AEs contributing to the asymmetry. These events have been selected to be labelled in the plot to draw attention to them. Labelling shows that the clear outlier is increased aspartate aminotransferase in the placebo arm (9 (12%) vs 7 (5%)). The greatest harm signal in the remdesivir arm is for rash (2 (3%) vs 11 (7%)).

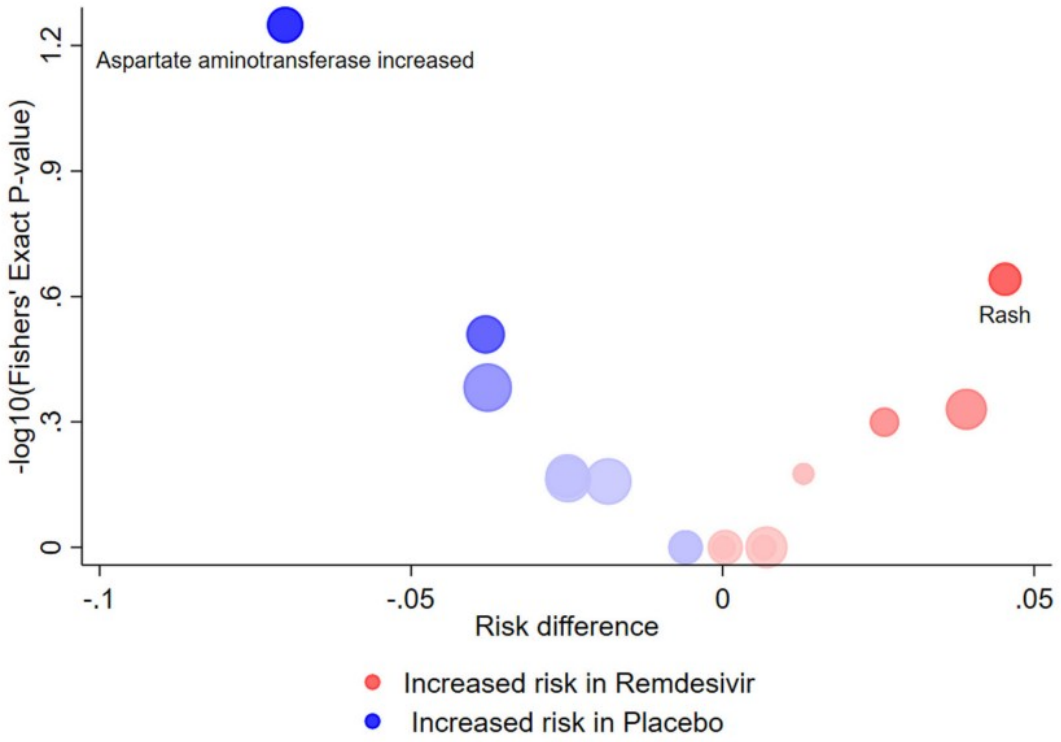


Figure 1: Example of a Volcano Plot for two treatment arms. Colour is used to indicate direction of treatment effect with red indicating greater risk in the Remdesivir group and blue indicating greater risk in the placebo group. Colour saturation corresponds to strength of p-value (level of evidence). The size of the bubble is proportional to the total number of events for both treatment groups combined.

A second valuable plot to display a summary across multiple outcomes is the Dot plot [Amit et.al. 2008]. Unlike the volcano plot, in which only one summary statistic can be displayed at one time, the Dot plot displays both the relative and absolute measure simultaneously. This is particularly important when the event rate is low. The same data in Figure 1 are shown in Figure 2. In this plot the relative risk (RR) with accompanying 95% CIs has been plotted on the central panel and AEs are ordered from the bottom to the top by increasing RR. We can see that the point estimates of the RRs are evenly distributed either side of the vertical line of no difference (RR = 1), and that there is great uncertainty in these estimates. The point estimate furthest away from one, communicates an elevated risk of increased aspartate aminotransferase in the placebo arm, and the upper 95% CI excludes the value of no difference. The corresponding absolute risks can be read from the left-hand side of the graph with the number of adverse events by arm included in the right-hand side of the plot. This detailed information on absolute values can be especially useful in situations such as when the event rate is low. In contrast to the volcano plot, the dot plot requires more assimilation but contains incorporates more information that is useful and we would expect to see in table.

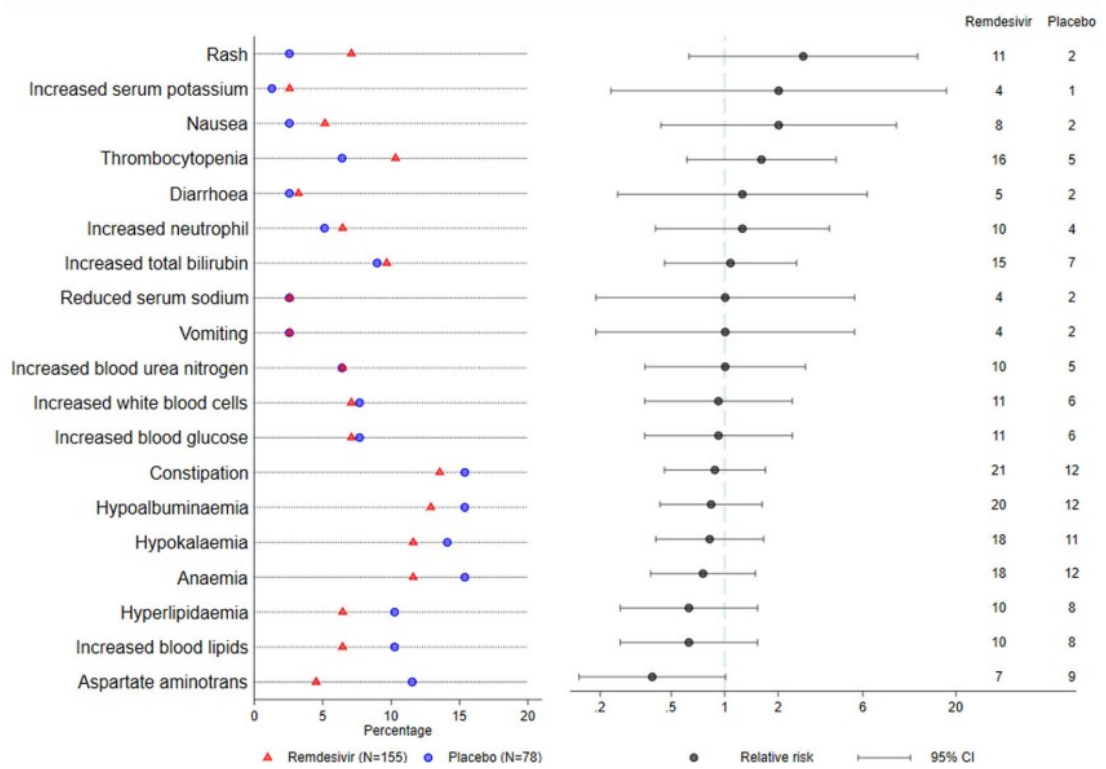


Figure 2. Example of a Dot Plot for two treatment arms. The left-hand side displays the absolute percentage value by treatment group. The right-hand side displays the relative risk and corresponding 95% CI on the log scale

While the dot and volcano plot are helpful to convey between arm imbalances, they do not provide a good summary measure of comparative overall burden. With some trials, such as trials of vaccines, we know the harm outcomes that we would like to compare in advance. When this is the case this enables us to use alternative approaches to the analysis. While rare events cannot be well detected within trials due to the relatively small sample size, we can reliably determine and compare reactogenicity events from vaccination. The standard approach to examining this burden is through a bar chart with stacks by severity. In order to interpret the information across multiple treatment arms in a stacked bar chart the reader has to visually compare several aspects (e.g. height of each section across rows and columns) and there is no direct between arm comparison.

In Figure 3 we present data from the COV-BOOST trial based on the principle of a radial graph. The figure compares different randomized vaccination arms for two different cohort of participants: participants who received two doses of Pfizer (BNT/BNT) prior to their 3rd randomized vaccine arm (right hand side); Participants who received two doses of AstraZeneca (ChAd/ChAd) prior to their 3rd randomized vaccine arm (left hand side). While information on the day that the event occurred is lost compared to a stacked bar chart, the maximum burden over 7 days is shown and the between arm comparison inference can be instantly observed. In Figure 3 it can be seen that the cohort who originally received AstraZeneca (ChAd/ChAd), for all booster vaccine arms, have few events and the rates are similar between arms as the coloured lines sit on top of one another. In contrast, participants who originally received two doses of Pfizer (BNT/BNT) then the third dose with AstraZeneca (ChAd) had higher reactogenicity events. Events of malaise, headache, feverish, fatigue and chill were higher.

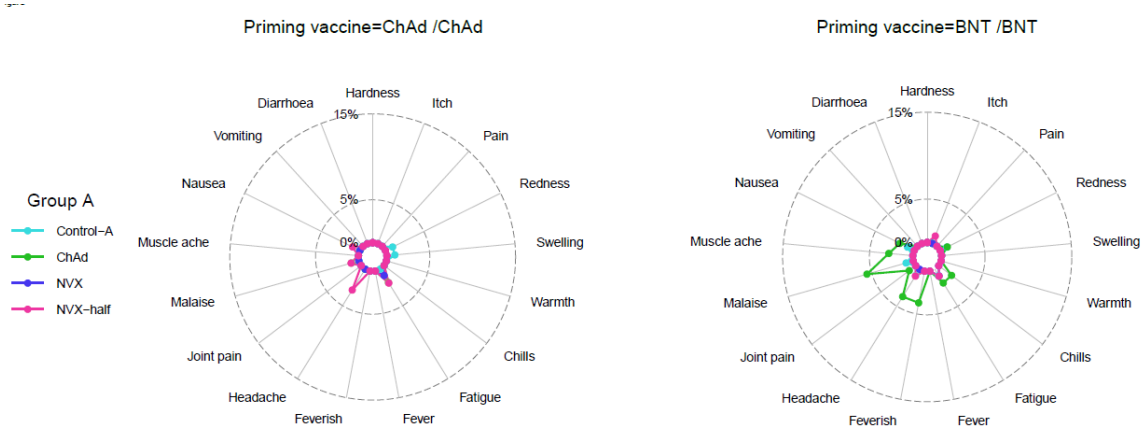


Figure 3: A fingerprint profile of reactogenicity events to facilitate between-arm comparison of the burden to participants. Each spoke is a pre-specified reactogenicity event, and the scale is the proportion of participants with a moderate or severe event. The area defined by each coloured circle is a summary measure of burden for each treatment arm.

Statistical software in Stata has developed to implement the volcano and dot plot [AEVOLCANO and AEDOT], the radial graph can be constructed in R using ggradar package

The volcano plot can be produced in Stata using the command aevolcano, which can be downloaded for use as follows, [AEVOLCANO]

```
ssc install aevolcano
```

The Dot plot can be produced using the command aedot [AEDOT]

```
ssc install aedot
```

Discussion

Assessing the benefit-harm balance is a critical part of a randomized controlled trial as this informs prescribing decisions and the design of future studies. Despite this, harm outcomes in trial publications are poorly presented. Poor presentation can obscure an informative assessment of the true harm profile. Graphical approaches offer a valuable way to summarize large amounts of emerging and pre-specified harm data from RCTs. However, the use of visualizations in journal articles to date has been limited. We have demonstrated the value of three approaches that are easy to implement with standard statistical software. We endorse the routine adoption of visualizations in trial publications as a powerful way to assess and communicate potential harms.

References

Amit O, Heiberger RM and Lane PW. Graphical approaches to the analysis of safety data from clinical trials. *Pharmaceutical Statistics* 2008; 7: 20-35.

Arksey H, O'Malley L. Scoping studies: towards a methodological framework. International Journal of Social Research Methodology. 2005;8(1):19-32.

Cornelius V, Cro S, Phillips R. Advantages of visualisations to evaluate and communicate adverse event information in randomised controlled trials. Trials. 2020; Vol: 21, ISSN: 1745-6215

Lineberry N, Berlin J, Mansi B, C et al. Recommendations to improve adverse event reporting in clinical trial publications: a joint pharmaceutical industry/journal editor perspective. BMJ. 2016; 355.

Munro APS, Janani L, Cornelius V, et al. Safety and immunogenicity of seven COVID-19 vaccines as a third dose (booster) following two doses of ChAdOx1 nCov-19 or BNT162b2 in the UK (COV-BOOST): a blinded, multicentre, randomised, controlled, phase 2 trial, The Lancet. 2021; Vol: 398,

Pitrou I, Boutron I, Ahman N, Ravaud P. Reporting of safety results in published reports of randomized controlled trials. Arch Intern Med 2009; 129 (19): 1756-61.

Phillips R, Hazell L, Suazet O, Cornelius V. Analysis and reporting of adverse events in randomised controlled trials: a review. BMJ Open. 2019.

Phillips R, Cornelius V, Sauzet O. Statistical methods for the analysis of adverse event data in randomised controlled trials: a scoping review and taxonomy, BMC Medical Research Methodology/ 2020; Vol: 20, ISSN: 1471-2288

AEVOLCANO. Phillips R, Cro S.: Stata module to produce volcano plot for adverse event data. https://ideas.repec.org/c/boc/bocode/s458736.html

AEDOT. Phillips R, Cro S.: Stata module to produce dot plot for adverse event data. https://ideas.repec.org/c/boc/bocode/s458735.html

Protocol: An overview of statistical methods developed to analyse adverse events in clinical trials: protocol for a methodological review: https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=97442

Zink RC, Wolfinger RD and Mann G. Summarizing the incidence of adverse events using volcano plots and time intervals. *Clinical Trials* 2013; 10: 398-406.

authors for pre-specified harm outcomes that is particularly valuable for multi-arm studies We explore their value using data from a COVID-19 treatment trial and COVID-19 vaccination trial.

Methods

We first examined two plots identified from a previous methodological review [Phillips et al 2020]. These were the **Volcano plot** [Zink et.al. 2013] and **Dot plot** [Amit et.al. 2008]. Both are suitable for any adverse events collected during the clinical trial (emerging events or pre-specified) and are suitable to display binary, count and time-to-event data. We reanalyzed data reported in two trials. The first was a Covid-19 treatment trial, which was a multi-centre, placebo-controlled trial of remdesivir in adults admitted to hospital with severe COVID-19 infections in Wuhan, China. We extracted aggregated data from the published report to produce the visualisations and compared the interpretation to the original trial report.

We also developed a new approach to visualize data for a COVID-19 vaccine trial we were working on that had a number of pre-specified 'reactogenicity' harm events of interest and multiple arms that we wanted to compare. A visulisation was required to enable us to compare a 'fingerprint' of these harms. In order to facilitate easier between arm comparison and display a summary measure of burden we based our visulisation on a radial graph. The trial compared 7 different vaccines as a 3rd booster vaccine (following 2 original doses) to a control arm. Three of the vaccines were at full and half dose. [Munro et al. 2021] Reactogenicity events are important to understand and compare between different vaccine options as they have economic burden in terms of days off work and school. This information is important for governments to use to inform their policy decisions for population level vaccination deployment.

We developed statistical software in Stata to implement the volcano and dot plot [AEVOLCANO and AEDOT], and the radial graph can be constructed in R using ggradar package.

Results

The **volcano plot** [Zink et.al. 2013] displays the between-arm statistic on the x-axis against the log transformed p-value from a between-arm statistical test of the researcher's choice on the y-axis. An example can be seen in Figure 1. In this plot the absolute risk difference is plotted on the x-axis and the log transformed p-value from the Fisher's exact test is plotted on the y-axis. The summary statistic plotted can be changed to a relative measure such as the risk ratio, odds ratio, incidence rate ratio etc. The corresponding x-axis would then display the log transformed p-value based on the chosen statistical test. To adjust for issues of multiple testing Zink et al. suggest incorporation of a multiplicity adjustment such as the false discovery rate.

Compared to the standard practice of a frequency table, the volcano plot provides a way to immediately identify outlying events. In the absence of any disproportionality in harm between arms we would expect to see a symmetrical curve, with a random scatter of events. The strength of the volcano plot is its provision of an overarching and immediate representation of the AE profile that allows the user to draw attention to a particular signal of interest, but it does not allow a detailed review of all AEs. For example, we can see in Figure 1 fairly symmetrical scatter for most AEs, with only two AEs contributing to the asymmetry. These events have been selected to be labelled in the plot to draw attention to them. Labelling shows that the clear outlier is increased aspartate aminotransferase in the placebo arm (9 (12%) vs 7 (5%)). The greatest harm signal in the remdesivir arm is for rash (2 (3%) vs 11 (7%)).

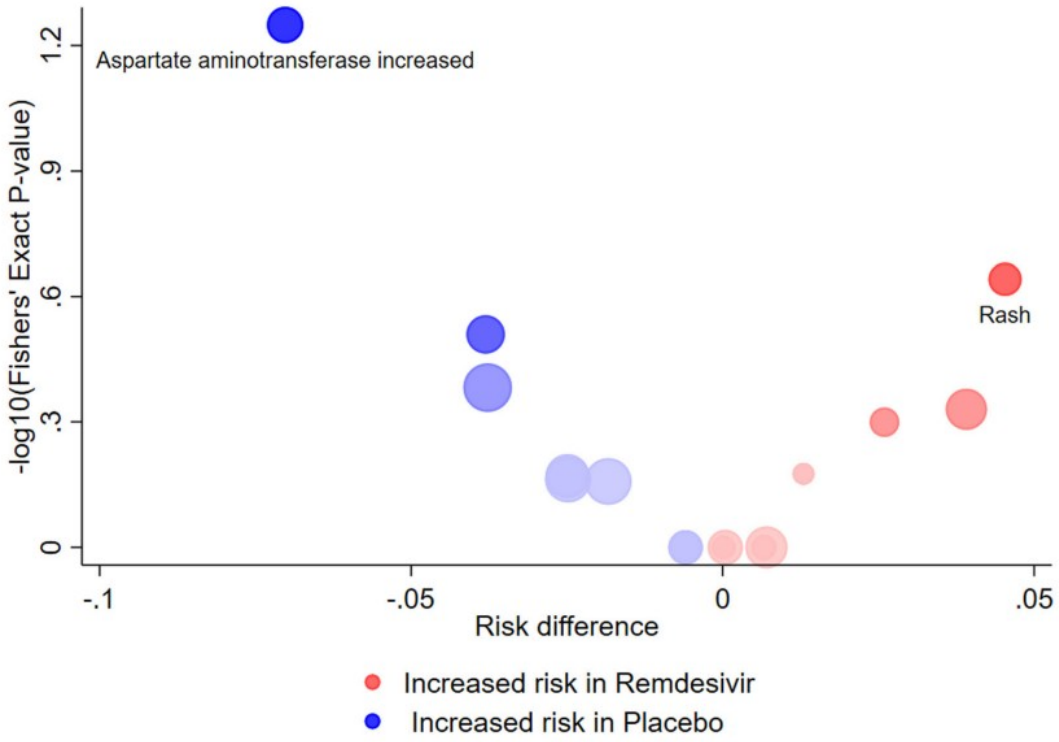


Figure 1: Example of a Volcano Plot for two treatment arms. Colour is used to indicate direction of treatment effect with red indicating greater risk in the Remdesivir group and blue indicating greater risk in the placebo group. Colour saturation corresponds to strength of p-value (level of evidence). The size of the bubble is proportional to the total number of events for both treatment groups combined.

A second valuable plot to display a summary across multiple outcomes is the Dot plot [Amit et.al. 2008]. Unlike the volcano plot, in which only one summary statistic can be displayed at one time, the Dot plot displays both the relative and absolute measure simultaneously. This is particularly important when the event rate is low. The same data in Figure 1 are shown in Figure 2. In this plot the relative risk (RR) with accompanying 95% CIs has been plotted on the central panel and AEs are ordered from the bottom to the top by increasing RR. We can see that the point estimates of the RRs are evenly distributed either side of the vertical line of no difference (RR = 1), and that there is great uncertainty in these estimates. The point estimate furthest away from one, communicates an elevated risk of increased aspartate aminotransferase in the placebo arm, and the upper 95% CI excludes the value of no difference. The corresponding absolute risks can be read from the left-hand side of the graph with the number of adverse events by arm included in the right-hand side of the plot. This detailed information on absolute values can be especially useful in situations such as when the event rate is low. In contrast to the volcano plot, the dot plot requires more assimilation but contains incorporates more information that is useful and we would expect to see in table.

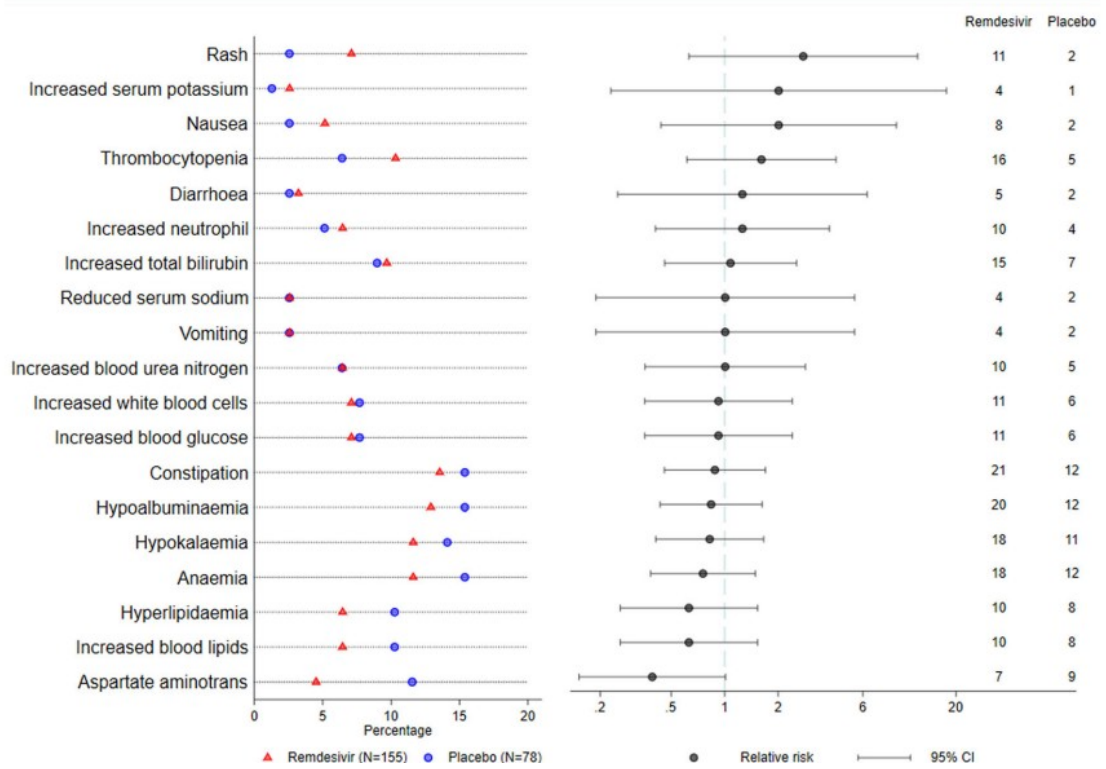


Figure 2. Example of a Dot Plot for two treatment arms. The left-hand side displays the absolute percentage value by treatment group. The right-hand side displays the relative risk and corresponding 95% CI on the log scale

While the dot and volcano plot are helpful to convey between arm imbalances, they do not provide a good summary measure of comparative overall burden. With some trials, such as trials of vaccines, we know the harm outcomes that we would like to compare in advance. When this is the case this enables us to use alternative approaches to the analysis. While rare events cannot be well detected within trials due to the relatively small sample size, we can reliably determine and compare reactogenicity events from vaccination. The standard approach to examining this burden is through a bar chart with stacks by severity. In order to interpret the information across multiple treatment arms in a stacked bar chart the reader has to visually compare several aspects (e.g. height of each section across rows and columns) and there is no direct between arm comparison.

In Figure 3 we present data from the COV-BOOST trial based on the principle of a radial graph. The figure compares different randomized vaccination arms for two different cohort of participants: participants who received two doses of Pfizer (BNT/BNT) prior to their 3rd randomized vaccine arm (right hand side); Participants who received two doses of AstraZeneca (ChAd/ChAd) prior to their 3rd randomized vaccine arm (left hand side). While information on the day that the event occurred is lost compared to a stacked bar chart, the maximum burden over 7 days is shown and the between arm comparison inference can be instantly observed. In Figure 3 it can be seen that the cohort who originally received AstraZeneca (ChAd/ChAd), for all booster vaccine arms, have few events and the rates are similar between arms as the coloured lines sit on top of one another. In contrast, participants who originally received two doses of Pfizer (BNT/BNT) then the third dose with AstraZeneca (ChAd) had higher reactogenicity events. Events of malaise, headache, feverish, fatigue and chill were higher.

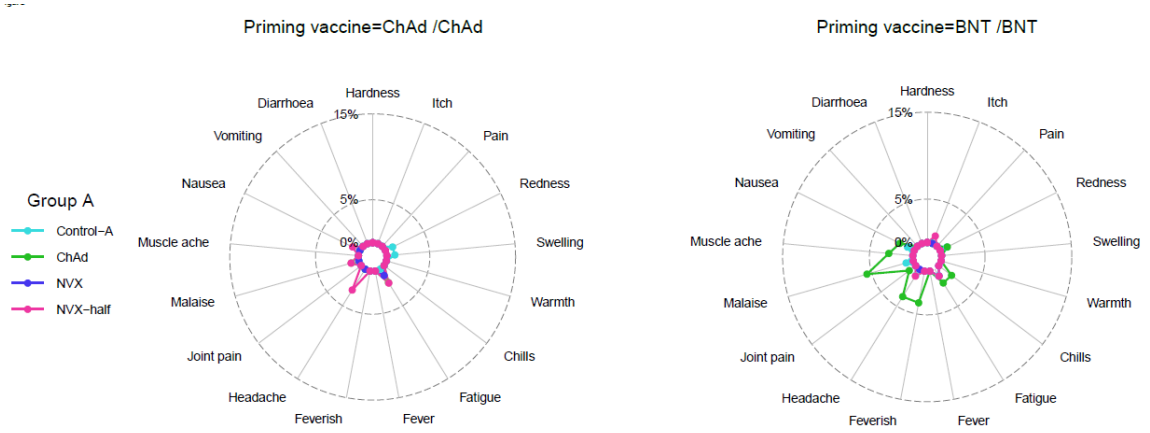


Figure 3: A fingerprint profile of reactogenicity events to facilitate between-arm comparison of the burden to participants. Each spoke is a pre-specified reactogenicity event, and the scale is the proportion of participants with a moderate or severe event. The area defined by each coloured circle is a summary measure of burden for each treatment arm.

Statistical software in Stata has developed to implement the volcano and dot plot [AEVOLCANO and AEDOT], the radial graph can be constructed in R using ggradar package

The volcano plot can be produced in Stata using the command aevolcano, which can be downloaded for use as follows, [AEVOLCANO]

```
ssc install aevolcano
```

The Dot plot can be produced using the command aedot [AEDOT]

```
ssc install aedot
```

Discussion

Assessing the benefit-harm balance is a critical part of a randomized controlled trial as this informs prescribing decisions and the design of future studies. Despite this, harm outcomes in trial publications are poorly presented. Poor presentation can obscure an informative assessment of the true harm profile. Graphical approaches offer a valuable way to summarize large amounts of emerging and pre-specified harm data from RCTs. However, the use of visualizations in journal articles to date has been limited. We have demonstrated the value of three approaches that are easy to implement with standard statistical software. We endorse the routine adoption of visualizations in trial publications as a powerful way to assess and communicate potential harms.

References

Amit O, Heiberger RM and Lane PW. Graphical approaches to the analysis of safety data from clinical trials. *Pharmaceutical Statistics* 2008; 7: 20-35.

Arksey H, O'Malley L. Scoping studies: towards a methodological framework. International Journal of Social Research Methodology. 2005;8(1):19-32.

Cornelius V, Cro S, Phillips R. Advantages of visualisations to evaluate and communicate adverse event information in randomised controlled trials. Trials. 2020; Vol: 21, ISSN: 1745-6215

Lineberry N, Berlin J, Mansi B, C et al. Recommendations to improve adverse event reporting in clinical trial publications: a joint pharmaceutical industry/journal editor perspective. BMJ. 2016; 355.

Munro APS, Janani L, Cornelius V, et al. Safety and immunogenicity of seven COVID-19 vaccines as a third dose (booster) following two doses of ChAdOx1 nCov-19 or BNT162b2 in the UK (COV-BOOST): a blinded, multicentre, randomised, controlled, phase 2 trial, The Lancet. 2021; Vol: 398,

Pitrou I, Boutron I, Ahman N, Ravaud P. Reporting of safety results in published reports of randomized controlled trials. Arch Intern Med 2009; 129 (19): 1756-61.

Phillips R, Hazell L, Suazet O, Cornelius V. Analysis and reporting of adverse events in randomised controlled trials: a review. BMJ Open. 2019.

Phillips R, Cornelius V, Sauzet O. Statistical methods for the analysis of adverse event data in randomised controlled trials: a scoping review and taxonomy, BMC Medical Research Methodology/ 2020; Vol: 20, ISSN: 1471-2288

AEVOLCANO. Phillips R, Cro S.: Stata module to produce volcano plot for adverse event data. https://ideas.repec.org/c/boc/bocode/s458736.html

AEDOT. Phillips R, Cro S.: Stata module to produce dot plot for adverse event data. https://ideas.repec.org/c/boc/bocode/s458735.html

Protocol: An overview of statistical methods developed to analyse adverse events in clinical trials: protocol for a methodological review: https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=97442

Zink RC, Wolfinger RD and Mann G. Summarizing the incidence of adverse events using volcano plots and time intervals. *Clinical Trials* 2013; 10: 398-406.

Statistique spatiale

Fully nonseparable Gneiting covariance functions for multivariate space-time data

Denis Allard[1] & Lucia Clarotto[2] & Xavier Emery[3]

[1] *Biostatistiques et Processus Spatiaux (BioSP), INRAE, 84914 Avignon, France*
[2] *Centre de Géosciences, Mines Paris, PSL University, 77300 Fontainebleau, France*
[3] *Department of Mining Engineering, University of Chile, and Advanced Mining Technology Center, University of Chile, Avenida Beauchef 850, Santiago 8370448, Chile*

Résumé. Nous élargissons la classe de Gneiting des fonctions de covariance spatio-temporelle en introduisant une classe très générale de fonctions de covariance multivariées entièrement non séparables. Chaque composante est modélisée par une fonction de covariance spatiale de la famille Matérn avec ses propres paramètres de lissage et d'échelle et, contrairement à tous les modèles actuellement disponibles, sa propre fonction de corrélation dans le temps. Ce modèle est illustré sur un jeu de données météorologiques trivariées. Notre nouveau modèle mène à un meilleur ajustement et de meilleurs scores prédictifs comparé à un modèle plus parcimonieux avec une fonction de corrélation temporelle commune.

Mots-clés. Modélisation spatio-temporelle, fonction de covariances maultivariées; covariances de Matérn

Abstract. We broaden the well-known Gneiting class of space-time covariance functions by introducing a very general parametric class of fully nonseparable direct and cross-covariance functions for multivariate random fields, where each component has a spatial covariance function from the Matérn family with its own smoothness and scale parameters and, unlike all currently available models, its own correlation function in time. The application of the proposed model is illustrated on a weather trivariate dataset over France. Our new model yields better fitting and better predictive scores compared to a more parsimonious model with common temporal correlation function.

Keywords. Spatio-temporal modeling; multivariate covariance function; Matérn covariance;

1 Introduction and background material

An increasing amount of multivariate data indexed by space-time coordinates are available in numerous scientific and engineering applications, which range from environmental and climate studies to remote sensing and monitoring systems in industrial operations and process management. For statisticians analyzing these data, one of the key issues is to

model the space-time dependence structure not only for each variable separately, but also across the variables. This requires versatile models that can account for different scale and smoothness parameters for each variable, and yet whose parameters can be accurately estimated. Bourotte et al. (2016) proposed classes of matrix-valued space-time covariance functions that are extensions of the univariate Gneiting class, where the completely monotone real-valued function φ is replaced by a matrix-valued function $\boldsymbol{\varphi} = [\varphi_{ij}]_{i,j=1}^{p}$, with each component φ_{ij} having specific parameters. In particular, they showed that the Gneiting-Matérn space-time model $\mathbf{C}^{\mathcal{M}} = \left[C_{ij}^{\mathcal{M}}\right]_{i,j=1}^{p}$ with

$$C_{ij}^{\mathcal{M}}(\boldsymbol{h}, u) = \frac{\sigma_i \sigma_j \rho_{ij}}{(\gamma(u)+1)^{d/2}} \mathcal{M}\left(\boldsymbol{h}; \frac{r_{ij}}{(\gamma(u)+1)^{1/2}}, \nu_{ij}\right), \quad (\boldsymbol{h}, u) \in \mathbb{R}^d \times \mathbb{R}, \tag{1}$$

is a valid second-order stationary matrix-valued covariance function if the matrices $\boldsymbol{\nu} = [\nu_{ij}]_{i,j=1}^{p}$ and $\boldsymbol{r}^2 = [r_{ij}^2]_{i,j=1}^{p}$ have positive entries such that $\nu_{ij} = \nu_i + \nu_j$, $r_{ij}^2 = 2r_i^2 + r_j^2$ with $\boldsymbol{\rho}\, \boldsymbol{r}^{2\boldsymbol{\nu}} / \Gamma(\boldsymbol{\nu}) \succeq 0$. Here and in the reste of this paper, all matrix operations are meant componentwise. Setting $\boldsymbol{h} = \mathbf{0}$, the entries of the associated temporal matrix-valued covariance functions are $C_{ij}^{\mathcal{M}}(\mathbf{0}, u) = \sigma_i \sigma_j \rho_{ij} (\gamma(u)+1)^{-d/2}$. These are proportional-in-time models, as they are the product of a covariance matrix with a single time correlation function $\rho_T(u) = (\gamma(u)+1)^{-d/2}$.

In this work, we present a new class of multivariate space-time covariance functions with Matérn spatial traces that allow, for each variable, different correlation functions in time and different smoothness and scale parameters in space. We present a very general class of covariance functions of the type

$$C_{ij}(\boldsymbol{h}, u) = \sigma_{ij}(u) \mathcal{M}(\boldsymbol{h}; r_{ij}(u), \nu_{ij}(u)), \qquad (\boldsymbol{h}, u) \in \mathbb{R}^d \times \mathbb{R},\ i, j = 1, \ldots, p,$$

where $\mathcal{M}(\cdot; r, \nu)$ denotes a univariate Matérn covariance function with scale and smoothness parameters r and ν. In order to build multivariate models that are neither proportional in space nor in time, an essential building block is to define admissible and relevant temporal matrix-valued covariances of the form $[(\gamma_{ij}(u)+1)^{-d/2}]_{i,j=1}^{p}$. A more detailed version of this work can be found on-line (Allard et al., 2022).

2 A general fully nonseparable Gneiting-Matérn class

2.1 Main result

This construction uses results on matrix-valued spatial covariance functions establihed in Emery et al. (2022).

Theorem 1. *Let $\boldsymbol{\nu}$, $\boldsymbol{b}^2$ and $\boldsymbol{a}^2$ be $p \times p$ symmetric conditionally negative semidefinite matrices, all with positive entries, such that $\boldsymbol{a}^{\boldsymbol{2}} - \boldsymbol{\nu} \preceq_c 0$. Let $u \mapsto \boldsymbol{\rho}(u)$ be a $p \times p$*

offset

matrix-valued covariance function on $\mathbb{R}$ *and let* $u \mapsto \boldsymbol{\eta}(u)$ *be a* $p \times p$ *matrix-valued pseudo-variogram on* $\mathbb{R}$*. Denoting* $\boldsymbol{r}(u) = [r_{ij}(u)]_{i,j=1}^p$ *and* $\boldsymbol{\sigma}(u) = [\sigma_{ij}(u)]_{i,j=1}^p$*, the matrix-valued function* $\boldsymbol{C} : (\boldsymbol{h}, u) \mapsto [C_{ij}(\boldsymbol{h}, u)]_{i,j=1}^p$ *with*

$$C_{ij}(\boldsymbol{h}, u) = \sigma_{ij}(u)\, \mathcal{M}(\boldsymbol{h}; r_{ij}(u), \nu_{ij}), \quad (\boldsymbol{h}, u) \in \mathbb{R}^d \times \mathbb{R},$$

is positive semidefinite in $\mathbb{R}^d \times \mathbb{R}$ *for*

$$\boldsymbol{r}(u) = \sqrt{\frac{\alpha\boldsymbol{\eta}(u) + \boldsymbol{a}^2}{\beta\boldsymbol{\eta}(u) + \boldsymbol{b}^2}} \text{ and } \boldsymbol{\sigma}(u) = \frac{\boldsymbol{\rho}(u)\Gamma(\boldsymbol{\nu})\exp(\boldsymbol{\nu})}{(\alpha\boldsymbol{\eta}(u) + \boldsymbol{a}^2)^{\boldsymbol{\nu}} \left(\beta\boldsymbol{\eta}(u) + \boldsymbol{b}^2\right)^{d/2}}, \tag{2}$$

where α *and* β *are nonnegative.*

Corollary 1. *Let* $\boldsymbol{\nu}$*,* $\boldsymbol{b}^2$*,* $\boldsymbol{a}^2$*,* $\boldsymbol{\eta}$*,* α *and* β *be as in Theorem 1. Let* $\boldsymbol{\tau}$ *be a* $p \times p$ *symmetric real matrix. The matrix-valued function* $\mathbf{C} : (\boldsymbol{h}, u) \mapsto [C_{ij}(\boldsymbol{h}, u)]_{i,j=1}^p$ *with*

$$C_{ij}(\boldsymbol{h}, u) = \frac{\tau_{ij}}{(\beta\eta_{ij}(u) + b_{ij}^2)^{d/2}(\alpha\eta_{ij}(u) + a_{ij}^2)^{\nu_{ij}}} \mathcal{M}\left(\boldsymbol{h}; \sqrt{\frac{\alpha\eta_{ij}(u) + a_{ij}^2}{\beta\eta_{ij}(u) + b_{ij}^2}}, \nu_{ij}\right), \tag{3}$$

is positive semidefinite in $\mathbb{R}^d \times \mathbb{R}$ *if the matrix* $\boldsymbol{\tau} e^{-\boldsymbol{\nu}}/\Gamma(\boldsymbol{\nu})$ *is positive semidefinite.*

See Allard et al. (2022) for a proof and more mathematical details. When $\alpha = 0$, accounting for the fact that $\exp(-\boldsymbol{\nu})$ is positive semidefinite, the models in (2) and (3) generalize the Gneiting-Matérn model (1). If, furthermore, the direct variograms η_{ii} are increasing functions for $i = 1, \ldots, p$, the scale parameters $r_{ii}(u) = a_{ii}(\beta\eta_{ii}(u) + b_{ii}^2)^{-1/2}$ decrease as u increases. The space-time direct correlations are thus higher than what would happen for a separable space-time covariance function, i.e., they are positively nonseparable. When $\beta = 0$ and under the same monotonicity assumption for the direct variograms, the scale parameters $r_{ii}(u) = \sqrt{\alpha\eta_{ii}(u) + a_{ii}^2}/b_{ii}$ increase with u. This a very original feature, since for all Gneiting-type spatio-temporal models of our knowledge, the scale parameter decreases with u. Applying the same reasoning as above, the resulting space-time direct covariances are thus negatively nonseparable. When α and β are both positive, the model is positively or negatively space-time nonseparable, depending on the relative values of the other parameters involved in $\boldsymbol{r}(u)$. However, in the common case where $\eta_{ij}(u) \to \infty$ as $u \to \infty$, $\boldsymbol{r}(u)$ asymptotically tends to a matrix with all elements equal to $\sqrt{\alpha/\beta}$. The space-time covariance is thus asymptotically separable as $|u| \to \infty$.

2.2 Models for unbounded pseudo-variograms

As a simple model for an unbounded matrix-valued pseudo-variogram with different diagonal entries (direct variograms), we propose the following construction. Let us define $W_i(u) = Y_0(u) + Y_i(u),\ u \in \mathbb{R}$, where $Y_1, \ldots, Y_p$ are second-order stationary random fields

with direct and cross-covariance functions $\boldsymbol{R} : u \mapsto [R_{ij}(u)]_{i,j=1}^{p}$ and Y_0 is an independent intrinsic random field with unbounded variogram γ_0. Then, the pseudo-variogram of $\boldsymbol{W} = [W_i]_{i=1}^{p}$ is such that $\eta_{ij}(u) = \gamma_0(u) + (R_{ii}(0) + R_{jj}(0))/2 - R_{ij}(u),\ u \in \mathbb{R}$. In matrix notation, a valid class of unbounded matrix-valued pseudo-variogram is thus

$$\boldsymbol{\eta}(u) = \gamma_0(u)\mathbf{1}_p + \boldsymbol{R}^0 - \boldsymbol{R}(u), \quad u \in \mathbb{R}, \tag{4}$$

where $\mathbf{1}_p$ is the all-ones matrix of size $p \times p$, $\boldsymbol{R}$ is any matrix-valued covariance function and $\boldsymbol{R}^0$ is the a-separable matrix with entries $R^0_{ij} = (R_{ii}(0) + R_{jj}(0))/2$.

3 Application to a weather dataset

3.1 Introduction

We now illustrate this model on a weather dataset consisting of three daily variables (solar radiation R, temperature T, and humidity H) recorded at 13 stations in Western France from 2003 to 2012. We shall compare a specific fully nonseparable multivariate model (FULL), as proposed in (3), to a Proportional-In-time (PIT) model. Both models will be applied to residuals, after centering in time and space, in order to filter out any seasonal and regional effect. For each of the three variables, the separability index $S(\boldsymbol{h}, u) = C(\boldsymbol{h}, u)C(\mathbf{0}, 0) - C_S(\boldsymbol{h})C_T(u)$ was computed. Empirical covariances were derived from the space-time empirical variograms computed with the function `EVariogram` from the package `CompRandFld` (Padoan and Bevilacqua, 2015), using spatial bins of 50 m. Since the variograms have been computed on residuals, $C(\mathbf{0}, 0) = 1$ and the separability index must be in the interval $[-1, 1]$, with $S(\boldsymbol{h}, u) = 0$ indicating separability of $C(\boldsymbol{h}, u)$. This analyses showed clearly that the residuals are uniformly positively space-time nonseparable for $|\boldsymbol{h}| \leq 500$ m and $u \leq 3$, which is actually often the case for climate variables. The parameter α in (3) is thus set to $\alpha = 0$. For identifiability reasons, we further set $\beta = 1$ and $b_{ij}^2 = 1 - R^0_{ij}$ with $\max_i\{R_{ii}(0)\} < 1$.

The pseudo-variogram $\boldsymbol{\eta}$ needs to be specified. For the unbounded part common to all components, we choose $\gamma_0(u) = (1 + |cu|^{2a})^b - 1, \quad u \in \mathbb{R}$ with $c > 0, 0 < a \leq 1$ and $0 \leq b \leq 1$, for ease of comparison with Bourotte et al. (2016). This variogram is reminiscent of the generalized Cauchy covariance model that will also be used for $\boldsymbol{R}$: $R_{ij}(u) = A_iA_j(1+|ru|^{2\lambda})^{-b},\ u \in \mathbb{R}$, with $r > 0$, $0 < \lambda < 1$ and $0 \leq A_i < 1$ for $i = 1, \ldots, p$, and where for the sake of parsimony the parameter b is set identical to the parameter b in γ_0. In this parameterization, the space–time nonseparability parameter b acts both inside and outside the Matérn function $\mathcal{M}$, implying a constant perfect correlation in time if $b = 0$. Following Gneiting (2002) and Bourotte et al. (2016), a reparameterization is thus useful. Multiplying the above equation by the temporal covariance function $\tilde{\rho}_{ij}(u) = \left[(1 + c|u|^{2a})^{\delta} - A_iA_j(1 + r|u|^{2\lambda})^{-\delta}\right]^{-1}$ with $0 \leq \delta \leq 1$ leads to the parametric family that will be used in the rest of this section. The parameterization is further simplified by

setting $\tau_{ii} = \sigma_i^2 a_{ii}^{2\nu_{ii}}(1 - A_i^2)^2$ and, for $i \neq j$, $\tau_{ij} = \rho_{ij}\sqrt{\tau_{ii}\tau_{jj}}$, with $[\rho_{ij}]_{i,j=1,\ldots,p}$ being a correlation matrix. With this parameterization, the matrix $\boldsymbol{\tau}$ is positive semidefinite by construction. Finally, the model reads

$$C_{ij}(\boldsymbol{h}, u) = \frac{1}{(1 + |cu|^{2a})^{\delta} - A_i A_j (1 + |ru|^{2\lambda})^{-\delta}} \frac{\rho_{ij} a_{ij}^{-2\nu_{ij}} \sqrt{\tau_{ii}\tau_{jj}}}{(1 + |cu|^{2a})^{b} - A_i A_j (1 + |ru|^{2\lambda})^{-b}} \times \mathcal{M}\left(\boldsymbol{h}; \frac{a_{ij}}{[(1 + |cu|^{2a})^{b} - A_i A_j (1 + |ru|^{2\lambda})^{-b}]^{1/2}}, \nu_{ij}\right), \quad (5)$$

with $(\boldsymbol{h}, u) \in \mathbb{R}^d \times \mathbb{R}$ and $i, j = 1, \ldots, p$. The model in (5) provides interpretable parameters for a fully nonseparable Gneiting-Matérn model.

3.2 Estimation

Composite likelihood methods have proven efficient in the context of spatio-temporal data, with less time-consuming steps and good asymptotic properties (Bevilacqua et al., 2012). Pairwise Likelihood (PL) is the product of marginal Gaussian likelihoods computed on all pairs $\{Z_i(\boldsymbol{s}_\alpha, t_\alpha), Z_j(\boldsymbol{s}_\beta, t_\beta)\}$ such that $|\boldsymbol{s}_\alpha - \boldsymbol{s}_\beta| \leq d_m$ and $|t_\alpha - t_\beta| \leq t_m$, where $i, j = 1, \ldots, p$ and $\alpha, \beta = 1, \ldots, N$. The data analyzed here are standardized residuals after centering for spatial and temporal trends. We thus set $\sigma_i = 1$, for $i = 1, \ldots, p$. There is a total of $p(p+2)+6$ parameters to estimate, where p is the number of variables. Here $p = 3$, leading to a total of 21 parameters. PL was maximized sequentially in subspaces of the parameter space corresponding to blocks of related parameters, while keeping all other parameters fixed to the values previously attained. Finally $\hat{b}$ is the value of b among $\{0, 0.1, \ldots, 0.9, 1\}$ corresponding to the highest maximized PL. To perform the maximization in the subspaces of the parameter space, we used the package `nlminb` implemented in `R` with lower and upper bounds for the parameters when mathematically necessary. Estimated parameters and fit of the model to the data can be found in Allard et al. (2022).

3.3 Prediction

We consider two different prediction settings. In the first setting, called *Spatial Interpolation* , prediction of the three variables at a validation station at day t is made knowing the trivariate data at the 11 estimation stations at day t, $t-1$ and $t-2$. In the second setting, called *Temporal Prediction*, prediction is made with data from days $t-1$ and $t-2$ only at all 13 stations (estimation and validation stations). Other settings, such as Spatio-temporal prediction (using data at 11 stations from days $t-1$ and $t-2$ for predicting at the 2 validations stations at day t) is possible, but not shown here because they are less common in practical situations. Predictions are computed from January 3rd to January 31s. In the Gaussian framework, the conditional expectation is the best

offset

linear unbiased prediction, also called (co)kriging in the geostatistics literature, while the conditional covariance matrix is the covariance matrix of the cokriging errors. The conditional variances are identical for all days since the configuration of the conditioning data remains identical. The two models (FULL and PIT) are compared by means of the Root Mean Square Error (RMSE), the Mean Absolute Error (MAE), the Continuous Ranked Probability Score (CRPS) and the Logarithmic Score (LogS). Table 1 reports our results. As a benchmark, we also report the scores obtained with a trivial prediction where all predicted values are set to the expectation (equal to 0) and all variances to the theoretical variance (equal to 1).

	RMSE			MAE			CRPS			LogS		
	R	T	H	R	T	H	R	T	H	R	T	H
	Spatial interpolation											
FULL	**0.746**	**0.582**	**0.864**	**0.587**	0.464	0.732	**0.427**	**0.335**	**0.509**	**0.268**	0.015	**0.393**
PIT	0.751	0.584	**0.864**	0.593	**0.462**	**0.731**	0.430	**0.335**	**0.509**	0.271	**0.010**	**0.393**
	Temporal prediction											
FULL	**0.919**	0.837	**0.813**	0.784	**0.706**	**0.667**	**0.531**	**0.482**	**0.470**	**0.421**	**0.329**	**0.311**
PIT	0.933	**0.835**	0.825	**0.780**	0.714	0.684	0.537	0.483	0.479	0.437	0.331	0.330
	Trivial interpolation											
	0.933	0.893	0.919	0.801	0.777	0.782	0.540	0.521	0.538	0.439	0.403	0.443

Table 1: Prediction scores at the validation stations at day t, using ML estimates. FULL: fully nonseparable model as in (5) with $b = 0.1$. PIT: simplified Proportional-In-Time model with $b = 0$. The best scores among FULL or PIT are shown in bold font.

References

Allard, D., Clarotto, L., and Emery, X. (2022). Fully nonseparable Gneiting covariance functions for multivariate space-time data. Preprint *HAL:* 03564931.

Bevilacqua, M., Gaetan, C., Mateu, J., and Porcu, E. (2012). Estimating space and space-time covariance functions for large data sets: a weighted composite likelihood approach. *Journal of the American Statistical Association*, 107(497):268–280.

Bourotte, M., Allard, D., and Porcu, E. (2016). A flexible class of non-separable cross-covariance functions for multivariate space–time data. *Spatial Statistics*, 18:125–146.

Emery, X., Porcu, E., and White, P. (2022). Flexible validity conditions for the multivariate Matérn covariance in any spatial dimension and for any number of components. *arXiv:2101.04235v2 [stat.ME].*

Gneiting, T. (2002). Nonseparable, stationary covariance functions for space–time data. *Journal of the American Statistical Association*, 97(458):590–600.

Padoan, S. A. and Bevilacqua, M. (2015). Analysis of random fields using CompRandFld. *Journal of Statistical Software*, 63:1–27.

Investigating spatial scan statistics for multivariate functional data

Camille Frévent [1,3] & Mohamed-Salem Ahmed [1] & Sophie Dabo-Niang [2,3] & Michaël Genin [1]

[1] *Univ. Lille, CHU Lille, ULR 2694 - METRICS: Évaluation des technologies de santé et des pratiques médicales, F-59000 Lille, France.* camille.frevent@univ-lille.fr, mohamed-salem.ahmed@univ-lille.fr, michael.genin@univ-lille.fr

[2] *Univ. Lille, CNRS, UMR 8524 - Laboratoire Paul Painlevé, F-59000 Lille, France* sophie.dabo@univ-lille.fr

[3] *INRIA-MODAL, University of Lille, France*

Résumé. Ce travail introduit de nouvelles statistiques de scan spatiales pour des données fonctionnelles multivariées indexées dans l'espace. Les méthodes sont dérivées d'un test de la MANOVA pour données fonctionnelles, une adaptation du test d'Hotelling, et une extention multivariée du test de Wilcoxon. Dans une étude de simulation les performances des méthodes sont investiguées. Nous les avons ensuite appliquées sur des données fonctionnelles multivariées mesurées avec une très fine résolution spatiale pour détecter des clusters de pollution dans le nord de la France en octobre 2021.
Mots-clés. Clusters spatiaux, données fonctionnelles multivariées, MANOVA, statistiques de scan spatiales, test d'Hotelling, test de Wilcoxon

Abstract. This paper introduces new spatial scan statistics for multivariate functional data indexed in space. The methods are derivated from a MANOVA test statistic for functional data, an adaptation of the Hotelling T^2-test statistic, and a multivariate extension of the Wilcoxon rank-sum test statistic. In a simulation study, the performances of the methods are investigated. Then we apply the methods on multivariate functional data measured at a very fine spatial scale to search for spatial clusters of abnormal daily concentrations of air pollutants in the north of France in October 2021.
Keywords. Hotelling T^2-test, MANOVA test, multivariate functional data, spatial clusters, spatial scan statistics, Wilcoxon rank-sum test

1 Introduction

Les méthodes de détection de clusters spatiaux ont été longuement étudiées. Le but est de développer des outils capables de détecter l'aggrégation de sites qui se comportent différemment. Les statistiques de scan spatiales permettent de détecter des clusters spatiaux sans information *a priori* quant à leur localisation. Elles ont été principalement

proposées par Kulldorff et Nagarwalla (1995) et Kulldorff (1997) dans le cas de modèles de Poisson et de Bernoulli. D'autres statistiques de scan ont ensuite été proposées pour d'autres modèles de distribution (Kulldorff *et al.* (2009), Cucala *et al.* (2017)).
Aujourd'hui les données sont de plus en plus mesurées en temps continu ou quasi-continu. Cela a conduit au développement de méthodes d'analyse de données fonctionnelles (Ramsay et Silverman (2005)). Dans le domaine des statistiques de scan spatiales, Frévent *et al.* (2021) et Smida *et al.* (2022) ont récemment proposé des méthodes pour données fonctionnelles univariées. Cependant il arrive que plusieurs variables soient conjointement mesurées au cours du temps, ces variables pouvant être de plus corrélées. L'application d'une méthode multivariée sur la moyenne au cours du temps des observations entraînerait une très grande perte d'information. L'approche non paramétrique proposée par Smida *et al.* (2022) peut être adaptée facilement au cadre fonctionnel multivarié, cependant l'approche n'a pas été testée dans ce contexte. De plus aucune méthode paramétrique de statistique de scan spatiale n'existe pour le moment à notre connaissance dans le cas fonctionnel multivarié. Dans ce travail nous allons proposer des adaptations pour données fonctionnelles multivariées aux statistiques de scan proposées par Frévent *et al.* (2021). Nous proposerons également une autre méthode basée sur un test de Wilcoxon-Mann-Whitney. Nous allons ensuite comparer les performances de nos méthodes à l'adaptation pour données fonctionnelles multivariées de l'approche de Smida *et al.* (2022).

Nous décrivons ici la statistique de scan spatial paramétrique pour données fonctionnelles basée sur une MANOVA fonctionnelle, une méthode basée sur un test d'Hotelling et une nouvelle approche basée sur un test de Wilcoxon-Mann-Whitney. La méthodologie est présentée en Section 2. Dans la Section 3, les performances des méthodes sont étudiées et comparées avec celles de l'adaptation au cadre fonctionnel multivarié de la méthode proposée par Smida, *et al.* (2022), et les méthodes sont appliquées sur des données réelles. Enfin, une discussion est présentée en Section 4.

2 Méthodologie

Soit $s_1, \ldots, s_n$, n sites d'un domaine d'observation $S \subset \mathbb{R}^2$ et $X_1, \ldots, X_n$ les observations de X dans $s_1, \ldots, s_n$. Ici, $\{X(t),\ t \in \mathcal{T}\}$ est un processus stochastique à valeurs dans $\mathbb{R}^p, p \geq 2$, avec $\mathcal{T}$ un intervalle de $\mathbb{R}$. A partir de maintenant les observations sont considérées indépendantes, ce qui est une hypothèse classique en statistique de scan.
Le but est de détecter des clusters spatiaux et de tester leur significativité. On cherche donc à tester $\mathcal{H}_0$ (l'absence de cluster) contre une hypothèse alternative $\mathcal{H}_1$ (la présence d'au moins un cluster $w \subset S$ de valeurs anormales pour X).

Cressie, N. (1977) définit une statistique de scan spatial comme le maximum d'un indice

de concentration sur un ensemble de clusters potentiels $\mathcal{W}$. Dans la suite sans perte de généralité, $\mathcal{W}$ est un ensemble de clusters circulaires contenant entre 1 et 50% des sites.

2.1 Statistique de scan paramétrique pour données fonctionnelles multivariées

On suppose que le processus X est à valeurs dans l'espace de Hilbert $L^2(\mathcal{T}, \mathbb{R}^p)$ des fonctions à valeurs dans $\mathbb{R}^p$ et de carré intégrable sur $\mathcal{T}$, équipé du produit scalaire $\langle X, Y \rangle = \int_{\mathcal{T}} X(t)^\top Y(t)\ \mathrm{d}t$.
Frévent *et al.* (2021) a proposé une statistique de scan spatiale pour données fonctionnelles univariées basée sur une ANOVA fonctionnelle. Une version multivariée de l'ANOVA est le test de la trace de Lawley-Hotelling. Górecki, T. et Smaga, Ł. (2017) a adapté ce test pour des processus de $L^2(\mathcal{T}, \mathbb{R}^p)$. Sans perte de généralité, en considérant deux groupes g_1 et g_2 d'observations indépendantes provenant de deux processus X_{g_1} et X_{g_2}, le test compare les fonctions moyennes μ_{g_1} et μ_{g_2} avec $\mu_{g_i}(t) = \mathbb{E}[X_{g_i}(t)]$, $i = 1, 2$.
Alors, pour la détection de clusters, $\mathcal{H}_0$ peut être définie par : $\mathcal{H}_0 : \forall w \in \mathcal{W},\ \mu_w = \mu_{w^\mathsf{c}} = \mu_S$, où μ_w, μ_{w^c} et μ_S sont les fonctions moyennes dans w, à l'extérieur de w et dans S, respectivement. L'hypothèse alternative associée au cluster potentiel w $\mathcal{H}_1^{(w)}$ se réécrit : $\mathcal{H}_1^{(w)} : \mu_w \neq \mu_{w^\mathsf{c}}$. Ensuite le test de la trace de Lawley-Hotelling compare les fonctions moyennes dans w et dans w^c en utilisant la statistique suivante :

$$\mathrm{LH}^{(w)} = \mathrm{Trace}(H_w E_w^{-1}) \tag{1}$$

où $H_w = |w| \int_{\mathcal{T}} [\bar{X}_w(t) - \bar{X}(t)][\bar{X}_w(t) - \bar{X}(t)]^\top\ \mathrm{d}t + |w^\mathsf{c}| \int_{\mathcal{T}} [\bar{X}_{w^\mathsf{c}}(t) - \bar{X}(t)][\bar{X}_{w^\mathsf{c}}(t) - \bar{X}(t)]^\top\ \mathrm{d}t$
et $E_w = \sum_{j, s_j \in w} \int_{\mathcal{T}} [X_j(t) - \bar{X}_w(t)][X_j(t) - \bar{X}_w(t)]^\top \mathrm{d}t + \sum_{j, s_j \in w^\mathsf{c}} \int_{\mathcal{T}} [X_j(t) - \bar{X}_{w^\mathsf{c}}(t)][X_j(t) - \bar{X}_{w^\mathsf{c}}(t)]^\top \mathrm{d}t.$
Alors, $\mathrm{LH}^{(w)}$ peut être considérée comme un indice de concentration et maximisée sur l'ensemble des clusters potentiels $\mathcal{W}$, ce qui amène à la définition suivante de la statistique de scan spatial paramétrique pour données fonctionnelles multivariées (MPFSS) : $\Lambda_{\mathrm{MPFSS}} = \max_{w \in \mathcal{W}} \mathrm{LH}^{(w)}$. Le cluster potentiel pour lequel ce maximum est atteint est appelé "cluster le plus probable" (*most likely cluster* (MLC)) et est donc $\mathrm{MLC} = \arg\max_{w \in \mathcal{W}} \mathrm{LH}^{(w)}$.

2.2 Statistique de scan *distribution-free* pour données fonctionnelles multivariées

Ici nous proposons d'adapter la statistique de scan *distribution-free* pour données fonctionnelles univariées proposée par Frévent *et al.* (2021).

Nous supposons que pour chaque temps t, $\mathbb{V}[X_i(t)] = \Sigma(t,t) \ \ \forall i \in [\![1;n]\!]$. Pour comparer la fonction moyenne entre deux groupes (w et w^{c}) Qiu *et al.* (2021) a proposé d'utiliser la statistique suivante :

$$T_{n,\max}^{(w)} = \sup_{t\in\mathcal{T}} T_n(t)^{(w)} \tag{2}$$

où $T_n(t)$ est une statistique pointwise définie par la statistique de test du T^2 d'Hotelling :

$$T_n(t)^{(w)} = \frac{|w||w^{\mathsf{c}}|}{n}(\bar{X}_w(t) - \bar{X}_{w^{\mathsf{c}}}(t))^\top \hat{\Sigma}(t,t)^{-1}(\bar{X}_w(t) - \bar{X}_{w^{\mathsf{c}}}(t)). \tag{3}$$

où $\hat{\Sigma}(s,t) = \dfrac{1}{n-2}\left[\displaystyle\sum_{i,s_i\in w}(X_i(s) - \bar{X}_w(s))(X_i(t) - \bar{X}_w(t))^\top + \sum_{i,s_i\in w^{\mathsf{c}}}(X_i(s) - \bar{X}_{w^{\mathsf{c}}}(s))(X_i(t) - \bar{X}_{w^{\mathsf{c}}}(t))^\top\right]$.
Pour la détection de clusters, comme pour le MPFSS, $\mathcal{H}_0$ peut être définie par $\mathcal{H}_0 : \forall w \in \mathcal{W},\ \mu_w = \mu_{w^{\mathsf{c}}} = \mu_S$. Et l'hypothèse alternative $\mathcal{H}_1^{(w)}$ associée au cluster potentiel w se réécrit : $\mathcal{H}_1^{(w)} : \mu_w \neq \mu_{w^{\mathsf{c}}}$.

$T_{n,\max}^{(w)}$ peut être considérée comme un indice de concentration et maximisée sur l'ensemble des clusters potentiels $\mathcal{W}$ pour obtenir la statistique de scan spatial *distribution-free* pour données fonctionnelles multivariées (MDFFSS) : $\Lambda_{\text{MDFFSS}} = \max_{w\in\mathcal{W}} T_{n,\max}^{(w)}$. Et pour cette méthode le *most likely cluster* est défini par $\text{MLC} = \arg\max_{w\in\mathcal{W}} T_{n,\max}^{(w)}$.

2.3 Statistique de scan basée sur les rangs pour données fonctionnelles multivariées

Oja, H. et Randles, R.H. (2004) ont développé une extension multivariée au test de Wilcoxon-Mann-Whitney. Nous considérons ici leur statistique de test en version pointwise c'est à dire

$$W(t)^{(w)} = \frac{pn}{\sum_{i=1}^n R_i(t)^\top R_i(t)}\left[\ |w|\ ||\bar{R}_w(t)||_2^2 + |w^{\mathsf{c}}|\ ||\bar{R}_{w^{\mathsf{c}}}(t)||_2^2\ \right] \tag{4}$$

où $R_i(t) = \frac{1}{n}\sum_{j=1}^n \text{sgn}(A_X(t)(X_i(t) - X_j(t)))$ avec

$$\begin{array}{rccl} \text{sgn} & : & \mathbb{R}^p & \to & \mathbb{R}^p \\ & & x & \mapsto & \left\{\begin{array}{ll} ||x||_2^{-1}x & \text{si } x \neq 0 \\ 0 & \text{sinon}\end{array}\right. \end{array}$$

et $A_x(t)$ une matrice de transformation des données telle que $\dfrac{p}{n}\displaystyle\sum_{i=1}^n R_i(t)R_i(t)^\top = \frac{1}{n}\sum_{i=1}^n R_i(t)^\top R_i(t) I_p$.
Ensuite comme précédemment on globalise l'information au cours du temps en considérant une nouvelle quantité :

$$W^{(w)} = \sup_{t\in\mathcal{T}} W(t)^{(w)}. \tag{5}$$

Finalement $W^{(w)}$ est considéré comme un indice de concentration ce qui conduit à la statistique de scan spatiale basée sur les rangs (MRBFSS) :

$$\Lambda_{\text{MRBFSS}} = \max_{w \in \mathcal{W}} W^{(w)}. \tag{6}$$

Et le *most likely cluster* est

$$\text{MLC} = \arg\max_{w \in \mathcal{W}} W^{(w)}. \tag{7}$$

2.4 Significativité du MLC

Une fois le MLC détecté, sa significativité doit être évaluée. Pour cela nous générons des données par permutation des données de départ (*random labelling*) et nous en déduisons une estimation de la p-valeur (Dwass (1957)). Enfin le MLC est considéré significatif si la p-valeur associée est inférieure à l'erreur de type I.

3 Résultats

3.1 Etude d'une simulation

Nous avons comparé (i) la statistique de scan spatial paramétrique pour données fonctionnelles multivariées (MPFSS) Λ_{MPFSS}, (ii) la statistique de scan spatial *distribution-free* pour données fonctionnelles multivariées (MDFFSS) Λ_{MDFFSS}, (iii) la statistique de scan spatial basée sur les rangs pour données fonctionnelles multivariées (MRBFSS) Λ_{MRBFSS}, et (iv) la statistique de scan spatial non paramétrique pour données fonctionnelles (Smida *et al.* (2022)) que nous adaptons au cadre multivariée (NPFSS) Λ_{NPFSS}.

3.1.1 Plan de la simulation

Les sites considérés sont les 94 départements de France métropolitaine, localisés par leur centre de gravité. Un vrai cluster composé des 8 départements d'Île-de-France est défini pour chacun des jeux de données simulés.
Les X_i sont générés avec le modèle suivant ($p = 2$) :

pour $i \in [\![1; 94]\!]$, $X_i(t) = (\sin[2\pi t^2]^5; 1 + 2.3t + 3.4t^2 + 1.5t^3)^\top + \Delta(t)\mathbb{1}_{s_i \in w} + \varepsilon_i(t)$, $t \in [0; 1]$.

où $\varepsilon_i(t) = \sum_{k=1}^{100} Z_{i,k}\sqrt{1.5 \times 0.2^k}\theta_k(t) + b_i(t)$ et $\theta_k(t) = \begin{cases} 1 & \text{si } k = 1 \\ \sqrt{2}\sin[k\pi t] & \text{si } k \text{ pair} \\ \sqrt{2}\cos[(k-1)\pi t] & \text{si } k \text{ impair, } k > 1 \end{cases}$,
et $b_i(t) \sim \mathcal{N}(0, 0.2^2 I_2)$.
Soit $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$ la matrice de covariance des $Z_{i,k}$, trois distributions pour les $Z_{i,k}$ sont considérées : une distribution normale, une distribution de Student et une distribution exponentielle. Trois valeurs de la corrélation ρ sont testées : $\rho = 0.2, 0.5, 0.8$.

Trois types de clusters sont simulés avec $\Delta_1(t) = \alpha(t;t)^\top$, $\Delta_2(t) = \alpha(t(1-t);t(1-t))^\top$ et $\Delta_3(t) = \alpha(\exp[-100(t-0.5)^2]/3;\exp[-100(t-0.5)^2]/3)^\top$, où α est un paramètre contrôlant l'intensité du cluster.
Pour chaque distribution, chaque Δ, chaque valeur de ρ et différentes valeurs de α, nous avons simulé 1000 jeux de données. La p-valeur associée à chaque MLC est estimée en générant 999 permutations des données et nous considérons une erreur de type I de 0.05. Puisque les données sont bruitées nous appliquons un lissage.

3.1.2 Résultats de la simulation

Les performances de toutes les méthodes diminuent légérement quand ρ augmente et le lissage améliore les performances des méthodes pointwise (MRBFSS et MDFFSS).
Les méthodes pointwise présentent de meilleures performances que les deux autres méthodes, surtout dans le cas du shift local Δ_3. Dans le cas gaussien le MDFFSS présente les meilleures performances mais ses performances diminuent quand on s'éloigne de la normalité. Dans ce cas le MRBFSS présente les meilleures performances.

3.2 Application sur données réelles

Nous avons cherché à détecter des clusters spatiaux de pollution de l'air dans le nord de la France avec une résolution spatiale de 2km sur 2km à partir de données quotidiennes sur l'ensemble du mois d'Octobre 2021. Les polluants considérés sont NO_2, O_3, PM_{10} et $PM_{2.5}$.
Les méthodes détectent des clusters significatifs de pollution élevée pour NO_2, PM_{10} et $PM_{2.5}$ à Lille, de faibles concentrations pour tous les polluants dans la zone rurale d'Avesnes-sur-Helpe, et de faibles concentrations pour NO_2 et de fortes concentrations pour O_3 dans la zone maritime de Calais.

4 Discussion

Nous nous sommes ici uniquement focalisés sur le cas de clusters circulaires. Il pourrait être intéressant d'investiguer une fenêtre de scan avec une autre forme. Par exemple une fenêtre de scan elliptique (Kulldorff *et al.* (2006)) peut être intéressante puisqu'une zone de forme allongée le long de la côte avec de fortes concentrations de O_3 semble présente.

Bibliographie

Cressie, N. (1977). On Some Properties of the Scan Statistic on the Circle and the Line, *Journal of Applied Probability*, 14, pp. 272-283.
Cucala, L., Genin, M., Lanier, C. et Occelli, F. (2017). A Multivariate Gaussian scan

statistic for spatial data, *Spatial Statistics*, 21, pp. 66-74.
Dwass, M. (1957). Modified Randomization Tests for Nonparametric Hypotheses, *Annals of Mathematical Statistics*, 28, pp. 181-187.
Frévent, C., Ahmed, M.-S., Marbac, M. et Genin, M. (2021). Detecting spatial clusters in functional data: new scan statistic approaches, *Spatial Statistics*, 46.
Górecki, T. et Smaga, Ł. (2017). Multivariate analysis of variance for functional data, *Journal of Applied Statistics*, 44(12), pp. 2172-2189.
Kulldorff, M. et Nagarwalla, N. (1995). Spatial disease clusters: Detection and inference, *Statistics in Medicine*, 14(8), pp. 799-810.
Kulldorff, M. (1997). A Spatial Scan Statistic, *Communications in Statistics - Theory and Methods*, 26, pp. 1481-1496.
Kulldorff, M., Huang, L., Pickle, L., et Duczmal, L. (2006). An elliptic spatial scan statistic, *Statistics in medicine*, 25(22), pp. 3929-3943. Kulldorff, M. et Huang, L. et Konty, K. (2009). A scan statistic for continuous data based on the normal probability model, *Int J Health Geogr*, 8(58).
Oja, H. et Randles, R. H. (2004). Multivariate nonparametric tests, *Statistical Science*, 18, pp. 598–605.
Qiu, Z., Chen, J. et Zhang, J.-T. (2021). Two-sample tests for multivariate functional data with applications, *Computational Statistics & Data Analysis*, 157.
Ramsay, JO et Silverman, BW (2005). Functional Data Analysis *Springer*.
Smida, Z., Cucala, L., Gannoun, A. et Durif, G. (2022). A Wilcoxon-Mann-Whitney spatial scan statistic for functional data, *Computational Statistics & Data Analysis*, 167.

Annulé

Spatial k-nearest neighbor regression estimate for random fields

Ahmad Younso

MISTEA, Universit Montpellier, INRAE, Institut Agro, Montpellier, France, ahmad.younso@inrae.fr

Résumé. L'estimation classique des k-plus proches voisins (k-NN) de la régression est étudiée avec des observations d'un processus spatialement dépendant satisfaisant une condition générale de α-mélange. La consistance et la vitesse de convergence de l'estimateur sont considérés sous des hypothèses classiques sur le paramètre de lissage k. La consistance de la règle k-NN de classification est établie.

Mots-clés. Champ aléatoire, condition de mélange, régression, estimateur de k-NN, condition de mélange, consistance.

Abstract. The classical k-nearest neighbor (k-NN) regression estimate is investigated with observations from a spatially dependent process satisfying a mild α-mixing condition. Consistency and rate of convergence of the estimator are considered under classical assumptions on the smoothing parameter k. Consistency of the (k-NN) classifier is investigated.

Keywords. Random field, mixing condition, regression , k-NN estimate, mixing condition, consistency.

1 Introduction

Dealing with spatial data in statistics has become important because of various areas of application such as in econometrics, in epidemiology, in environmental science, in geology, in oceanography and many others in which the data of interest are collected across space. In recent years, interest in spatial estimation, prediction and classification using nonparametric methods has increased. Most of the works use kernel method and depend on the mixing concept to handle the spatial data dependency. The literature dealing with the nonparametric k-nearest neighbor (k-NN) method in the

spatial case is not extensive in comparaison with that dealing with kernel method. In the *i.i.d* case, the standard *k-NN* method for estimating density and regression functions or data classification has been widely used. The real interest in the k-nearest neighbor method comes from the nature of the smoothing parameter k which takes its values in a discrete set in addition to the flexibility of the *k-NN* technique to all kinds of heterogeneity in the covariate. In order to introduce the model, consider $\{(X_{\mathbf{i}}, Y_{\mathbf{i}})\}_{\mathbf{i}\in\mathbb{Z}^N}$ a strictly stationary random field defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and taking values in $\mathbb{R}^d \times \mathbb{R}$. A point $\mathbf{i} = (i_1, ..., i_N) \in \mathbb{Z}^N$ will be referred to as a site. For $\mathbf{n} = (n_1, ..., n_N) \in (\mathbb{N}^*)^N$, we define the rectangular region $\mathcal{I}_{\mathbf{n}}$ by $\mathcal{I}_{\mathbf{n}} = \{\mathbf{i} \in \mathbb{Z}^N : \ 1 \leq i_l \leq n_l, \ \forall l = 1, ..., N\}$ and we set $\hat{\mathbf{n}} = n_1 \times ... \times n_N = \text{card}\,(\mathcal{I}_{\mathbf{n}})$ with $\text{card}(A)$ denotes cardinal of the set A. We will write $\mathbf{n} \to \infty$ if $\min_{1\leq l\leq N} n_l \to \infty$. We assume that for each $\mathbf{i} \in \mathbb{Z}^N$, $(X_{\mathbf{i}}, Y_{\mathbf{i}})$ has the same probability distribution as the pair (X, Y) and that the random field is observed on $\mathcal{I}_{\mathbf{n}}$ with $D_{\mathbf{n}} = \{(X_{\mathbf{i}}, Y_{\mathbf{i}}) : \mathbf{i} \in \mathcal{I}_{\mathbf{n}}\}$ denotes the observations set. One wishes to predict the response $Y_{\mathbf{j}}$ of a new observation $X_{\mathbf{j}} = x$ at an unsampled site $\mathbf{j}$. In other word, we want to estimate $\mathbb{E}(Y_{\mathbf{j}}/X_{\mathbf{j}} = x)$ based on the data $D_{\mathbf{n}}$. The problem gets more complicated when $X_{\mathbf{j}}$ is missing. Before introducing this problem, we first define α-mixing random field. Generally, the α-mixing coefficient of the random field $\{(X_{\mathbf{i}}, Y_{\mathbf{i}})\}_{\mathbf{i}\in\mathbb{Z}^N}$ is defined as follows, for any $\mathbf{i} \neq \mathbf{j}$,

$$\alpha(\sigma((X_{\mathbf{i}}, Y_{\mathbf{i}})), \sigma((X_{\mathbf{j}}, Y_{\mathbf{j}}))) = \sup_{A\in\sigma((X_{\mathbf{i}},Y_{\mathbf{i}})), B\in\sigma((X_{\mathbf{j}},Y_{\mathbf{j}}))} |\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)|,$$

with $\sigma((X_{\mathbf{i}}, Y_{\mathbf{i}}))$ $(\sigma((X_{\mathbf{j}}, Y_{\mathbf{j}})))$ denotes the σ-algebra generated by $(X_{\mathbf{i}}, Y_{\mathbf{i}})$ $((X_{\mathbf{j}}, Y_{\mathbf{j}}))$. The random field $\{(X_{\mathbf{i}}, Y_{\mathbf{i}})\}_{\mathbf{i}\in\mathbb{Z}^N}$ is said to be polynomially α-mixing if

$$\alpha(t) = \sup_{\|\mathbf{i}-\mathbf{j}\|\geq t>0} \alpha(\sigma((X_{\mathbf{i}}, Y_{\mathbf{i}})), \sigma((X_{\mathbf{j}}, Y_{\mathbf{j}}))) = O(t^{-\theta}) \text{ for some } \theta > 0 \qquad (1)$$

where $\|.\|$ denotes the Euclidean norm. This mixing condition allows to give more importance to the nearest observations to the site to be predicted. To study the finite sample behavior of the model, we suppose that we want to predict $Y_{\mathbf{j}}$ in the unsampled site $\mathbf{j}$ when $X_{\mathbf{j}}$ is missing. For example, in image restoration, one wants to fill the missing data in a given image. In this case, we assume that the random field is observed only on a subset $\mathcal{S}_{\mathbf{n}} \subset \mathcal{I}_{\mathbf{n}}$. One wishes to predict $Y_{\mathbf{j}}$ in unsampled site $\mathbf{j} \in \mathcal{I}_{\mathbf{n}} - \mathcal{S}_{\mathbf{n}}$ based on observations $X_{\mathbf{i}}$ in a vicinity $v_{\mathbf{j}}$ of $\mathbf{j}$ with $v_{\mathbf{j}}$ is not containing $\mathbf{j}$. Let $v_{\mathbf{j}} = \mathbf{j} + v$, where

$v \subset \mathbb{Z}^N$ is a fixed bounded set not containing $\mathbf{0}$ with $\text{card}(v) = s$ (s is also the cardinal of each $v_{\mathbf{j}}$). We denote $X_{(\mathbf{j})} = (X_{\mathbf{i}} : \mathbf{i} \in v_{\mathbf{j}})$. Clearly, the vector $X_{(\mathbf{j})}$ takes values in $\mathbb{R}^{\tilde{d}}$ with $\tilde{d} = sd$. We set $\mathcal{J}_{\mathbf{n}} = \{\mathbf{i} \in \mathcal{S}_{\mathbf{n}} : v_{\mathbf{i}} \subset \mathcal{S}_{\mathbf{n}}\}$ and we use the data set $D'_{\mathbf{n}} = \{(X_{(\mathbf{i})}, Y_{\mathbf{i}}) : \mathbf{i} \in \mathcal{J}_{\mathbf{n}}\}$ to define the k-NN estimator of the regression function $m(x) = \mathbb{E}\left(Y_{\mathbf{j}}/X_{(\mathbf{j})} = x\right)$ as follows

$$m_{\mathbf{n}}(x) = \sum_{\mathbf{i} \in \mathcal{J}_{\mathbf{n}}} w_{\mathbf{ni}} Y_{\mathbf{i}}, \tag{2}$$

where $w_{\mathbf{ni}} = w_{\mathbf{ni}}(x; D_{\mathbf{n}})$ is $1/k$ if $X_{(\mathbf{i})}$ is one of the k-nearest neighbor of x in $D'_{\mathbf{n}}$ and $w_{\mathbf{ni}}$ is zero otherwise with $k = k(\mathbf{n})$, the smoothing parameter, is a positive integer satisfying the following classical conditions

$$k \longrightarrow \infty \quad \text{and} \quad k/\hat{\mathbf{n}} \longrightarrow 0 \; as \; \mathbf{n} \to \infty. \tag{3}$$

The distance between two elements of $\mathbb{R}^{\tilde{d}}$ or $\mathbb{Z}^N$ will be computed by the Euclidean distance. It is important to mention that the sum in (2) is taken over $\mathcal{J}_{\mathbf{n}}$ instead of $\mathcal{I}_{\mathbf{n}}$ in order to ensure that the sum makes sense. In addition, to make this sum has the same asymptotic behavior as the sum over $\mathcal{I}_{\mathbf{n}}$, it will be assumed that $\mathcal{I}_{\mathbf{n}} - \mathcal{S}_{\mathbf{n}}$ is a finite set. It will be also assumed that the random vectors $X_{(\mathbf{i})}$ has a density f with respect to the Lebesgue measure λ to avoid messy technicalities necessary to handle distance ties. Figure 1 shows three typical examples of the regions defined in the above model for $n_1 = n_2 = 5$ when there is only one unsampled site $\mathbf{j}$.

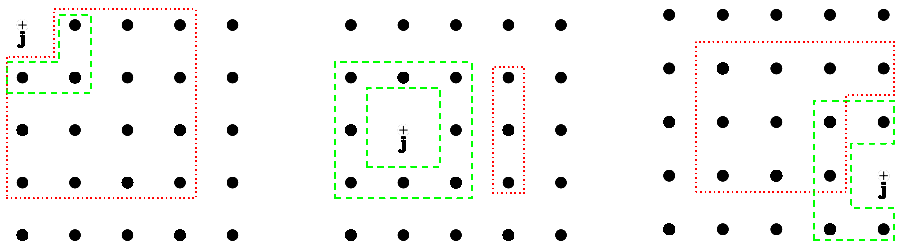


Figure 1: Three regions $\mathcal{I}_{(5,5)}$ with $\mathcal{J}_{(5,5)}$ in red dotted lines, $v_{\mathbf{j}}$ in green dashed lines and $\mathcal{S}_{(5,5)} = \mathcal{I}_{(5,5)} - \{\mathbf{j}\}$.

For example, if $\{Z_{\mathbf{i}}\}_{\mathbf{i} \in \mathbb{N}^N}$ is a strictly stationary real random field observed on $\mathcal{I}_{\mathbf{n}}$ except for a site $\mathbf{j}$ and satisfies the above mixing condition. To predict the value of the random field in the missing site $\mathbf{j}$ with the help of the above

model, it is sufficient to take $Y_{\mathbf{j}} = Z_{\mathbf{j}}$ and $X_{(\mathbf{j})} = (Z_{\mathbf{i}}, \mathbf{i} \in v_{\mathbf{j}})$. In the temporal context ($N = 1$), the model corresponds to a Markov process of order s. However, the optimal smoothing parameter k can be chosen by the cross-validation criterion.

Classification problem. In the binary classification problem, the variable $Y_{\mathbf{i}}$ takes value in $\{0, 1\}$. It is called the label of $X_{\mathbf{i}}$. In the spatial approach, one wants to predict the label at unsampled site $\mathbf{j}$ when $X_{\mathbf{j}}$ is missing. This problem exists in a wide range of fields, including image labelling, computer vision and others. In the present model, the set of observations in the vicinity $v_{\mathbf{j}}$ will be used to create a classifier $g : \mathbb{R}^{\tilde{d}} \longrightarrow \{0, 1\}$ mapping $X_{(\mathbf{j})}$ into the predicted label of $X_{\mathbf{j}}$. Misclassification happens when the associated label of $X_{\mathbf{j}}$ is misspecified. In this case, $g(X_{(\mathbf{j})}) \neq Y_{\mathbf{j}}$ and the probability of error $L(g) = \mathbb{P}\{g(X_{(\mathbf{j})}) \neq Y_{\mathbf{j}}\}$. The best classifier for $X_{(\mathbf{j})} = x$, called Bayes classifier, defined by $g^*(x) = I_{(m(x)>1/2)}$ with I_A is the indicator function of A and $m(x)$ is the regression function defined above. Let $L^* = L(g^*)$ the probability of error of g^*. It is well known that $L^* \leq L(g)$ for all possible classifiers g. In practice, the optimal classifier g^* is virtually always unknown since it depends on the probability distribution of $(X_{(\mathbf{j})}, Y_{\mathbf{j}})$ which is not available. The task of classification is then to use the data $D'_{\mathbf{n}} = \{(X_{(\mathbf{i})}, Y_{\mathbf{i}}) : \mathbf{i} \in \mathcal{J}_{\mathbf{n}}\}$ to construct a classifier $g_{\mathbf{n}}$ that can predict the label of $X_{\mathbf{j}}$ with low error rates. Define the k-NN classifier $g_{\mathbf{n}}(x) = I_{(m_{\mathbf{n}}(x)>1/2)}$ with $m_{\mathbf{n}}(x)$ is the k-NN regression estimate defined in (2). Denote $L_{\mathbf{n}}$ the probability of error of $g_{\mathbf{n}}$. The best we can expect from $g_{\mathbf{n}}$ is to achieve the Bayes error. In the *i.i.d.* case, the consistency of $g_{\mathbf{n}}$ is established (see Devroye et *al.* 1996). The classifier $g_{\mathbf{n}}$ is called consistent if $\mathbb{E}(L_{\mathbf{n}}) \to L^*$ as $\mathbf{n} \to \infty$. In this work, the consistency result of $g_{\mathbf{n}}$ will be extended to the spatial case, see Younso (2017) for the consistency of the spatial kernel classifier.

2 Consistency of spatial k-NN regression estimate and classifier

We are concerned with the L_1 convergence of $m_{\mathbf{n}}$ defined in (2) to m measured by

$$J_{\mathbf{n}} = \int_{\mathbb{R}^{\tilde{d}}} |m_{\mathbf{n}}(x) - m(x)| \, \mu(dx),$$

with μ denote the probability measure of $X_{(\mathbf{i})}$. Devroye et *al.* (1994) establish the almost surely convergence of $J_{\mathbf{n}}$ to 0 in the *i.i.d.* case. In a more general context, we establish some consistency results on $J_{\mathbf{n}}$ for spatially dependent data. More specifically, we investigate the convergence of $J_{\mathbf{n}}$ and the rate of convergence in L_1. In the spatially dependent case, we need to suppose that for each $\mathbf{i} \neq \mathbf{j}$ with $v_{\mathbf{i}} \cap v_{\mathbf{j}} = \phi$, $(X_{(\mathbf{i})}, X_{(\mathbf{j})})$ has a density $f_{\mathbf{i},\mathbf{j}}$ such that

$$\sup_{u,v \in \mathbb{R}^d} |f_{\mathbf{i},\mathbf{j}}(u,v) - f(u)f(v)| \leq C, \text{ for some } C > 0. \tag{4}$$

This condition is used by Carbon et *al.* (2007) in a similar context with the kernel regression estimate.

Theorem 1. *Suppose that the variables* $\mathbb{E}|Y| < \infty$ *and that* $D_{\mathbf{n}}$ *are observations of* α*-mixing random field. Suppose furhermore that (1) and (3)-(4) are satisfied with* $\theta > 2N$*. Then, as* $\mathbf{n} \to \infty$,

$$\mathbb{E}(J_{\mathbf{n}}) \longrightarrow 0.$$

The convergence of Theorem 1 is obtained under classical assumptions on both the mixing coefficient and the smoothing parameter k. In order to investigate the rate of convergence of $\mathbb{E}(J_{\mathbf{n}})$ towards 0, it will be assumed that the following smoothness condition is fulfilled, there exist a μ-integrable function K on $\mathbb{R}^d$ and a constant $\zeta > 0$, such that, for all $r > 0$,

$$\left| m(x) - \frac{1}{\mu(S_{x,r})} \int_{S_{x,r}} m(z)\mu(dz) \right| \leq K(x) r^{\zeta} \tag{5}$$

for μ-almost all $x \in \mathbb{R}^{\tilde{d}}$. Condition (5) is a weak condition used by Györfi (1981) in the *i.i.d.* case. The order of smoothness ζ is similar in spirit to the Holder parameter, but does not require m to be continuous. Györfi (1981) obtains asymptotic rates in terms of ζ. In the following theorem, the rate of convergence for the k-NN regression estimate in the L_1 space is given.

Theorem 2. *Suppose that the variables* $(Y_{\mathbf{i}}, \mathbf{i} \in \mathcal{I}_{\mathbf{n}})$ *are bounded and that* $D_{\mathbf{n}}$ *are observations of* α*-mixing random field. Suppose furthermore that (1) and (3)-(5) are fullfiled with* $\theta > 2N$*. Then,* $\mathbb{E}(J_{\mathbf{n}}) = O\left(1/k + (k/\hat{\mathbf{n}})^{\zeta/d}\right)$. Consequently, for higher order of smoothness, one gets better rates of convergence. This rate is faster than that obtained by Györfi (1981). Observe that if we let $k = (\hat{\mathbf{n}}/k)^{\zeta/d}$, then $k = \hat{\mathbf{n}}^{1/(1+(d/\zeta))}$ and $\mathbb{E}(J_{\mathbf{n}}) = O\left(\hat{\mathbf{n}}^{-1/(1+(d/\zeta))}\right)$.

Theorem 3. *Suppose that the variables* $(Y_{\mathbf{i}}, \mathbf{i} \in \mathcal{I}_{\mathbf{n}})$ *take values in* $\{0,1\}$ *and that* $D_{\mathbf{n}}$ *are observations of* α*-mixing random field. Suppose furhermore that (1) and (3)-(4) are satisfied with* $\theta > 2N$*. Then, the* k*-NN classifier is consistent, i.e.,* $\mathbb{E}(L_{\mathbf{n}}) \longrightarrow L^*$ *as* $\mathbf{n} \to \infty$*. If in addition (5) is satisfied, then,* $\mathbb{E}(L_{\mathbf{n}}) - L^* = O\left(1/k + (k/\hat{\mathbf{n}})^{\zeta/d}\right)$.

Conclusion. The asymptotic results of this work extend, under some general assymptions, previous results ontained in the *i.i.d.* case. If we suppose furtheremore that the random field is β-mixing, the almost surely convergence of both $J_{\mathbf{n}}$ and $L_{\mathbf{n}}$ can be established with the help of the Hoeffding's inequality together with some coupling lemmas. To evaluate the performance of the model, some simulation studies are conducted. The optimal value of the smoothing parameter k is obtained by the cross-validation criterion.

Bibliographie

Biau, G. (2002). Optimal asymptotic quadratic errors of density estimators on random fields.*Statistics & Probability Letters*, 60, 297–307.

Carbon, M., Francq, C. and Tran, L. T. (2007). Kernel regression estimation for random fields. *Journal of Statistical Planning and Inference*, 137, 778–798.

Cheng, P. E. (1984). Strong consistency of nearest neighbor regression function estimators.*Journal of Multivariate Analysis*, 15, 63–72.

Devroye, L. , Györfi, L., Krzyzak, A. and Lugosi, G. (1994). On the strong universal consistency of nearest neighbor regression function estimates. *The Annals of Statistics*, 22, 1371–1385.

Devroye, L., Györfi, L. and Lugosi, G. (1996).*A probabilitic Theory of Pattern Recognition.* Springer-Verlag, New York.

Györfi, L. (1981). The rate of convergence of k-NN regression estimates and classification rules. *IEEE Transactions on Information Theory*, 27, 362–364.

Neaderhouser, C. C. (1980). Convergence of block spins defined on random fields. *J. Statist. Phys.*, 22, 673–684.

Rosenblatt, M. (1985). *Stationary sequences and random fields* . Birkhauser, Boston.

Younso, A. (2017). On the consistency of a new kernel rule for spatially dependent data. *Statistics & Probability Letters*, 131, 64–71.

Estimation de processus de mélanges spatio-temporels pour détecter des changements dynamiques de populations

Solange Pruilh [1,2] & Anne-Sophie Jannot [2]& Stéphanie Allassonnière [2]

[1] *Centre de Mathématiques Appliquées, Ecole polytechnique, Institut Polytechnique de Paris, 91128 Palaiseau Cedex, France et solange.pruilh@polytechnique.edu*

[2] *HeKA, Parisanté Campus, CRC, Université de Paris, INRIA, INSERM, Sorbonne Université et anne-sophie.jannot@aphp.fr, stephanie.allassonniere@inria.fr*

Résumé. Le suivi des populations est un défi dans de nombreux domaines tels que la santé publique ou l'écologie. Nous proposons une méthode pour modéliser et surveiller les distributions de populations dans l'espace et le temps, afin de construire un système d'alerte de changements spatio-temporels dans la répartition des données. Ce pipeline, appelé *spatio-temporal mixture process*, combine un suivi temporel avec des estimations par algorithmes de type Expectation-Maximisation (EM). Nous améliorons les algorithmes existants pour éviter des fausses alertes dues à des maxima locaux et à des données à forte variabilité.

Mots-clés. Modèle de Mélange Gaussien, Algorithme EM, donnée spatio-temporelle

Abstract. Population monitoring is a challenge in many fields such as public health or ecology. We propose a method to model and monitor population distributions in space and time, in order to build an alert system for spatio-temporal changes in the distribution. This pipeline, called *spatio-temporal mixture process*, combines temporal monitoring with estimates by Expectation-Maximization(EM)-type algorithms. We improve existing algorithms to avoid false alarms due to local maxima and high variability data.

Keywords. Gaussian Mixture Model, EM Algorithm, Spatio-temporal data

1 Introduction

La croissance rapide des systèmes d'information de santé a engendré une disponibilité de données permettant le suivi de patients ou de populations affectées par diverses maladies. Un challenge existant est de développer des méthodes qui utilisent ces données pour améliorer les stratégies de santé publique et transformer ces données observées en systèmes d'aide à la décision exploitables. Ces méthodes doivent exploiter la composante spatiale des patients (lieu de naissance, de diagnostique, de résidence, etc), mais aussi la composante temporelle, qui requiert de considérer la distribution de la population comme changeant au cours du temps.

Nous proposons ici un système d'aide à la décision exploitable qui met en évidence des groupes de patients nouveaux ou évoluant. Cela permet par exemple de rapidement identifier une nouvelle source de contamination pour une maladie transmissible, dès l'apparition des premiers cas affectés dans les systèmes d'information de santé.

1.1 Modèles statistiques spatiaux et temporels

Les analyses statistiques spatio-temporelles éprouvées dans la recherche en épidémiologie sont principalement basées sur des tests statistiques, couplés, ou non, avec l'estimation de la densité de noyaux spatio-temporels, comme présenté par Kirby et al. (2017). On y trouve notamment les méthodes de balayage statistique comme celle proposée originellement par Kulldorff (1997). Les balayages statistiques détectent des clusters spatiaux et/ou temporels en utilisant des fenêtres glissantes dans le temps et l'espace. Ces méthodes demandent des efforts de sélection pour trouver un groupe de contrôle adéquat et nécessitent de fixer des paramètres sur la fenêtre glissante considérée.

Les modèles de mélange de lois sont un autre champ permettant de modéliser des données spatiales. Ils présentent des avantages tels qu'en premier lieu la flexibilité car on peut définir la densité de probabilité de chaque cluster en fonction du type d'observations (scalaires, vecteurs, mesures positives, etc.). Deuxièmement, les résultats sont interprétables car les sujets peuvent être attribués à des classes estimées a posteriori, ce qui permet de distinguer des groupes homogènes au sein de la population. Troisièmement, ils ne reposent pas sur l'estimation d'une distribution de référence de la population, contrairement aux méthodes statistiques de balayage. Enfin, ces modèles de mélange sont paramétriques. Le processus que nous présentons par la suite inclut des estimations de modèles de mélange sur des données spatiales.

2 Méthodes

Lorsque les données sont des observations multivariées à valeur réelle, la densité de probabilité habituelle pour chaque groupe est la distribution gaussienne multivariée. Ceci est particulièrement pertinent lorsque l'on considère des données géographiques. Nous supposons pour la suite que la population est modélisée suivant des modèles de mélanges gaussiens (GMMs) décrits ci-dessous.

Soit un jeu de données $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ avec $\mathbf{x}_i \in \mathbb{R}^d$. Soit $\mathcal{N}_d(\cdot|\mu_k, \Sigma_k)$ la densité de la loi gaussienne de dimension d de moyenne μ_k et de matrice de covariance Σ_k. Soient $(z_i)_{i=1,\ldots,n}$, variables latentes telles que chaque z_i suit une distribution catégorielle de paramètre π. Cette information est codée sous la forme d'une variable binaire à K dimensions $\mathbf{z}_i$ pour chaque $i \in \{1, \ldots, n\}$ avec $z_i^k = 1$ si la donnée x_i appartient au cluster k, 0 sinon. Le modèle statistique génératif s'écrit :

$$\begin{cases} z_i & \sim \text{ Categorical}(\pi_1, \ldots, \pi_k)\,, \\ x_i | z_i^k = 1 & \sim \mathcal{N}_d(\mu_k, \Sigma_k)\,. \end{cases} \tag{1}$$

Nous avons ici deux objectifs. On cherche premièrement à estimer le nombre de composants K dans le modèle. Ensuite, sachant K, le second objectif est d'estimer le vecteur des paramètres θ, contenant les paramètres des distributions gaussiennes et les proportions du mélange π, tout cela uniquement à partir des donnés observées.

Nous proposons une amélioration d'un algorithme existant pour répondre à ces deux enjeux conjointement. Cet algorithme est ensuite combiné à un processus complet que nous construisons afin de suivre l'évolution temporelle de la population et de pouvoir proposer un outil de levées d'alertes.

2.1 Estimer des mélanges de lois

Le principe de l'algorithme EM, introduit par Dempster et al. (1977), est de produire une suite de paramètres $(\hat{\theta}^p)_{p\in\mathbb{N}}$ qui converge vers un ensemble de points critiques de la vraisemblance observée du modèle. Ce qui dans le cas du mélange de gaussienne s'écrit :

$$p_\theta(\mathbf{x}) = \prod_{i=1}^{n} \left[\sum_{k=1}^{K} \pi_k \mathcal{N}_d(x_i | \mu_k\, \Sigma_k) \right] . \tag{2}$$

L'algorithme EM alterne entre une étape de calcul d'espérance et une étape de maximisation qui met à jour les paramètres du mélange, jusqu'à ce qu'un critère de convergence soit atteint. La construction de la suite garantit que les points critiques sont des maxima, mais il peut s'agir de maxima globaux ou locaux. D'autres limites apparaissent avec l'EM, telle la dépendance à l'initialisation, mais aussi le choix du nombre de clusters K a-priori ou son estimation a-posteriori par sélection d'un modèle au sein d'une collection finie avec des critères tels que l'AIC de Akaike (1973) ou le BIC de Schwarz (1978). Par la suite, plusieurs méthodes que l'on peut qualifier d'hybrides ou dynamiques ont été proposées afin de s'affranchir d'une collection de modèles en estimant simultanément le nombre optimal de classes et les paramètres du mélange. Les travaux de Yang et al. (2012) font partie de ces nouvelles méthodes. Ils combinent l'exécution d'un algorithme EM unique permettant l'estimation des paramètres mais aussi du nombre de composants, avec une robustesse à l'initialisation et un temps de calcul raisonnablement faible. Mais cet algorithme peut toujours atteindre des maxima locaux incorrects, impliquant des clusters superposés, avec donc des paramètres identiques, conduisant à l'arrêt prématuré de l'algorithme. Yang et al. (2012) introduisent une régularisation sur les proportions du mélange dans la fonction objectif, impliquant l'entropie de ces proportions. Partant de la log-vraisemblance complète, la fonction objectif à maximiser devient alors:

$$\mathcal{L}'(\theta, \mathbf{x}, \mathbf{z}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_i^k \log(\pi_k \mathcal{N}(x_i | \mu_k, \Sigma_k)) + \beta \sum_{i=1}^{n} \sum_{k=1}^{K} \pi_k \log \pi_k, \text{avec } \beta \geq 0\,. \tag{3}$$

Les composantes de proportions très faibles sont éliminées à chaque itération, ce qui dynamiquement diminue K. Des expériences sur données réelles nous ont montré que cet algorithme appelé Robust EM produit parfois des clusters superposés, conduisant à des solutions dégénérées.

Afin de résoudre ce problème, nous rajoutons dans l'algorithme Robust EM une étape de détection de ces clusters jumeaux, ainsi qu'une gestion plus automatique du critère d'arrêt pour convergence. Si l'algorithme n'arrive toujours pas à éviter ces clusters superposés, alors nous forçons la fusion de ces derniers. Nous nommons cette nouvelle version Modified Robust EM.

De nombreux ensembles de données de santé publique présentant une forte variabilité spatiale, ceci entraîne des estimations incorrectes avec des matrices de covariances singulières. Nous incluons une régularisation des matrices de covariance au sein de notre Modified Robust EM, similaire à celle présente dans Fraley and Raftery (2007) qui repose sur un prior d'Inverse-Wishart. Cette régularisation permet d'éviter des dégénérescences dans l'estimation des matrices de covariances des composants.

Nous proposons aussi un Constrained EM (C-EM), qui repose sur l'introduction de contraintes dans l'estimation des paramètres des distributions gaussiennes. Ces nouveaux paramètres doivent rester dans un voisinage de paramètres $\theta^{(old)}$. Dans notre cas $\theta^{(old)} = \theta^{(t-1)}$, paramètres d'un mélange estimé à l'instant $t-1$. Avec ce modèle contraint nous gardons une trace indirecte du modèle estimé au temps précédent tout en l'adaptant aux nouvelles observations. Cette adaptation permet de prendre en compte le fait que chaque modèle n'est estimé que sur un échantillon parfois de petite taille et donc sujet à variabilité. Maintenant que nous avons un outil d'estimation automatique de la distribution multimodale de la population, il s'agit de l'utiliser pour regarder si celle-ci évolue au cours du temps.

2.2 Un pipeline spatio-temporel

Nous considérons un modèle à temps discret indexé par $t = 1, \ldots, T$. A chaque pas de temps, notons le vecteur de données $\mathbf{X}^{(t)} = (\mathbf{X}_1^{(t)}, \ldots, \mathbf{X}_{n_t}^{(t)})$ avec $\mathbf{X}_i^{(t)} \in \mathbb{R}^d$. Nous supposons que ces données sont échantillonnées suivant un modèle statistique dépendant du temps. Nous modélisons les données à chaque temps t par un mélange de distributions de probabilité, paramétré par un vecteur $\theta^{(t)}$, définissant le modèle courant $M^{(t)}$.

Nous détaillons ici notre pipeline complet, nommé *spatio-temporal mixture process* (STMP). A chaque temps t se présente un nouveau vecteur $\mathbf{X}^{(t)}$, indépendant du précédent $\mathbf{X}^{(t-1)}$. Étant donné ce nouvel échantillon, nous voulons évaluer si le modèle précédent $M^{(t-1)}$, défini comme un modèle de mélange estimé sur $\mathbf{X}^{(t-1)}$, est susceptible de s'adapter au nouvel ensemble $\mathbf{X}^{(t)}$. Nous proposons d'estimer deux modèles à chaque temps t :

- un modèle contraint (ou modèle candidat) M' pour approcher $\mathbf{X}^{(t)}$ où M' est un ajustement de $M^{(t-1)}$, obtenu par l'estimation de θ' proche de $\theta^{(t-1)}$.

- un modèle alternatif, M^a avec pour paramètres θ^a, estimé uniquement à partir de l'ensemble de données $\mathbf{X}^{(t)}$.

Avec ces deux modèles estimés en main, nous sommes maintenant en mesure de suivre les changements de la distribution de la population, et de déterminer s'il y a une modification de la répartition géographique de la population. Le système d'alerte que nous proposons est défini comme suit. Si au temps t, le modèle M', proche de $M^{(t-1)}$, n'est pas adapté pour décrire $\mathbf{X}^{(t)}$, on propose le modèle indépendant M^a comme nouvelle description de la population actuelle et on lance une alerte. L'objectif est maintenant de définir la règle de décision permettant de sélectionner l'un ou l'autre des modèles et de déclencher ou non l'alerte en conséquence.

Une façon simple de quantifier la qualité de l'ajustement d'un modèle statistique aux données est sa vraisemblance. La règle de décision est donnée par le rapport de vraisemblance ou facteur de Bayes, défini à partir des vraisemblances des modèles de mélange estimés M' et M^a, données respectivement par $p_{\theta'}(\mathbf{X}^{(t)})$ et $p_{\theta^a}(\mathbf{X}^{(t)})$:

$$r_t(M', M^a) = \frac{p_{\theta^a}(\mathbf{X}^{(t)})}{p_{\theta'}(\mathbf{X}^{(t)})}\,. \tag{4}$$

Le modèle alternatif étant non contraint, $p_{\theta^a}(\mathbf{X}^{(t)})$ est la valeur maximale de la vraisemblance des données sans hypothèse. $p_{\theta'}(\mathbf{X}^{(t)})$ est la valeur maximale de la vraisemblance lorsque les paramètres θ' sont restreints pour rester dans un voisinage de $\theta^{(t-1)}$. Dans le cas où le modèle contraint M' s'ajuste bien à $\mathbf{X}^{(t)}$, le rapport de vraisemblance sera proche de 1 car les vraisemblance devraient être similaires. En revanche, si $\mathbf{X}^{(t)}$ est échantillonné à partir d'une distribution très différente de celle de $M^{(t-1)}$, alors M' aura une vraisemblance inférieure à celle de M^a qui, par définition, sera capable de mieux s'adapter au nouvel ensemble de points et il devrait y avoir une notification lorsque ce rapport est largement supérieur à un.

Afin d'accepter ou de rejeter le modèle alternatif au temps t, nous définissons un seuil τ tel que si $r_t(M', M^a) \geq \tau$, le modèle alternatif est sélectionné et une alerte est émise.

Notre pipeline combine le Modified Robust EM présenté précédemment pour estimer M^a, avec le Constrained EM pour estimer M'. Notons que ce pipeline offre une grande flexibilité sur les distributions choisies dans le modèle de mélange et sur les algorithmes d'estimation utilisés dans la première étape.

3 Applications

Nous comparons notre Modified Robust EM à d'autres algorithmes existants notamment celui de Yang et al. (2012). Avec des performances similaires au Robust EM, le Modified Robust EM règle le problème des clusters superposés et permet l'estimation de données à forte variabilité, grâce aux régularisations sur les matrices de covariances. Des expériences

sur des données simulées permettent de fixer le seuil d'alerte τ, interne à notre STMP. Finalement nous présentons des résultats sur des données réelles de santé publique en particulier sur le suivi de la première vague COVID19 à Paris.

4 Conclusion

Notre application sur ces données de santé publique montre que le STMP modélise correctement les distributions de population et lève des alertes compréhensibles et interprétables. Notre pipeline complet et générique permet de modéliser l'évolution d'une distribution de populations, et de détecter des changements anormaux dans cette distribution. Ce STMP est combiné avec de nouvelles variantes de l'EM, qui nous permettent d'éviter plusieurs écueils de l'EM original comme l'initialisation, l'estimation de K et l'estimation sur données à forte dispersion.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle of the Maximum Likelihood Principle. In *2nd International Symposium on Information Theory*, pp. 267–281. Akadémiai Kiadó.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological) 39*(1), 1–22.

Fraley, C. and A. E. Raftery (2007). Bayesian Regularization for Normal Mixture Estimation and Model-Based Clustering. *Journal of Classification*, 155–181.

Kirby, R. S., E. Delmelle, and J. M. Eberth (2017). Advances in spatial epidemiology and geographic information systems. *Annals of epidemiology 27*(1), 1–9.

Kulldorff, M. (1997, January). A spatial scan statistic. *Communications in Statistics - Theory and Methods 26*(6), 1481–1496.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics 6*(2), 461–464.

Yang, M.-S., C.-Y. Lai, and C.-Y. Lin (2012, November). A robust EM clustering algorithm for Gaussian mixture models. *Pattern Recognition 45*(11), 3950–3961.

Modélisation spatiale des accidents routiers de Besançon (France) à l'aide de processus de Cox log-Gaussien

Cécile Spychala [a,1], Camelia Goga [a,2] & Clément Dombry [a,3]

[a] *Laboratoire de Mathématiques de Besançon, 16 route de Gray, 25000, Besançon*
[1] *cecile.spychala@univ-fcomte.fr*
[2] *camelia.goga@univ-fcomte.fr*
[3] *clement.dombry@univ-fcomte.fr*

Résumé. Dans un objectif de prévention et/ou d'anticipation des accidents routiers, la modélisation statistique de la dépendance spatiale et des facteurs de risque potentiels représente un atout majeur. L'intérêt de cette étude se porte sur la localisation géo-référencée des accidents. Nous avons croisé ces événements avec des covariables caractérisant la zone géographique d'étude (socio-démographiques et infrastructures par exemple). Après une sélection de variables (aggrégation de modèles de Poisson), la survenue des accidents a été modélisée par un processus de Cox log-Gaussien spatial. Les résultats de cette analyse permettent l'identification des principaux facteurs de risques d'accident et l'identification des zones critiques. Les données mises en application sont les accidents routiers s'étant produits entre 2017 et 2019 dans la CAGB (communauté urbaine de Besançon).

Mots-clés. accidents routiers, données spatiales, processus de Cox log-Gaussien

Abstract. In order to prevent and/or forecast road accidents, the statistical modelling of spatial dependence and potential risk factors is a major asset. The focus of this study is on the georeferenced location of accidents. We crossed these events with covariates characterizing the study geographical area such as sociodemographic and infrastructure measures. After a variable selection (Poisson models aggregation), the occurrence of accidents was modelled by using a spatial log-Gaussian Cox process. The results of this analysis enable us to identify principal risk factors of road accidents and critical areas. The data used are road accidents that occurred between 2017 and 2019 in the CAGB (urban community of Besançon).

Keywords. road accidents, spatial data, point pattern, point process, log-Gaussian Cox process

1 Introduction

Les accidents de la route étant l'une des principales causes de décès dans le monde font l'objet de vives inquiétudes. Les analyses statistiques des données sur les accidents de

la route sont des outils majeurs pour les forces de l'ordre qui peuvent agir pour prévenir et anticiper ces événements. Différentes méthodes statistiques peuvent être appliquées selon différents objectifs d'étude. L'objectif de cette étude ici est d'identifier les zones de la CAGB (Communauté d'Agglomération du Grand Besançon) où la probabilité qu'un accident survienne est élevée. La prise en compte des caractéristiques environnementales de ces accidents apparaît donc essentielle. Les données d'accidents de la route de cette étude sont des points géo-référencés, appelés configuration de points (ou *point pattern* en anglais), et la principale motivation est de comprendre comment ces points sont disposés dans l'espace. Les point pattern sont générés par un mécanisme aléatoire sous-jacent, qu'on appelle processus ponctuel, et c'est en réalité ce processus qui fait objet d'intérêt.

La localisation des accidents de la route étant au cœur du sujet, ces événements ont été modélisés par un processus ponctuel spatial nommé processus de Cox log-Gaussien. De plus, afin d'identifier les facteurs de risque potentiels influençant la survenue des accidents, des ressources environnementales ont été apportées en tant que covariables dans le modèle.

2 Matériels et méthodes

2.1 Étude de cas: accidents routiers

L'étude porte sur les accidents corporels de la route survenus entre 2017 et 2019 dans la CAGB (*Communauté d'Agglomération du Grand Besançon*), siège de la région Franche-Comté. Cette communauté urbaine est composée de 68 communes dont Besançon qui est la commune centrale. Ces accidents de la route sont géo-référencés et représentent donc notre point pattern à analyser.

2.2 Processus de Cox log-Gaussien

Une donnée sous la forme d'un ensemble de points $\mathbf{x} = \{x_1, \ldots, x_n\}$ de $\mathbb{R}^2$, répartis irrégulièrement sur une zone d'étude, est appelée un *spatial point pattern* (configuration spatiale de point) et fait référence aux localisations en tant qu'événements (Diggle, 2013). Cet ensemble de points est considéré comme le résultat de la réalisation d'un mécanisme stochastique sous-jacent appelé processus ponctuel. Comme l'analyse ne porte pas sur les points eux-mêmes, mais sur la façon dont les points ont été générés, l'ajustement d'un processus ponctuel spatial est généralement l'approche adoptée pour modéliser une telle situation. Le traitement des processus ponctuels repose sur l'étude de l'intensité du processus ponctuel et de l'interaction entre les points. L'analyse de l'intensité est généralement la plus importante, elle représente le nombre moyen d'événements par unité de surface ou autrement dit, le taux d'occurence. Ensuite, l'interaction entre les points consiste à déterminer si les points semblent avoir été placés indépendamment les uns des autres ou s'ils présentent une sorte de dépendance.

Nous proposons dans cette étude de modéliser notre point pattern par un processus de Cox et plus particulièrement par un processus de Cox log-Gaussien (*Log-Gaussian Cox Processes* en anglais, LGCP). Un LGCP est un processus de Cox dont la log-intensité est modélisée par un processus gaussien latent. Pour plus de détails, voir par exemple Moller et al. (1998), Diggle et al. (2013) ou Taylor et al. (2015). En pratique, l'ajustement s'effectue sur une grille de calcul (considérée comme régulière). Le LGCP spatial est noté comme suit :

$$\begin{aligned} X(u) &\sim Poisson(\Lambda(u)) \\ \Lambda(u) &= C_A \exp(Z(u)\beta + Y(u)) \end{aligned}$$

où $X(u)$ désigne le nombre d'événements dans la cellule de la grille de calcul contenant $u \in \mathbb{R}^2$, C_A la superficie de la cellule, $Z(u)$ vecteur de covariables avec les effets associés β et $Y = \{Y(u) : u \in \mathbb{R}^2\}$ est un processus Gaussien (avec une variance $\sigma^2 > 0$ et une fonction de corrélation exponentielle $r(t) = \exp(-t/\phi)$ avec $\phi > 0$ un paramètre d'échelle).

L'inférence bayésienne pour les paramètres Y, β, σ et ϕ est effectuée à l'aide des méthodes de Monte Carlo par chaînes de Markov. Le lecteur peut trouver une description complète dans Robert (1996), Robert and Casella (2004) ou Taylor et al. (2015).

2.3 Caractéristiques environnementales des accidents

Comme vu précédemment, de l'information par le biais de covariables peut être incluse dans la définition de l'intensité du LGCP. Ainsi, des ressources environnementales provenant de sources de données telles que l'INSEE (*Institut National de la Statistique et des Etudes Economiques*), data.gouv ou OpenStreetMap ont été collectées. Ces données ont été fusionnées en un seul support de données spatiales, correspondant à une grille, par le biais de méthodes d'interpolation.

Les covariables peuvent être classées en trois groupes : socio-démographiques, infrastructures routières et infrastructures globales. Dans un premier temps, les composantes socio-démographiques utilisées sont la proportion des personnes âgées de 18 à 24 ans et la proportion des personnes âgées de plus de 65 ans. Dans un second temps quatre covariables liées aux infrastructures routières ont été construites : le nombre d'intersection, le nombre de radars, la longueur totale des routes municipales et la longueur totale des routes nationales (qui englobe routes départementales, nationales et autoroutes). Enfin, sept covariables constituent l'aspect "infrastructures globales" : le nombre d'établissements de santé, le nombre d'écoles, le nombre d'écoles supérieures, le nombre d'établissements de loisirs, le nombre de stations services, le nombre de gares ferroviaires (ou station de taxi) et le nombre de commerces (alimentaires et non alimentaires). Toutes ces informations sont données par cellule de la grille, support sur lequel elles ont été interpolées. Le nombre de covariables est ainsi de 13.

Certaines de ces informations ont été choisies car elles montraient des caractéristiques pertinentes dans les études d'accident comme par exemple celles de Ramírez and Valencia (2021) et Spychala et al. (2021). D'autres ont été prises car elles pourraient hypothétiquement montrer des résultats significatifs.

2.4 Sélection de variables

Le nombre de covariables s'avère être assez important au vue du modèle de processus ponctuel LGCP qui représente un challenge en terme de computation. En effet, comme mentionné précédemment, des méthodes bayésiennes sont associées à des méthodes de Monte Carlo par chaînes de Markov afin d'estimer les paramètres du modèle. Une sélection de variables a alors été proposée afin de réduire le nombre de covariables à introduire dans le modèle. L'apprentissage supervisé est utilisé afin d'analyser les dépendances entre la variable de réponse, qui donne le nombre d'accidents, et les covariables mentionnées précédemment. Une méthode d'aggrégation de modèles de Poisson fut employée, inspirée de la méthode Random Forest.

3 Résultats

Après ajustement du modèle LGCP, nous pouvons interpréter les résultats et ainsi répondre aux objectifs fixés. Pour quantifier les risques des facteurs potentiels, l'accent est mis sur les effets des covariables. Les estimations des paramètres β permettent de conclure sur l'effet significatif ou non significatif d'une covariable en question et ainsi donc d'identifier les facteurs les plus dangereux.

Enfin, afin d'identifier les zones géographiques critiques de plus ou moins grande occurrence, les quantités suivantes $\mathbb{P}(\exp(Y) > k|X)$ et $\mathbb{P}(\exp(Z\beta + Y) > k|X)$ sont tracées sur la fenêtre d'étude (CAGB) de manière similaire à Taylor et al. (2015). Les cartes obtenues représentent respectivement le risque relatif (à un seuil k) du processus gaussien latent et la probabilité que le nombre d'accident soit supérieur à k.

References

Diggle, P. (2013). *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns*. Chapman & Hall/CRC, third edition.

Diggle, P., Moraga, P., Rowlingson, B., and Taylor, B. (2013). Spatial and spatio-temporal log-gaussian cox processes: Extending the geostatistical paradigm. *Statistical Science*, 28(4):105848.

Moller, J., Syversveen, A., and Waagepetersen, R. (1998). Log gaussian cox processes. *Scandinavian Journal of Statistics*, 25(3):451–482.

Ramírez, A. and Valencia, C. (2021). Spatiotemporal correlation study of traffic accidents with fatalities and injuries in bogota (colombia). *Accident Analysis & Prevention*, 149:105848.

Robert, C. (1996). *Méthodes de Monte Carlo par chaînes de Markov*. Economica.

Robert, C. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer.

Spychala, C., Armand, J., Dombry, C., and Goga, C. (2021). Multivariate statistical analysis for exploring road crash-related factors in the franche-comté region of france. *Communications in Statistics: Case Studies, Data Analysis and Applications*, 7(3):442–474.

Taylor, B. M., Davies, T. M., Rowlingson, B. S., and Diggle, P. J. (2015). Bayesian inference and data augmentation schemes for spatial, spatiotemporal and multivariate log-gaussian cox processes in r. *Journal of Statistical Software*, 63(7).

Statistique bayésienne

Sélection de variables bayésienne en grande dimension dans les modèles non-linéaires à effets mixtes utilisant l'algorithme SAEM

Marion Naveau [1,2] & Guillaume Kon Kam King [2] & Laure Sansonnet [3] & Maud Delattre [2]

[1] *Université Paris-Saclay, INRAE, UMR MIA-Paris, 75005, Paris, France - E-mail : marion.naveau@inrae.fr*

[2] *Université Paris-Saclay, INRAE, UMR MaIAGE, 78350, Jouy-en-Josas, France E-mail : guillaume.kon-kam-king@inrae.fr, maud.delattre@inrae.fr*

[3] *Université Paris-Saclay, INRAE, AgroParisTech, UMR MIA-Paris, 75005, Paris, France - E-mail : laure.sansonnet@agroparistech.fr*

Résumé. Les données de grande dimension, avec beaucoup plus de covariables que d'observations, comme les données génomiques par exemple, sont maintenant couramment analysées. Dans ce contexte, il est souvent souhaitable de pouvoir se concentrer sur les quelques covariables les plus pertinentes grâce à une procédure de sélection de variables. La question de la sélection de variables en grande dimension est largement documentée dans les modèles de régression standard, mais il existe encore peu d'outils pour y répondre dans le cadre des modèles à effets mixtes non linéaires. Dans ce travail, nous abordons la sélection de variables sous un angle bayésien et proposons une procédure de sélection combinant l'utilisation de priors *spike-and-slab* et l'algorithme SAEM. Comme pour la régression Lasso, l'ensemble des covariables pertinentes est sélectionné en explorant une grille de valeurs pour le paramètre de pénalisation. L'approche proposée est plus rapide qu'un algorithme MCMC classique et montre de très bonnes performances de sélection sur des données simulées.

Mots-clés. Sélection de variables bayésienne, modèles non linéaires à effets mixtes, données de grande dimension, *spike-and-slab*, algorithme SAEM.

Abstract. High-dimensional data, with many more covariates than observations, such as genomic data for example, are now commonly analyzed. In this context, it is often desirable to be able to focus on the few most relevant covariates through a variable selection procedure. High dimensional variable selection is widely documented in standard regression models, but there are still few tools to address it in the context of nonlinear mixed effects models. In this work, we approach variable selection from a Bayesian perspective and propose a selection procedure combining the use of *spike-and-slab* priors and the SAEM algorithm. Similarly to Lasso regression, the set of relevant covariates is selected by exploring a grid of values for the penalization parameter. The proposed approach is much faster than a classical MCMC algorithm and shows very good selection performances on simulated data.

Keywords. Bayesian variable selection, nonlinear mixed effects models, high-dimensional data, *spike-and-slab*, SAEM algorithm.

1 Introduction

1.1 Modèle statistique

On se place dans le contexte général d'observations collectées de façon répétée sur plusieurs individus au sein d'une population d'intérêt. On note n le nombre d'individus et n_i le nombre d'observations pour l'individu i. Sans perte de généralité et simplement pour alléger les notations, nous supposons que $n_i = J$ pour tout $i \in \{1, \dots n\}$. Les modèles non linéaires à effets mixtes (NLMEM) ont principalement été introduits pour modéliser les réponses d'individus ayant le même comportement global mais avec des variations individuelles ([Lavielle, 2014]). Pour tout $i \in \{1, \dots n\}$ et $j \in \{1, \dots J\}$, la réponse de l'individu i au temps t_{ij}, que l'on note y_{ij}, est modélisée de la façon suivante :

$$\begin{cases} y_{ij} = g(\varphi_i, t_{ij}) + \varepsilon_{ij} & , \varepsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2) \quad (1a) \\ \varphi_i = \mu + {}^t\beta V_i + \xi_i & , \xi_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \Gamma^2) \quad (1b) \end{cases}$$

où g est une fonction connue qui dépend de manière non linéaire d'un paramètre individuel φ_i non-observé. Dans ce modèle, on suppose donc que toutes les réponses suivent une même fonctionnelle g mais dépendant de paramètres variables entre individus. Cette variabilité inter-individuelle est décrite au moyen de l'équation (1b) qui sépare, d'une part, la variabilité qui peut être expliquée par des covariables $V_i = {}^t(V_{i1}, \dots, V_{ip}) \in \mathbb{R}^p$, et d'autre part, la variabilité aléatoire, non attribuable aux covariables mesurées dans la population, elle-même décrite par l'effet aléatoire ξ_i.

Dans la suite, nous considérons les situations où $\varphi_i \in \mathbb{R}$. Nous notons $\theta = (\mu, \beta, \Gamma^2, \sigma^2)$ le paramètre de population, inconnu, $y = ((y_{ij})_{1 \le j \le J})_{1 \le i \le n}$ l'ensemble des observations disponibles et $\varphi = (\varphi_i)_{1 \le i \le n}$ l'ensemble des paramètres individuels.

1.2 Objectif et formulation bayésienne

L'objectif principal de ce travail est de sélectionner les covariables pertinentes, *i.e.* celles qui décrivent le mieux la variabilité entre les individus. Il s'agit donc d'estimer le support du vecteur d'effets fixes $\beta \in \mathbb{R}^p$, noté S^*_β :

$$S^*_\beta = \left\{ \ell \in \{1, \dots, p\} \middle| \beta_\ell \neq 0 \right\}$$

La difficulté ici est que le modèle (1) est un modèle à données incomplètes. En effet, bien que la première couche (1a) soit observée, ce n'est pas le cas pour les paramètres individuels $(\varphi_i)_i$. L'inférence est donc difficile car la vraisemblance et les estimateurs classiques n'ont pas de forme explicite.

L'estimation du paramètre de population θ dans les NLMEM a fait l'objet de nombreux travaux. Une solution très répandue est d'utiliser l'algorithme EM (*Expectation-Maximization*, [Dempster *et al.*, 1977]), ou une de ses variantes, comme l'algorithme SAEM

(*Stochastic Approximation EM*, [Delyon *et al.*, 1999]), pour calculer l'estimateur du maximum de vraisemblance ou l'estimateur du maximum *a posteriori*.

La sélection de variables dans les modèles à effets mixtes est elle aussi très étudiée. Les outils proposés sont très différents en petite dimension, *i.e.* $p < n$, et en grande dimension, *i.e.* $p >> n$, mais aussi selon que la fonction g est linéaire ou non linéaire en φ_i (voir par exemple [Müller *et al.*, 2013], [Delattre *et al.*, 2014], [Schelldorfer *et al.*, 2011], [Schelldorfer *et al.*, 2014]). Plus précisément, supposer g linéaire en φ_i dans (1) permet de développer des critères dont le calcul et/ou l'étude théorique font appel à des quantités explicites, ce qui est rarement le cas lorsque g est non linéaire en φ_i. En conséquence, la sélection de variables en grande dimension est peu traitée dans le cadre des NLMEM ([Bertrand et Balding, 2013], [Ollier, 2021]).

Dans ce travail, nous développons une nouvelle procédure de sélection de covariables en grande dimension dans les NLMEM. En s'inspirant des travaux de [Ročková et George, 2014] pour les modèles linéaires gaussiens, nous avons adopté une approche bayésienne utilisant un prior *spike-and-slab* pour le vecteur de régression β ([George et McCulloch, 1993]).

Pour cela, nous introduisons un vecteur binaire de variables latentes, $\delta = (\delta_\ell)_{1\leq\ell\leq p}$, correspondant au support de β, tel que :

$$\forall 1 \leq \ell \leq p \ , \ \delta_\ell = \begin{cases} 1 & \text{si la covariable } \ell \text{ est à inclure dans le modèle,} \\ 0 & \text{sinon.} \end{cases}$$

On peut remarquer qu'identifier les coefficients non nuls de β revient à identifier les composantes de δ égales à 1. Les priors bayésiens utilisés dans ce travail sont les suivants :

$$\begin{cases} \pi(\beta|\delta) & = \mathcal{N}_p(0, D_\delta), \text{ où } D_\delta = \text{diag}(a_1, \ldots, a_p),\ a_\ell = (1-\delta_\ell)\nu_0 + \delta_\ell\nu_1 \ , \ 0 < \nu_0 < \nu_1. & \text{(2a)} \\ \pi(\mu) & = \mathcal{N}(0, \sigma_\mu^2), \text{ avec } \sigma_\mu^2 > 0. & \text{(2b)} \\ \pi(\sigma^2) & = \mathcal{IG}\left(\frac{\nu_\sigma}{2}, \frac{\nu_\sigma\lambda_\sigma}{2}\right), \text{ avec } \nu_\sigma, \lambda_\sigma > 0. & \text{(2c)} \\ \pi(\Gamma^2) & = \mathcal{IG}\left(\frac{\nu_\Gamma}{2}, \frac{\nu_\Gamma\lambda_\Gamma}{2}\right), \text{ avec } \nu_\Gamma, \lambda_\Gamma > 0. & \text{(2d)} \\ \pi(\delta|\alpha) & = \alpha^{|\delta|}(1-\alpha)^{p-|\delta|}, \text{ avec } \alpha \in [0,1] \text{ et } |\delta| = \sum_{\ell=1}^p \delta_\ell. & \text{(2e)} \\ \pi(\alpha) & = \mathit{Beta}(a, b), \text{ avec } a, b > 0. & \text{(2f)} \end{cases}$$

où $(\nu_0, \nu_1, \sigma_\mu^2, \nu_\sigma, \lambda_\sigma, \nu_\Gamma, \lambda_\Gamma, a, b)$ sont des hyperparamètres fixés. Le prior fondamental est le prior *spike-and-slab* (2a) qui va induire la parcimonie dans le vecteur β. Une recommandation générale pour ce type de prior est de fixer l'hyperparamètre ν_0 petit et l'hyperparamètre ν_1 grand pour favoriser respectivement l'exclusion des effets non significatifs et l'inclusion des covariables influentes. Le choix des valeurs de ces hyperparamètres reste néanmoins difficile en pratique.

2 Procédure de sélection de variables

Nous proposons une procédure en deux étapes.

1. La première étape consiste à réduire le nombre de modèles candidats. De manière similaire à la régression Lasso, notre procédure explore une grille de valeurs pour la variance de la distribution spike ν_0 plutôt que de se concentrer sur une seule valeur. En effet, cet hyperparamètre contrôle le support de β, qui n'est pas connu. Ainsi, cette grille permet d'étudier différents niveaux de parcimonie dans le vecteur β. Notons Δ cette grille. Pour chaque $\nu_0 \in \Delta$, l'algorithme MCMC-SAEM, introduit par [Kuhn et Lavielle, 2004], est utilisé pour obtenir l'estimation par maximum *a posteriori* de $\Theta = (\theta, \alpha)$, notée $\widehat{\Theta}_{\nu_0}^{MAP}$. Celle-ci est ensuite utilisée pour obtenir une estimation $\widehat{\delta}$ des variables indicatrices δ en calculant la probabilité d'inclusion *a posteriori* de chaque covariable ℓ :

$$\widehat{\delta}_\ell = 1 \Longleftrightarrow \mathbb{P}(\delta_\ell = 1 | y, \widehat{\Theta}_{\nu_0}^{MAP}) \geq 0.5$$

On en déduit alors un sous-ensemble de covariables pertinentes :

$$\widehat{S}_{\nu_0} = \left\{ \ell \in \{1, \ldots, p\} \;\middle|\; |(\widehat{\beta}_{\nu_0}^{MAP})_\ell| \geq s_\beta(\nu_0, \nu_1, \widehat{\alpha}_{\nu_0}^{MAP}) \right\}$$

où le seuil $s_\beta(\nu_0, \nu_1, \widehat{\alpha}_{\nu_0}^{MAP})$ est explicite et indépendant de la covariable ℓ considérée. Cette première étape réduit la collection de modèles à une collection de sous-modèles prometteurs $(\widehat{S}_{\nu_0})_{\nu_0 \in \Delta}$.

2. Ensuite, la deuxième étape consiste à sélectionner le "meilleur" modèle parmi ceux conservés à la fin de la première étape. On utilise une extension du BIC, le critère eBIC ([Chen et Chen, 2008]), dont la pénalité permet de tenir compte du fait que, comme $p >> n$, le nombre de modèles d'une certaine taille augmente très vite avec la taille considérée. Finalement le modèle sélectionné est le modèle $\widehat{S}_{\hat{\nu}_0}$ où :

$$\hat{\nu}_0 = \underset{\nu_0 \in \Delta}{\operatorname{argmin}} \left\{ eBIC(\widehat{S}_{\nu_0}) \right\}$$

3 Applications numériques

Les performances de sélection de cette procédure ont été étudiées sur des données simulées selon un modèle de croissance logistique. Les résultats sont très encourageants, le bon support est sélectionné par notre procédure dans une grande majorité des cas. Comme attendu, pour différents nombres de covariables p fixés, le bon support est d'autant plus souvent sélectionné que le nombre d'individus n augmente et que la variance inter-individuelle Γ^2 diminue. Nous avons également l'intention de comparer notre méthode

à une procédure MCMC (Markov Chain Monte-Carlo) classique, et nous sommes assez confiants concernant la vitesse de notre méthode par rapport au MCMC.

Enfin, une application sur des données de sénescence de feuilles de blé est en cours. On s'intéresse à un stress azoté imposé à un panel de 220 variétés de blé tendre. Les variétés y répondent différemment, notamment on observe que certaines supportent mieux le stress et la sénescence est retardée. L'objectif est de sélectionner des marqueurs moléculaires, parmi plusieurs milliers, qui seraient associés à cette tolérance.

4 Remerciements

Ce travail a été financé par le projet Stat4Plant ANR-20-CE45-0012. Les auteurs remercient la plateforme INRAE MIGALE (MIGALE, INRAE, 2020. Migale bioinformatics Facility, doi : 10.15454/1.5572390655343293E12) pour les moyens de calcul et les capacités de stockage.

Bibliographie

[Bertrand et Balding, 2013] Bertrand, J. et Balding, D. J. (2013). Multiple single nucleotide polymorphism analysis using penalized regression in nonlinear mixed-effect pharmacokinetic models. *Pharmacogenetics and genomics*, 23(3):167–174.

[Chen et Chen, 2008] Chen, J. et Chen, Z. (2008). Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771.

[Delattre *et al.*, 2014] Delattre, M., Lavielle, M., Poursat, M.-A. *et al.* (2014). A note on bic in mixed-effects models. *Electronic journal of statistics*, 8(1):456–475.

[Delyon *et al.*, 1999] Delyon, B., Lavielle, M. et Moulines, E. (1999). Convergence of a stochastic approximation version of the em algorithm. *Annals of statistics*, pages 94–128.

[Dempster *et al.*, 1977] Dempster, A. P., Laird, N. M. et Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society : Series B (Methodological)*, 39(1):1–22.

[George et McCulloch, 1993] George, E. I. et McCulloch, R. E. (1993). Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423): 881–889.

[Kuhn et Lavielle, 2004] Kuhn, E. et Lavielle, M. (2004). Coupling a stochastic approximation version of em with an mcmc procedure. *ESAIM : Probability and Statistics*, 8:115–131.

[Lavielle, 2014] Lavielle, M. (2014). *Mixed effects models for the population approach : models, tasks, methods and tools*. CRC press.

[Müller *et al.*, 2013] Müller, S., Scealy, J. L. et Welsh, A. H. (2013). Model selection in linear mixed models. *Statistical Science*, 28(2):135–167. Publisher : Institute of Mathematical Statistics.

[Ollier, 2021] Ollier, E. (2021). Fast selection of nonlinear mixed effect models using penalized likelihood. *arXiv preprint arXiv :2103.01621*.

[Ročková et George, 2014] Ročková, V. et George, E. I. (2014). Emvs : The em approach to bayesian variable selection. *Journal of the American Statistical Association*, 109(506):828–846.

[Schelldorfer *et al.*, 2011] Schelldorfer, J., Bühlmann, P. et DE GEER, S. V. (2011). Estimation for high-dimensional linear mixed-effects models using 1-penalization. *Scandinavian Journal of Statistics*, 38(2):197–214.

[Schelldorfer *et al.*, 2014] Schelldorfer, J., Meier, L. et Bühlmann, P. (2014). Glmmlasso : an algorithm for high-dimensional generalized linear mixed models using 1-penalization. *Journal of Computational and Graphical Statistics*, 23(2):460–477.

Mixture of experts posterior surrogates for approximate Bayesian computation

Florence Forbes [1], Hien Duy Nguyen [3], TrungTin Nguyen [1] & Julyan Arbel [1]

[1] *Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France*
florence.forbes@inria.fr, trung-tin.nguyen@inria.fr, julyan.arbel@inria.fr
[3] *School of Mathematics and Physics, University of Queensland, St. Lucia, Australia.*
h.nguyen7@uq.edu.au

Résumé. Les procédures de calcul bayésien approché (ABC) reposent sur l'évaluation de l'écart entre les données simulées et les données observées. Cet écart est souvent évalué en comparant des statistiques résumées plutôt que directement les données. Le choix d'une distance et de résumés appropriés est donc une étape cruciale qui peut affecter la qualité des approximations. Dans ce travail, nous introduisons une étape d'apprentissage préliminaire dans laquelle des lois de substitution, issues d'un modèle de mélange d'experts, sont construites pour approximer les lois a posteriori visées. Ces lois a posteriori de substitution sont ensuite utilisées à la place des statistiques résumées et comparées à l'aide de métriques entre distributions. On montre que la quasi-loi a posteriori résultante converge vers la vraie loi a posteriori, sous des conditions standard. Des expériences montrent que notre approche est particulièrement performante lorsque la loi a posteriori est multimodale.

Mots-clés. Approximate Bayesian computation, Mélanges d'experts, statistiques résumées, distance de Wasserstein, modèles de substitution, lois a posteriori mulitmodales.

Abstract. A key ingredient in approximate Bayesian computation (ABC) procedures is the choice of a discrepancy that describes how different the simulated and observed data are, often based on a set of summary statistics when the data cannot be compared directly. Unless discrepancies and summaries are available from expert or prior knowledge, which seldom occurs, they have to be chosen and this can affect the quality of approximations. The choice between discrepancies is an active research topic, which has mainly considered data discrepancies requiring samples of observations or distances between summary statistics. In this work, we introduce a preliminary learning step in which surrogate posteriors are built using a specific instance of a Mixture of Experts model. These surrogate posteriors are then used in place of summary statistics and compared using metrics between distributions in place of data discrepancies. The resulting ABC quasi-posterior distribution is shown to converge to the true one, under standard conditions. Experiments show that our approach is particularly useful when the posterior is multimodal.

Keywords. Approximate Bayesian computation, Gaussian mixtures, summary statistics, Wasserstein distance, surrogate models, multimodal posterior distributions.

1 Introduction

Approximate Bayesian computation (ABC) appears as a natural candidate for addressing estimation problems where there is a lack of availability or tractability of the likelihood, but when the data generating process is available as a tractable simulation procedure. The fundamental idea of ABC is to generate parameter proposals $\boldsymbol{\theta}$ in a parameter space Θ using a prior distribution $\pi(\boldsymbol{\theta})$ and accept a proposal if the simulated data $\mathbf{z}$ for that proposal is similar to the observed data $\mathbf{y}$, where $\mathbf{z}, \mathbf{y} \in \mathcal{Y}$. This similarity is usually measured using a distance or discriminative measure D, whereby a simulated sample $\mathbf{z}$ is retained if $D(\mathbf{z}, \mathbf{y})$ is smaller than a given threshold ϵ. In this simple form, the procedure is generally referred to as rejection ABC. Other variants are possible and often recommended, for instance using MCMC or sequential procedures. We will focus on the rejection version as all developments can be easily adapted to more sophisticated variants.

In the case of a rejection algorithm, selected samples are drawn from the so-called ABC quasi-posterior, which is an approximation to the true posterior $\pi(\boldsymbol{\theta} \mid \mathbf{y})$. Under conditions similar to those of Bernton et al. (2019), regarding the existence of a probability density function (pdf) $f_{\boldsymbol{\theta}}(\mathbf{z})$ for the likelihood, the ABC quasi-posterior depends on D and on a threshold ϵ, and can be written as

$$\pi_\epsilon(\boldsymbol{\theta} \mid \mathbf{y}) \propto \pi(\boldsymbol{\theta}) \int_{\mathcal{Y}} \mathbf{1}_{\{D(\mathbf{y},\mathbf{z}) \leq \epsilon\}} \, f_{\boldsymbol{\theta}}(\mathbf{z}) \, d\mathbf{z} \, . \tag{1.1}$$

More specifically, the similarity between $\mathbf{z}$ and $\mathbf{y}$ is evaluated based on two components: the choice of summary statistics $s(\cdot)$ to account for the data in a more robust manner, and the choice of a distance to compare the summary statistics. That is, $D(\mathbf{y}, \mathbf{z})$ in (1.1) should then be replaced by $D(s(\mathbf{y}), s(\mathbf{z}))$, whereupon we overload D to also denote the distance between summary statistics $s(\cdot)$.

However, there is no general rule for constructing good summary statistics for complex models and if a summary statistic does not capture important characteristics of the data, the ABC algorithm is likely to yield samples from an incorrect posterior (Fearnhead and Prangle, 2012). Great insight has been gained through the work of Fearnhead and Prangle (2012), who introduced the *semi-automatic* ABC framework and showed that under a quadratic loss, the optimal choice for the summary statistic of $\mathbf{y}$ was the true posterior mean of the parameter: $s(\mathbf{y}) = \mathbb{E}[\boldsymbol{\theta} \mid \mathbf{y}]$. This conditional expectation cannot be calculated analytically but can be estimated by regression using a learning data set prior to the ABC procedure itself.

Our first contribution is to investigate an alternative efficient way to construct summary statistics, in the same vein as semi-automatic ABC, but based on combinations of posterior moments, not restricted to the posterior means. For this purpose, the Gaussian Locally Linear Mapping (GLLiM) method (Deleforge et al., 2015) appears as a good candidate regression model, with properties that balance between computationally expensive neural networks and simple standard regression techniques. GLLiM provides, at

low cost, a parametric estimation of the true posterior distributions. Using a learning set of parameter and observation pairs, GLLiM learns a family of finite Gaussian mixtures whose parameters depend analytically on the observation to be inverted. Such models correspond to Gaussian mixture of experts whose moments can be easily computed and used as summary statistics.

Our second contribution is to propose to compare directly the full surrogate posterior distributions provided by GLLiM, without reducing them to their moments. This requires the specification of a distance to compare such distributions. The recent Mixture-Wasserstein distance, denoted throughout the text as MW_2, designed for Gaussian mixtures (Delon and Desolneux, 2020) match perfectly this need. There exist other distances between mixtures that are tractable such as the L_2 distance, which is also considered in this work.

A remarkable feature of our approach is that it can be equally applied to settings where a sample of *i.i.d.* observations is available (*e.g.* Bernton et al. 2019) and to settings where a single observation is available as a vector, a time series realization, or a data set that is reduced to a vector of summary statistics (*e.g.* Fearnhead and Prangle 2012).

Regarding the approach's theoretical properties, we provide two results. In the first result, the true posterior is used to compare samples $\mathbf{y}$ and $\mathbf{z}$. In the second result, a surrogate posterior is learned and used to compare samples. Conditions are specified under which the resulting ABC quasi-posterior converges to the true posterior. More details can be found in Forbes et al. (2021).

2 Extended semi-automatic ABC

A learning set $\mathcal{D}_N = \{(\boldsymbol{\theta}_n, \mathbf{y}_n), n \in [N]\}$ is built from the joint distribution that results from the prior $\pi(\boldsymbol{\theta})$ on $\boldsymbol{\theta}$ and the likelihood $f_{\boldsymbol{\theta}}$, where $[N] = \{1, \ldots, N\}$. The idea is to capture the relationship between $\boldsymbol{\theta}$ and $\mathbf{y}$ with a joint probability model with computationally inexpensive and straightforward conditional distributions and moments. For the choice of the model to fit to $\mathcal{D}_N$, we propose to use the so-called GLLiM model (Deleforge et al., 2015) for its ability to capture non-linear relationships in a tractable manner, based on flexible mixtures of Gaussian distributions. GLLiM provides, for each observed $\mathbf{y}$, a full posterior probability distribution within a family of parametric Mixture of Expert models $\{p_G(\boldsymbol{\theta} \mid \mathbf{y}; \boldsymbol{\phi}), \boldsymbol{\phi} \in \boldsymbol{\Phi}\}$. To model non-linear relationships, it uses a mixture of K linear models. More specifically, the expression of $p_G(\boldsymbol{\theta} \mid \mathbf{y}; \boldsymbol{\phi})$ is analytical and available for all $\mathbf{y}$ with $\boldsymbol{\phi}$ being independent of $\mathbf{y}$:

$$p_G(\boldsymbol{\theta} \mid \mathbf{y}; \boldsymbol{\phi}) = \sum_{k=1}^{K} \eta_k(\mathbf{y})\, \mathcal{N}(\boldsymbol{\theta}; \boldsymbol{A}_k\mathbf{y} + \boldsymbol{b}_k, \boldsymbol{\Sigma}_k), \tag{2.1}$$

where $\mathcal{N}(\,\cdot\,; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the Gaussian pdf with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ and $\eta_k(\mathbf{y}) = \pi_k \mathcal{N}(\mathbf{y}; \boldsymbol{c}_k, \boldsymbol{\Gamma}_k) / \sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{y}; \boldsymbol{c}_j, \boldsymbol{\Gamma}_j)$. This distribution involves parameters:

$\boldsymbol{\phi} = \{\pi_k, \boldsymbol{c}_k, \boldsymbol{\Gamma}_k, \boldsymbol{A}_k, \boldsymbol{b}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K$. The parameter $\boldsymbol{\phi}$ can be estimated using an Expectation-Maximization (EM) algorithm (Deleforge et al., 2015). Fitting a GLLiM model to $\mathcal{D}_N$ therefore results in a set of parametric distributions $\{p_G(\boldsymbol{\theta} \mid \mathbf{y}; \boldsymbol{\phi}^*_{K,N}),\ \mathbf{y} \in \mathcal{Y}\}$, which can be seen as a parametric mapping from $\mathbf{y}$ values to posterior pdfs on $\boldsymbol{\theta}$. The parameter $\boldsymbol{\phi}^*_{K,N}$ is the same for all conditional distributions and does not need to be re-estimated for each new instance of $\mathbf{y}$.

Instead of comparing simulated $\mathbf{z}$'s to the observed $\mathbf{y}$, or comparing their summary statistics, we propose to compare $p_G(\boldsymbol{\theta} \mid \mathbf{z}; \boldsymbol{\phi}^*_{K,N})$'s and $p_G(\boldsymbol{\theta} \mid \mathbf{y}; \boldsymbol{\phi}^*_{K,N})$, as given by (2.1). We then derive two procedures, referred to as GLLiM-MW2-ABC and GLLiM-L2-ABC in Algorithm 1, respectively.

Algorithm 1 GLLiM-ABC algorithms – Vector and functional variants

1: **Inverse operator learning.** Apply GLLiM on a training set $\mathcal{D}_N = \{(\boldsymbol{\theta}_n, \mathbf{y}_n), n \in [N]\}$ to estimate, for any $\mathbf{z} \in \mathcal{Y}$, the K-Gaussian mixture $p_G(\boldsymbol{\theta} \mid \mathbf{z}; \boldsymbol{\phi}^*_{K,N})$ in (2.1) as a first approximation of the true posterior $\pi(\boldsymbol{\theta} \mid \mathbf{z})$, where $\boldsymbol{\phi}^*_{K,N}$ does not depend on $\mathbf{z}$.

2: **Distances computation.** Consider another set $\mathcal{E}_M = \{(\boldsymbol{\theta}_m, \mathbf{z}_m), m \in [M]\}$. For a given observed $\mathbf{y}$, do one of the following for $m \in [M]$:

Vector summary statistics.

GLLiM-E-ABC: Compute statistics $s_1(\mathbf{z}_m) = \mathbb{E}_G[\boldsymbol{\theta} \mid \mathbf{z}_m; \boldsymbol{\phi}^*_{K,N}]$.

GLLiM-EV-ABC: Compute both $s_1(\mathbf{z}_m)$ and $s_2(\mathbf{z}_m) = \mathrm{Var}_G[\boldsymbol{\theta} \mid \mathbf{z}_m; \boldsymbol{\phi}^*_{K,N}]$.

In both cases, compute standard distances between summary statistics.

Functional summary statistics.

GLLiM-MW2-ABC: Compute $\mathrm{MW}_2(p_G(\cdot \mid \mathbf{z}_m; \boldsymbol{\phi}^*_{K,N}), p_G(\cdot \mid \mathbf{y}; \boldsymbol{\phi}^*_{K,N}))$.

GLLiM-L2-ABC: Compute $\mathrm{L}_2(p_G(\cdot \mid \mathbf{z}_m; \boldsymbol{\phi}^*_{K,N}), p_G(\cdot \mid \mathbf{y}; \boldsymbol{\phi}^*_{K,N}))$.

3: **Sample selection.** Select $\boldsymbol{\theta}_m$ values that lead to distances under an ϵ threshold (rejection ABC) or apply an ABC procedure that can handle distances, directly.

4: **Sample use.** For a given observed $\mathbf{y}$, use the produced sample of $\boldsymbol{\theta}$ values to compute a closer approximation of $\pi(\boldsymbol{\theta} \mid \mathbf{y})$.

3 Theoretical properties

One important question is the proximity of the resulting so-called ABC quasi-posterior to the true posterior. We assume a fixed given observed $\mathbf{y}$ and the dependence on $\mathbf{y}$ is omitted from the notation. Let us first recall the standard form of the ABC quasi-posterior, omitting summary statistics from the notation:

$$\pi_\epsilon(\boldsymbol{\theta} \mid \mathbf{y}) = \frac{\int_{\mathcal{Y}} \mathbf{1}_{\{D(\mathbf{y},\mathbf{z}) \leq \epsilon\}}\ \pi(\boldsymbol{\theta} \mid \mathbf{z})\ \pi(\mathbf{z})\ d\mathbf{z}}{\int_{\mathcal{Y}} \mathbf{1}_{\{D(\mathbf{y},\mathbf{z}) \leq \epsilon\}}\ \pi(\mathbf{z})\ d\mathbf{z}}\ . \tag{3.1}$$

If D is a distance and $D(\mathbf{y}, \mathbf{z})$ is continuous in $\mathbf{z}$, the ABC posterior in (3.1) can be shown to have the desirable property of converging to the true posterior when ϵ tends to 0. The proof is based on the fact that when ϵ tends to 0, due to the property of the distance D, the set $\{\mathbf{z} \in \mathcal{Y} : D(\mathbf{y}, \mathbf{z}) \leq \epsilon\}$, defining the indicator function in (3.1), tends to the singleton $\{\mathbf{y}\}$ so that consequently $\mathbf{z}$ in the likelihood can be replaced by the observed $\mathbf{y}$, which then leads to an ABC quasi-posterior proportional to $\pi(\boldsymbol{\theta}) f_{\boldsymbol{\theta}}(\mathbf{y})$ and therefore to the true posterior as desired. It is interesting to note that this proof is based on working on the term under the integral only and is using the equality, at the limit, of $\mathbf{z}$ to $\mathbf{y}$, which is actually a stronger assumption than necessary for the result to hold.

We can then replace $D(\mathbf{y}, \mathbf{z})$ by $D(\pi(\cdot \mid \mathbf{y}), \pi(\cdot \mid \mathbf{z}))$, with D now denoting a distance on densities, and obtain the same convergence result when ϵ tends to 0. More specifically, we can show the following general result. Let us define our ABC quasi-posterior as,

$$q_\epsilon(\boldsymbol{\theta} \mid \mathbf{y}) = \frac{\int_{\mathcal{Y}} \mathbf{1}_{\{D(\pi(\cdot|\mathbf{y}), \pi(\cdot|\mathbf{z})) \leq \epsilon\}} \pi(\boldsymbol{\theta} \mid \mathbf{z}) \, \pi(\mathbf{z}) \, d\mathbf{z}}{\int_{\mathcal{Y}} \mathbf{1}_{\{D(\pi(\cdot|\mathbf{y}), \pi(\cdot|\mathbf{z})) \leq \epsilon\}} \pi(\mathbf{z}) \, d\mathbf{z}} \,. \tag{3.2}$$

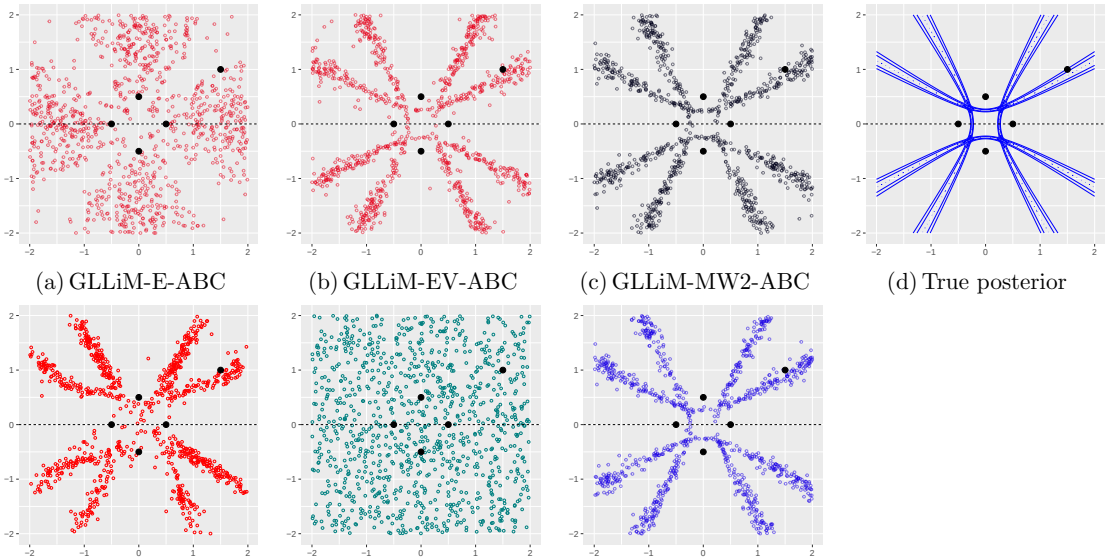
The following Theorem 3.1, proved in Forbes et al. (2021), shows that $q_\epsilon(\cdot \mid \mathbf{y})$ converges to $\pi(\cdot \mid \mathbf{y})$ in total variation, for fixed $\mathbf{y}$. For $\epsilon > 0$, $A_\epsilon = \{\mathbf{z} \in \mathcal{Y} : D(\pi(\cdot \mid \mathbf{y}), \pi(\cdot \mid \mathbf{z})) \leq \epsilon\}$.

Theorem 3.1. *If $\pi(\boldsymbol{\theta} \mid \cdot)$ is continuous for all $\boldsymbol{\theta} \in \Theta$, and $\sup_{\boldsymbol{\theta} \in \Theta} \pi(\boldsymbol{\theta} \mid \mathbf{y}) < \infty$; there exists a $\gamma > 0$ such that $\sup_{\boldsymbol{\theta} \in \Theta} \sup_{\mathbf{z} \in A_\gamma} \pi(\boldsymbol{\theta} \mid \mathbf{z}) < \infty$; $D(\cdot, \cdot) : \Pi \times \Pi \to \mathbb{R}_+$ is a metric on the functional class $\Pi = \{\pi(\cdot \mid \mathbf{y}) : \mathbf{y} \in \mathcal{Y}\}$; and $D(\pi(\cdot \mid \mathbf{y}), \pi(\cdot \mid \mathbf{z}))$ is continuous, with respect to $\mathbf{z}$. Then, $q_\epsilon(\cdot \mid \mathbf{y})$ in (3.2) converges in total variation to $\pi(\cdot \mid \mathbf{y})$, for fixed $\mathbf{y}$, as $\epsilon \to 0$.*

In most ABC settings, based on data discrepancy or summary statistics, the above consideration and result are not useful because the true posterior is unknown by construction and cannot be used to compare samples. However this principle becomes useful in our setting, which is based on surrogate posteriors. While the previous result can be seen as an oracle of sorts, it is more interesting in practice to investigate whether a similar result holds when using surrogate posteriors in the ABC likelihood. This is the goal of Theorem 2 in Forbes et al. (2021) which shows the convergence of the ABC quasi-posterior for a restricted class of target distributions and surrogate posteriors learned as mixtures.

4 Illustration

Numerical experiments, available in Forbes et al. (2021), show the versatility of our approach, applicable on both *i.i.d.* samples and single observation settings. The example below shows that it is particularly useful, when the posterior is multimodal. The object of interest is an unknown parameter $\boldsymbol{\theta} = (x, y)$ that can be interpreted as a source location in a 2D scene. To create a multimodal posterior, we consider a likelihood detailed in Forbes et al. (2021). The true posterior is shown in Figure 1 (d).



(a) GLLiM-E-ABC (b) GLLiM-EV-ABC (c) GLLiM-MW2-ABC (d) True posterior

(e) GLLiM mixture (f) Semi-automatic ABC (g) GLLiM-L2-ABC

Figure 1: GLLiM is learned with $K = 38$ and $N = 10^5$ while ABC is run usiing $M = 10^6$ simulations for (a,b,f,h) and $M = 10^5$ for (c,g). (a) GLLiM-E-ABC, (b) GLLiM-EV-ABC, (c) GLLiM-MW2-ABC, (d) contours of the true posterior, (e) approximate GLLiM posterior for the observed data, (f) semi-automatic ABC and (g) GLLiM-L2-ABC.

References

Bernton, E., Jacob, P. E., Gerber, M., and Robert, C. P. (2019). Approximate Bayesian computation with the Wasserstein distance. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81:235–269. 2, 3

Deleforge, A., Forbes, F., and Horaud, R. (2015). High-Dimensional Regression with Gaussian Mixtures and Partially-Latent Response Variables. *Statistics and Computing*, 25(5):893–911. 2, 3, 4

Delon, J. and Desolneux, A. (2020). A Wasserstein-type distance in the space of Gaussian Mixture Models. *SIAM Journal on Imaging Sciences*. 3

Fearnhead, P. and Prangle, D. (2012). Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(3):419–474. 2, 3

Forbes, F., Nguyen, H. D., Nguyen, T. T., and Arbel, J. (2021). Approximate Bayesian computation with surrogate posteriors. *Preprint. hal-03139256*. 3, 5

Bayesian nonparametric mixtures inconsistency for the number of clusters

Louise Alamichel[1,*], Daria Bystrova[1], Julyan Arbel[1] & Guillaume Kon Kam King[2]

[1] *Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP, LJK, 38000 Grenoble, France*
{louise.alamichel, daria.bystrova, julyan.arbel}@inria.fr
[2] *Université Paris-Saclay, INRAE, MaIAGE, 78350 Jouy-en-Josas, France*
guillaume.kon-kam-king@inrae.fr

Résumé. Les modèles de mélange bayésiens non paramétriques sont souvent utilisés pour modéliser des données complexes. Si ces modèles sont bien adaptés à l'estimation de densité, leur application au clustering, bien que courante, reste plus discutée. En effet, Miller and Harrison (2014) montrent l'inconsistance a posteriori du nombre de clusters lorsque le nombre réel de clusters est fini pour les modèles de mélange à processus de Dirichlet et à processus de Pitman–Yor. Dans ce travail, nous étendons ce résultat à d'autres priors bayésiens non paramétriques tels que les processus de type Gibbs et la représentation en dimension finie de ces différents priors. Ces représentations finies comprennent le processus multinomial de Dirichlet, de Pitman–Yor et celui de gamma généralisé normalisé, récemment proposés. Plus précisément, nous montrons que les modèles de mélange basés sur tous ces processus sont également inconsistants quant au nombre de clusters.

Mots-clés. Bayésien non-paramétrique, consistance, mélange fini

Abstract. Bayesian nonparametric mixture models are often employed for modelling complex data. While these models are well-suited for density estimation, their application for clustering has some limitations. Miller and Harrison (2014) proved posterior inconsistency in the number of clusters when the true number of clusters is finite for Dirichlet process and Pitman–Yor process mixture models. In this work, we extend this result to additional Bayesian nonparametric priors such as Gibbs-type processes and finite-dimensional representations of them. The latter include the Dirichlet multinomial process and the recently emerged Pitman–Yor and normalized generalized gamma multinomial processes. We show that mixture models based on these processes are also inconsistent in the number of clusters.

Keywords. Bayesian nonparametric, consistency, finite mixture

*This work has been partially supported by the LabEx PERSYVAL-Lab (ANR-11-LABX-0025-01) funded by the French program Investissements d'avenir.

1 Introduction

Finite mixture models are hugely popular in applications, but choosing the appropriate number of components remains challenging. Using Bayesian nonparametric (BNP) priors (such as the Dirichlet process) for mixture modelling allows to avoid this choice, by assuming an infinite number of components, which implies that the number of clusters found in a dataset is flexibly estimated depending on the amount and the structure of the data. While BNP mixture models are well-suited for density estimation, they have some limitations when employed for clustering data known to originate from a finite mixture. Indeed, Miller and Harrison (2014) states that when using a Dirichlet process or Pitman–Yor process mixture the posterior will not be consistent for the number of clusters if the data is generated according to a finite mixture. Here we show that this type of inconsistency result generalizes to Gibbs-type process mixtures and finite-dimensional representations of BNP priors.

We start by introducing the notion of a partition-based mixture model, before stating the inconsistency results of Miller and Harrison (2014) on Dirichlet process mixtures and Pitman–Yor process mixtures. We then present our generalization of this result to Gibbs-type processes and to several finite dimensional representations of BNP processes. For clarity, we omit the proofs of our results. However, the proofs are presented in article in preparation, which is a part of PhD project of Louise Alamichel.

1.1 Partition-based mixture model

We consider partition-based mixture model as defined in Miller and Harrison (2014). Firstly, we define the distribution on partitions. Let $\mathcal{A}_k(n)$ be the set of ordered partitions of $\{1, \dots, n\}$ into k ($k \in \{1, \dots, n\}$) nonempty sets:

$$\mathcal{A}_k(n) := \left\{ (A_1, \dots, A_k) : A_1, \dots, A_k \text{ disjoint}, \bigcup_{i=1}^{k} A_i = \{1, \dots, n\}, \ |A_i| \geq 1 \ \forall i \right\}.$$

We denote $n_i := |A_i|$. Then, we consider a distribution $p(A)$ on $\bigcup_{k=1}^{n} \mathcal{A}_k(n)$, which induces a distribution on k as $p(k \mid A) = \mathbb{I}(A \in \mathcal{A}_k(n))$ where $\mathbb{I}$ is the indicator function.

Next we introduce the hierarchical mixture model. We denote π a prior density on the d-dimensional parameters $\theta \in \Theta \subset \mathbb{R}^d$ and p_θ a parametrized component density. The hierarchical structure of *partition-based mixture model* is then:

$$p(A, k) = p(A)\mathbb{I}(A \in \mathcal{A}_k(n)), \quad p(\theta_{1:k}|A, k) = \prod_{i=1}^{k} \pi(\theta_i), \quad p(x_{1:n}|A, k, \theta_{1:k}) = \prod_{i=1}^{k} \prod_{j \in A_i} p_{\theta_i}(x_j),$$

where $x_{1:n} = (x_1, \dots, x_n)$ with $x_i \in \mathcal{X}$, $\theta_{1:k} = (\theta_1, \dots, \theta_k)$ with $\theta_i \in \Theta$, and $A \in \mathcal{A}_k(n)$. In

particular, it is a Dirichlet process mixture model when

$$p(A) = \frac{\alpha^k}{k!(\alpha)_n} \prod_{i=1}^{k} (n_i - 1)!$$

for $A \in \mathcal{A}_k(n)$, where $\alpha > 0$ and $(x)_n = x(x+1)\cdots(x+n-1)$, with $(x)_0 = 1$ by convention.

Finally, we denote by K_n the number of clusters.

1.2 Inconsistency theorem

The central result of Miller and Harrison (2014, Theorem 6) is reproduced below as Theorem 1.1. This result depends on two conditions which are given thereafter.

Firstly, we introduce some notations for Condition 1. For $A \in \mathcal{A}_k(n)$, we define $R_A = \bigcup_{i:|A_i|>2} A_i$, the union of all clusters except singletons. For $j \in R_A$, we define $B(A, j)$ to be the ordered partition $B \in \mathcal{A}_{k+1}(n)$ obtained by removing j from its cluster and creating a new singleton for it. Then $B_i = A_i \setminus \{j\}$, $i = 1, \ldots, k$, and $B_{k+1} = \{j\}$. Let $\mathcal{Z}_A := \{B(A, j) : j \in R_A\}$, for $n > k \geq 1$, we define

$$c_n(k) := \frac{1}{n} \max_{A \in \mathcal{A}_k(n)} \max_{B \in \mathcal{Z}_A} \frac{p(A)}{p(B)},$$

with the convention that $0/0 = 0$ and $y/0 = \infty$ for $y > 0$.

Condition 1. *Assume* $\limsup_{n\to\infty} c_n(k) < \infty$*, given some particular* $k \in \{1, 2, \ldots\}$.

The second condition induces a control on the likelihood through the control of single-cluster marginals. A single-cluster marginal of the model is $m(x_{A_i}) = \int_\Theta \left(\prod_{j \in A_i} p_\theta(x_j)\right) \pi(\theta) \mathrm{d}\theta$. Given $c \in [0, \infty)$, we introduce

$$\varphi_k(x_{1:n}, c) := \min_{A \in \mathcal{A}_k(n)} \frac{1}{n} \left| S_A(x_{1:n}, c) \right|,$$

where $S_A(x_{1:n}, c)$ is the set of indices $j \in \{1, \ldots, n\}$ such that the part A_ℓ containing j satisfies $m(x_{A_\ell}) \leq c m(x_{A_\ell \setminus j}) m(x_j)$, i.e. the set of observations for which the marginals of the new clusters obtained after taking out that observation and creating a new singleton cluster dominates the marginal of the original cluster up to a constant c.

Condition 2. *Given a sequence of random variables* $X_1, X_2, \ldots \in \mathcal{X}$*, and* $k \geq 1$*, assume* $\sup_{c \in [0,\infty)} \liminf_{n\to\infty} \varphi_k(X_{1:n}, c) > 0$ *a.s.*

For $c = 1$, Condition 2 means that as $n \to \infty$, there is always a non-vanishing proportion of the observations for which creating a singleton cluster will increase its cluster marginal.

Theorem 1.1 (Miller and Harrison, 2014)**.** *Let $X_1, X_2, \ldots \in \mathcal{X}$ be a sequence of r.v., and consider a partition-based model. Then, if Conditions 1 and 2 hold for any $k \geq 1$, we have*

$$\limsup_{n\to\infty} p(K_n = k|X_{1:n}) < 1 \quad \textit{with probability } 1.$$

The first condition is only related to partition distribution, while the second condition only involves the data-distribution and single cluster marginals. Hence, to generalize this inconsistency result to other processes, it is enough to show that Condition 1 also holds for these different processes.

2 Gibbs-type process

Gibbs-type processes are a natural generalization of the Dirichlet and Pitman–Yor processes. Gibbs-type processes of order: $\sigma, \sigma < 1$, can be characterized through the probability distribution of the induced random partition, which has the following form:

$$p(n_1, \ldots, n_k) = V_{n,k} \prod_{j=1}^{k} (1-\sigma)_{n_j-1}, \tag{1}$$

where $V_{n,k}$ are nonnegative numbers that satisfy the recurrence relation

$$V_{n,k} = (n - \sigma k)V_{n+1,k} + V_{n+1,k+1}, \quad V_{1,1} = 1. \tag{2}$$

In this case, we have a probability distribution for ordered partition $A \in \mathcal{A}_k(n)$ which can be deduced from (1) by dividing on $k!$ to adjust for order: $p(A) = \frac{V_{n,k}}{k!} \prod_{j=1}^{k} (1-\sigma)_{n_j-1}$.

Proposition 2.1. *Consider a Gibbs-type process with $0 < \sigma < 1$, then Condition 1 holds for any $k \in \{1, 2, \ldots\}$, so does inconsistency of Theorem 1.1.*

Idea of proof. For $B = B(A, j)$ as defined in section 1.2, we want to prove that

$$c_n(k) = \frac{1}{n} \max_{A\in\mathcal{A}_k(n)} \max_{B\in\mathcal{Z}_A} \frac{p(A)}{p(B)} < \infty$$

As $\frac{1}{n}\frac{p(A)}{p(B)} \leq \frac{V_{n,k}}{V_{n,k+1}}(k+1)$, we just have to prove that the sequence $\left(\frac{V_{n,k}}{V_{n,k+1}}\right)_{n\geq 0}$ is bounded. Coefficients $V_{n,k}$ can be equivalently defined through the density f_σ of a positive stable random variable

$$V_{n,k} = \frac{\sigma^k}{\Gamma(n-k\sigma)} \int_0^{+\infty} \int_0^1 t^{-k\sigma} p^{n-k\sigma-1} h(t) f_\sigma((1-p)t) \mathrm{d}t \mathrm{d}p.$$

Applying the Laplace approximation method as in Arbel and Favaro (2021), we obtain that the sequence $\left(\frac{V_{n,k}}{V_{n,k+1}}\right)_{n\geq 0}$ is bounded and Condition 1 is satisfied. □

3 Finite dimensional representation

The recent works by Lijoi et al. (2020a,b) develop finite-dimensional versions of the Pitman–Yor process and normalized random measures with independent increments (NRMI). The latter includes the Dirichlet and normalized generalized gamma multinomial processes as special cases. These processes can be seen as approximations for the corresponding infinite dimensional processes.

3.1 Pitman–Yor multinomial process

The Pitman–Yor multinomial process is defined in Lijoi et al. (2020b) as a discrete random probability measure $\tilde{p}_H$ such that $\tilde{p}_H|\tilde{p}_{0,H} \sim \mathrm{PY}(\sigma, \alpha; \tilde{p}_{0,H})$, $\tilde{p}_{0,H} = \frac{1}{H}\sum_{h=1}^{H} \delta_{\tilde{\theta}_h}$ where $H \geq 1$, $\tilde{\theta}_h \overset{\text{iid}}{\sim} P$, $\sigma \in [0, 1)$ and $\alpha > -\sigma$. The distribution on ordered partitions for the Pitman–Yor multinomial process is as follows

$$p(A) = \frac{H!}{(H-k)!k!} \frac{1}{(\alpha+1)_{n-1}} \sum_{(\ell_1,\ldots,\ell_k)} \frac{\Gamma(\alpha/\sigma + |\ell^{(k)}|)}{\sigma\Gamma(\alpha/\sigma+1)} \prod_{i=1}^{k} \frac{C(n_i, \ell_i; \sigma)}{H^{\ell_i}},$$

where the sum runs over the vectors $\ell^{(k)} = (\ell_1, \ldots, \ell_k)$ such that $\ell_i \in \{1, \ldots, n_i\}$ and $|\ell^{(k)}| = \ell_1 + \cdots + \ell_k$. The $C(n, k; \sigma)$ are the generalised factorial coefficients defined as

$$C(n, k; \sigma) = \frac{1}{k!} \sum_{j=0}^{k} (-1)^j \binom{k}{j} (-j\sigma)_n. \tag{3}$$

Proposition 3.1. *Consider a Pitman–Yor multinomial process, then Condition 1 holds for any $k < \min(n, H)$, so does the inconsistency of Theorem 1.1.*

3.2 Normalized infinitely divisible multinomial processes

Normalized infinitely divisible multinomial (NIDM) processes were introduced in Lijoi et al. (2020a). NIDM processes can be described through NRMI measures using following hierarchical structure: $(\tilde{p}_H \mid \tilde{p}_{0,H}) \sim \mathrm{NRMI}(c, \rho; \tilde{p}_{0,H})$, $\tilde{p}_{0,H} = \frac{1}{H}\sum_{h=1}^{H} \delta_{\tilde{\theta}_h}$.

If the NRMI process in the definition of the NIDM process is a Dirichlet process, then we can we obtain the Dirichlet multinomial process and the distribution on the ordered partitions is defined as:

$$p(A) = \Pi_H(n_1, \ldots, n_k) = \frac{H!}{(H-k)!k!} \frac{\prod_{j=1}^{k} (c/H)_{n_j}}{(c)_n}$$

Proposition 3.2. *Consider a Dirichlet multinomial process, then Condition 1 holds for any $k < \min(n, H)$, so does inconsistency of Theorem 1.1.*

Similarly, when we consider normalized generalized gamma as NRMI process in the definition, we get NGG multinomial process.

In this case the probability on the ordered partition is slightly more involved: We have with $k \leq \min(n, H)$ (see Example 2 Lijoi et al., 2020a):

$$P(A) = \frac{H!}{(H-k)!k!} \sum_{(\ell_1,\ldots,\ell_k)} \frac{V_{n,|\ell^{(k)}|}}{H^{|\ell^{(k)}|}} \prod_{i=1}^{k} \frac{C(n_i, \ell_i; \sigma)}{\sigma^{\ell_i}},$$

Here, the $C(n, k; \sigma)$ are again the generalized factorial coefficients defined in (3). The $V_{n,k}$ are the parameters defined in (2) for the particular case of NGG processes.

Proposition 3.3. *Consider a normalized generalized gamma multinomial process, then Condition 1 holds for any $k < \min(n, H)$, so does inconsistency of Theorem 1.1.*

4 Discussion

We have proved that Gibbs-type process mixtures are inconsistent a posteriori for the number of clusters, when the true number of components is finite. It is also the case, for some finite-dimensional representations of BNP priors as the Dirichlet multinomial process. However, we did not prove inconsistency in general for NIDM. A recent proposal by Guha et al. (2021) shows that this inconsistency problem can be resolved for Dirichlet process mixtures using a simple post-processing procedure based on merging neighbouring clusters. This procedure is effective as long as the posterior contracts at a known rate under the Wasserstein metric to the true mixing measure, which raises hope concerning its applicability to a larger class of processes than the Dirichlet process mixture.

References

Arbel, J. and Favaro, S. (2021). "Approximating predictive probabilities of Gibbs-type priors." *Sankhya A*, 83(1), 496–519.

Guha, A., Ho, N., and Nguyen, X. (2021). "On posterior contraction of parameters and interpretability in Bayesian mixture modeling." *Bernoulli*, 27(4), 2159–2188.

Lijoi, A., Prünster, I., and Rigon, T. (2020a). "Finite-dimensional discrete random structures and Bayesian clustering." *Preprint*.

— (2020b). "The Pitman–Yor multinomial process for mixture modelling." *Biometrika*, 107(4), 891–906.

Miller, J. W. and Harrison, M. T. (2014). "Inconsistency of Pitman-Yor process mixtures for the number of components." *The Journal of Machine Learning Research*, 15(1), 3333–3370.

Annulé

Pedestrian traffic mapping over time at the scale of a city

Mounia Zaouche [1] & Nikolai Bode [2]

[1] *Department of Engineering Mathematics, University of Bristol, Ada Lovelace Building, University Walk, Bristol, BS8 1TW. Email : fh20286@bristol.ac.uk*

[2] *Department of Engineering Mathematics, University of Bristol, Ada Lovelace Building, University Walk, Bristol, BS8 1TW. Email : nikolai.bode@bristol.ac.uk*

Résumé. La marche est une bonne alternative aux déplacements motorisés car elle permet de réduire les embouteillages et la pollution atmosphérique, mais aussi de renforcer la cohésion sociale. L'activité des piétons est par ailleurs un indicateur du dynamisme et de la prospérité financière d'une ville. La compréhension de la distribution du trafic dans le temps et l'espace est une question fondamentale en ingénierie des transports. Cependant la cartographie du trafic piéton à de grandes échelles spatio-temporelles bien que nécessaire reste rare. Cette étude vise à utiliser des données de comptage de piétons localisées, pour cartographier le trafic piéton dans le temps à l'échelle d'une ville en utilisant la modélisation bayésienne. Nous avons utilisé la méthode d'inférence puissante Integrated Nested Laplace Approximations (INLA) combinée à l'approche Stochastic Partial Differential Equation (INLA-SPDE) pour exploiter un grand nombre de données de comptage horaire enregistrées par plusieurs capteurs installés dans le centre-ville de Melbourne, mais aussi pour obtenir des résultats *a posteriori* rapides et précis. Nous avons développé et testé deux modèles spatio-temporels sophistiqués décrivant la variabilité spatiale au sein des données mais intégrant également différents termes d'interaction, y compris des interactions spatio-temporelles. Les critères de l'erreur quadratique moyenne ou Root Mean Square Error (RMSE) et du Deviance Information Criterion (DIC) utilisés pour analyser et comparer leurs performances ont permis de mettre en lumière l'un de ces modèles. Ce dernier a permis d'observer une amélioration de plus de 39 % en termes de RMSE en comparaison avec l'écart-type des données utilisées. Des cartes de prédiction et d'incertitude dans le temps ont finalement été produites.

Mots-clés. Cartographie du trafic piéton, dynamique piétonne, modélisation spatio-temporelle, intéractions espace-temps, modèles macroscopiques, INLA, SPDE

Abstract. Walking is a good alternative to motorized vehicle trips as it helps reducing traffic congestion and air pollution, but it also strengthens community cohesion. Pedestrian activity is an indicator of the dynamism and the financial prosperity of a city. Understanding the distribution of traffic in time and space over available infrastructures is an important issue in transport engineering, but pedestrian traffic mapping over large spatio-temporal scales is often missing and remains necessary. This study aims to make use of available localised pedestrian count data to map pedestrian traffic over time at the scale of cities using Bayesian modelling. We employed the powerful Integrated Nested

Laplace Approximations inference method combined with the Stochastic Partial Differential Equation (INLA-SPDE) to exploit a large hourly count data collected from several sensors installed across the city centre of Melbourne, and to obtain fast and accurate posterior results. We developed and tested two sophisticated spatio-temporal models describing the spatial variability within the data set but also integrating different interaction terms, including space-time interactions. The Root Mean Square Error (RMSE) and the Deviation Information Criterion (DIC) criteria used to analyse and compare their performances allowed to make one of these models stand out. This model resulted in an improvement in terms of RMSE of more than 39% considering the standard deviation of the data set used. Prediction and uncertainty maps over time were then produced.

Keywords. Pedestrian traffic maps, pedestrian dynamics, spatio-temporal modelling, space-time interactions, macroscopic models, INLA, SPDE.

1 Introduction

Pedestrian activity is a key research topic in transport engineering as it depicts the vitality and vibrancy of city but it also benefits ecology and increases community cohesion. In constrast to vehicular traffic, pedestrian dynamics mapping at large spatio-temporal is missing, as most studies either focus on small spatial scales at a high temporal resolution (Chraibi et al. (2019)) or on long term trends over a large spatial scale. Direct observation approaches have been employed to map pedestrian traffic via signals from geolocated personal devices, being mainly mobile phones (Zhang et al., (2016)). Although this method has shown promise, it overly relies on the cooperation of individuals or companies and does not account for individuals without personal devices. Some researchers have considered other techniques such as regression based approaches (Cooper et al. (2021), Pfiester et al. (2021), Wang et al. (2017)) or the Geographically Weighted Regression (Piefster et al. (2021)) using generally locally observed data (e.g. pedestrian counts recorded at a junction), hence preserving the privacy of passers-by. However, most of this work has failed to describe the spatial variations across the domain or have not fully captured temporal changes in pedestrian traffic dynamics by resorting to coarse temporal partitioning. One of the main focus of this study is to integrate as much information as possible and to benefit from the usefulness of Bayesian hierarchical models in explaining simultaneous variation across space-time. As an alternative to the simulation-based Markov Chain Monte Carlo (MCMC) method computationally intensive when it deals with high dimensional data, we used the Integrated Nested Laplace Approximations (INLA) combined with the Stochastic Partial Differential Equation (SPDE) approach (Rue et al. (2009), Lindgren (2022)), to integrate hourly pedestrian count data collected over 2 months in 2019 by pedestrian sensor device installed throughout the city centre of Melbourne and to get accurate and fast posterior estimates. We built up two types of spatio-temporal

models enabling the incorporation of : (i) the spatial auto-correlation within the data set relying on the Euclidean distance between sensors across the area; (ii) a fine temporal resolution using every day of the week and every hour of the day; (iii) different interaction terms, including space-time interactions of different shape. These models have been compared via the Root Mean Square Error (RMSE) and the Deviance Information Criterion (DIC;Spiegelhater et al. (2002)) which measures the goodness of fit of Bayesian models.

2 Data

The City of Melbourne gives access to hourly pedestrian count data recorded from 2009 to date, at various sensors installed throughout the area (data.melbourne.vic.gov.au). We selected for this study 34 active sensors shown in Figure 1 (red, and green locations) to gather hourly data recorded every day in January and February of 2019 (59 days).
We represented the infrastructure of streets by a network consisting of nodes connected to each other by edges and used the free geographic OpenStreetMap (OSM) data. The nodes are modeled as geographical points and the edges as geographical lines (Figure 1 on the left).

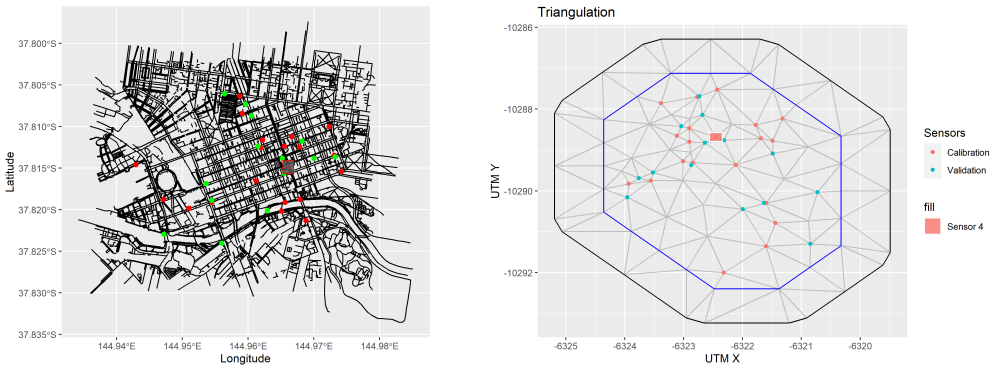


Figure 1: Study area, 20 calibration sensors in red and 14 validation sensors in green. Sensor 4 is framed by a red square in both plots. Left : street network of the city centre of Melbourne (longitude/latitude coordinates). Right : Triangulation of the domain in the UTM coordinate system (kms).

3 Spatio-temporal models

We assume that the 34 time-series (related to the 34 sensors) share the same type of periodicity (daily and weekly periods), but have a different amplitude at each location. The models need to account for the daily and weekly seasonalities but also for spatial

variations across the study area, and space-time interactions.
We denote by y_{idh} the number of pedestrians measured at sensor i, at day d ($1 \leq d \leq 7$) of the week and at hour h ($1 \leq h \leq 24$) of the day.

3.1 Spatio-temporal model I

The first model denoted M^{STI} integrates a separate spatial effect and 2 space-time terms with an exchangeable model on the space component (e.g with all of the spatial time-series being correlated in the same way) :

$$\begin{array}{rcl} y_{idh} & \sim & Poisson(\lambda^I_{idh}), \\ \eta^I_{idh} & = & \log(\lambda^I_{idh}) = \beta^I_0 + \epsilon^I_i + \xi_i + \alpha_{ih} + \omega_{id}. \end{array} \tag{1}$$

with β^I_0 the intercept, ϵ^I_i the error term at site i, and ξ_i being the i-th realization of the Gaussian field ξ with an isotropic Matérn covariance function.
The interaction terms α and ω describe respectively the hourly and daily temporal processes over space. On the time components, we assume autoregressive processes and a uniform correlation over space. Here, the only spatially structured process is ξ.

3.2 Spatio-temporal model II

This model denoted M^{STII} accounts for a day-hour interaction term and a space-time interaction but with this time a structured spatial effect on the space component :

$$\begin{array}{rcl} y_{idh} & \sim & Poisson(\lambda^{II}_{idh}), \\ \eta^{II}_{idh} & = & \log(\lambda^{II}_{idh}) = \beta^{II}_0 + \epsilon^{II}_i + \delta_{ih} + \gamma_{hd}. \end{array} \tag{2}$$

with β^{II}_0 the intercept, ϵ^{II}_i the error term at site i. For the δ space-time effect, the spatial process is a Matérn Gaussian field evolving in time following an autoregressive dynamics. The γ component describes the interaction between the hourly and daily effects both being autoregressive processes.

4 Inference method : INLA-SPDE

The Integrated Nested Laplace Approximations (INLA) is an algorithm designed for the class of Latent Gaussian Models (LGM) and whose objective is to estimate the marginal posterior distribution of the parameters through efficient numerical algorithms and Laplace approximations in a minimum time (Blangiardo et al., (2015)). In the spatial and spatio-temporal modelling field, INLA is generally combined with the Stochastic Partial Differential Equation approach (SPDE; Lindgren et al. (2022)) to approximate the solution of a partial differential equation being a Gaussian field with Matérn correlation by a Gaussian Markov Random Field (GMRF; Rue et al. (2005)) characterized by local

neighborhood and a sparse precision matrix. Thus, for beneficial computational purposes, ξ (1) and the space component of δ (2) are Gaussian fields estimated through the SPDE approach.
An approximate solution of the SPDE is obtained by using the finite element method resulting in the triangulation of the study area (right plot of Figure 1).

5 Results

We calibrated the spatio-temporel models on a subset of 20 sensors (red locations in Figure 1), hence by exploiting $59 * 24 * 20 = 28320$ measurements, and used an external validation set of counts collected over the same period at 14 sensors (green locations in Figure 1). The RMSE and DIC criteria clearly emphasize the performance of M^{STII} as it yields the lowest RMSE (499.66 versus 598.58) and DIC (2573350.24 versus 2679619.90) values. It also represents an improvement of 39.2 % versurs 27.1 % considering the standard deviation of the data set (820.66).This favourable effect can only be attributed to interactions between space and time, as well as between time and day. The M^{STI} is to have the same spatial predictions for each time point. The time, as well as the day of week, simply vary the amplitude. Conversely, looking at the maps produced by the M^{STII} model on the street network (Figure 2), the spatial pattern clearly changes over time and the predictions stay in line with what the data suggest.

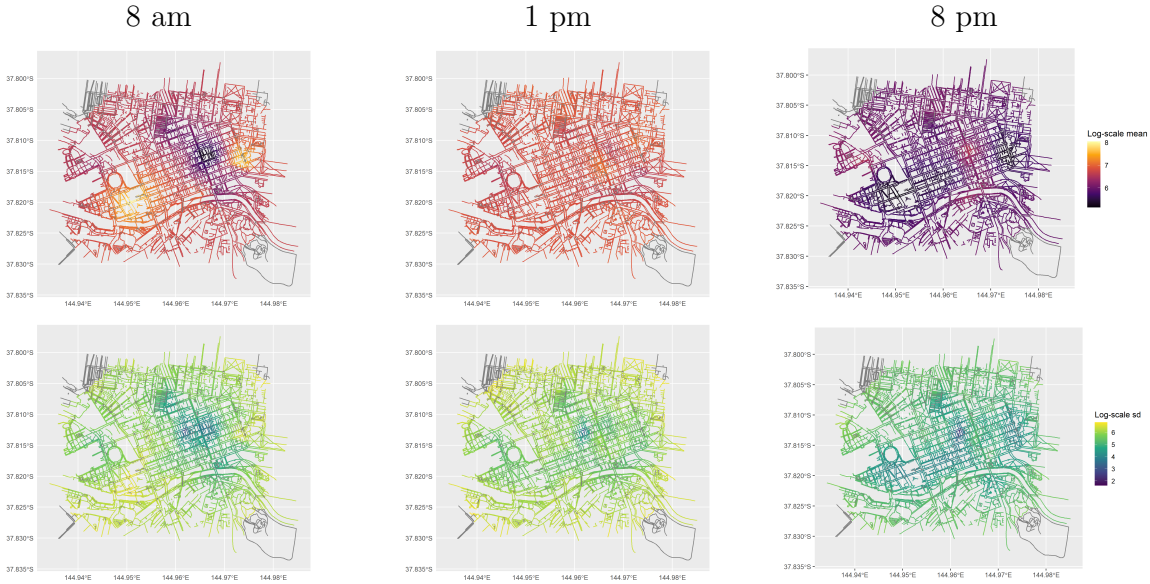


Figure 2: Maps of the pedestrian counts log-scale mean (first row) and standard deviation (second row) estimated by the M^{STII} on the street network, on Monday at 8 am, 1 pm and 8 pm.

6 Conclusion

In this study, we built up and tested sophisticated Bayesian spatio-temporal models and mapped pedestrian traffic at the scale of a city using a fine temporal resolution. We also demonstrated the beneficial inclusion of the structured interaction effects such as space-time or time-time (day of the week and time of the day). However, these models only rely on Euclidean distances and essentially capture background levels of pedestrian traffic. The next step consists in considering street network covariates as well as different distance metrics depicting more realistically pedestrian movement. Further work will be to investigate the adaptability of the M^{STII} model to changing dynamics over time using the Covid-19 pandemic.

Bibliographie

Blangiardo, M., Cameletti, M. (2015). Spatial and spatio-temporal Bayesian models with R-INLA, *John Wiley Sons*.

Chraibi, M., Tordeux, A., Schadschneider, A., Seyfried, A. (2019), *Complex Dynamics of Traffic Management.*, pp. 649-669

Cooper, C. H., Harvey, I., Orford, S., Chiaradia, A. J. (2021). Using multiple hybrid spatial design network analysis to predict longitudinal effect of a major city centre redevelopment on pedestrian flows, *Transportation*, 48(2), pp. 643-672.

Lindgren, F., Bolin, D., Rue, H. (2022). The SPDE approach for Gaussian and non-Gaussian fields: 10 years and still running, *Spatial Statistics*, 95, 100599.

Pfiester, L. M., Thompson, R. G., Zhang, L. (2021). Spatiotemporal exploration of Melbourne pedestrian demand, *Journal of transport geography*, 95, 103151.

Rue, H., Held, L. (2005). Gaussian Markov random fields: theory and application, *CRC press*.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., Van Der Linde, A. (2002). Bayesian measures of model complexity and fit, *Journal of the royal statistical society: Series b (statistical methodology)*, 64(4), pp. 583-639.

Wang, X., Liono, J., McIntosh, W., Salim, F. D. (2017). A Predicting the city foot traffic with pedestrian sensor data, *In Proceedings of the 14th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*, pp. 1-10.

Zhang, K., Wang, M., Wei, B., Sun, D. (2016). Identification and prediction of large pedestrian flow in urban areas based on a hybrid detection approach, *Sustainability*, 9, 36.

Bayesian functional linear regression estimation. Extension to scalar and categorical covariates

Feriel Bouhadjera [1], Meïli Baragatti [1], Nadine Hilgert [1], Nathalie Smits [2] & Paul-Marie Grollemund [3]

[1] *MISTEA, Université Montpellier, INRAE, Institut Agro, Montpellier, France*
[2] *ABSys, Université Montpellier, INRAE, Institut Agro, Montpellier, France*
[3] *LMBP, Université Clermont Auvergne, Aubière, France*

Résumé. Nous considérons un modèle de régression linéaire où la variable à expliquer est réelle et les co-variables sont catégorielles, scalaires et fonctionnelles et agissent de manière additive dans le modèle. Notre objectif est d'estimer les paramètres de ce modèle, tout en conservant un caractère interprétable pour la partie du modèle incluant les variables fonctionnelles. Ce travail est une extension du modèle Bliss (*Bayesian functional Linear regression with Sparse Step function*), voir Grollemund et al. (2019), développé dans un cadre Bayésien et qui ne contient que des variables fonctionnelles. Nous proposons d'étendre la méthode Bliss à un modèle plus général contenant également des co-variables catégorielles et scalaires. Dans la suite, nous explicitons le modèle étendu et expliquons comment estimer les paramètres d'une manière interprétable. Une illustration est faite sur des données simulées et un jeu de données réelles sur le dépérissement de la vigne.

Mots-clés. Données fonctionnelles, méthode Bliss, régression linéaire, statistique Bayésienne.

Abstract. We consider a linear regression model with a scalar response variable and categorical, scalar and functional covariates, that act additively in the model. Our objective is to estimate the parameters of this model, while maintaining an interpretability for the part of the model including the functional covariates. This work is an extension of the Bliss model (*Bayesian functional Linear regression with Sparse Step function*), see Grollemund et al. (2019), developed in a Bayesian framework with only functional covariates. We propose to extend the Bliss method to a more general model that also contains categorical and scalar covariates. In the following, we explain the extended model and how to estimate the parameters in an interpretable way. An illustration is made on simulated data and a real data set obtained in the field of vine dieback.

Keywords. Bayesian statistics, Bliss method, functional data, linear regression.

1 Introduction

Technological advances allow us to collect large amount of data in many areas. In agronomy, a major challenge is to extract information from these massive data and establish links with final characteristics such as production yields or quality. The motivation of our study comes from a project on the vine dieback. A unique database obtained by the *Bureau National Interprofessionnel du Cognac* on the monitoring of 55 plots since 1977 offers the opportunity to analyse the determinants of multi-year vine dieback trajectories. In this project, the objectives are to be able to identify (1) the factors and interactions of biotic, abiotic and technical factors that contribute to the decline in plot yield and to the mortality of individual grapevines and (2) the time period over which these factors have an impact, both in the short term at the scale of the crop cycle and in the long term since the plot was planted. One way of modeling this problem is to build a regression model to explain the effect of climatic conditions and cultivation managements on the vine yields. The covariates involved in such a study are of different types : functional, categorical or scalar. An 'interpretable' linear regression model with functional covariates already exists in the Bayesian framework : the Bliss model, see Grollemund et al. (2019). In this model, the *prior* distribution assumes that the coefficient functions are step functions. Bliss allows to recover periods of time that influence the most the outcome, which helps answering objective (2) mentioned above. To take into account scalar and categorical covariates as well, we have to extend the Bliss model as follows.

From a statistical point of view, consider a p-dimensional scalar covariate vector $(z_1, \cdots, z_p)$, q functional covariates $(x_1(t), \cdots, x_q(t))$ on $\mathcal{C}$, which is an interval of $\mathbb{R}$, a categorical covariate α with r levels, and a scalar variable to be explained y. The statistical linear model under consideration is as follows, for a n-sample of individuals doubly indexed by (ij), where $i = 1, \cdots, r$ stands for the categorical levels and $j = 1, \cdots, n_i$ stands for the repetitions :

$$y^{(ij)} = \beta_0 + \alpha_i + \sum_{s=1}^{p} \gamma_s z_s^{(ij)} + \sum_{v=1}^{q} \int_{\mathcal{C}} \beta_v(t) x_v^{(ij)}(t) dt + \epsilon^{(ij)}, \tag{1}$$

where β_0 is the intercept, $(\gamma_1, \cdots, \gamma_p)$ and $(\alpha_1, \cdots, \alpha_r)$ are scalar parameters, and $(\beta_1(t), \cdots, \beta_q(t))$ are functional parameters. Finally, $\epsilon^{(ij)}$ are random variables. Note that this model could also be extended to 2 or more categorical covariates.

The problem we are interested in is to explain the response variable from all these covariates of different type. Statistical models that involve functional covariates with scalar and categorical covariates are not common.

Functional data models have been popularized by works such as Ramsey and Silverman (1997) and Ferraty and Vieu (2006). There are different derivatives of this model in the literature. Not to be exhaustive, we quote the generalized functional regression model (see, Müller et Stadtmüller (2005)) and the functional regression model with a functional

response (see, Yao et al. (2005)). It is known that the correlation between the covariates decreases the estimation quality. In the functional part of our model, there are obviously large correlations. To overcome this problem, robust estimators based on the minimization of a penalized least squares criterion such as Ridge or Lasso regression exist in the literature. However, despite the robustness of the proposed estimators, it is difficult to interpret the estimate.
From a practical point of view, it is important to estimate periods of time during which the climatic conditions have influenced the variable of interest. For a fixed v, if the coefficient function $\beta_v(t)$ is null over a period $\mathcal{C}$ then the integral $\int_{\mathcal{C}} \beta_v(t)x_v(t)dt$ can be written without using the values of x_v over $\mathcal{C}$ and if it is positive (resp. negative) over a period then the increase (resp. decrease) of the value of x_v over this period induces an increase (resp. decrease) of the value of y. It is important to highlight the periods where $\beta_v(t)$ is positive or negative. That's why we are interested in the estimation of the support of the coefficient function, which is an important feature of the interpretation. To that aim, the Bliss method focuses on the estimation of the support of the functional parameter, which makes the interpretation easier. The functional parameter is identified through a piecewise constant function appropriately chosen. The complex functional linear regression model is then simplified to a linear regression model. The Bliss model is estimated in a Bayesian framework which allows to take into account expert knowledge.
To fit linear regression models, several Bayesian approaches have been considered, and without being exhaustive we can list Wang et al. (2007) and Goldsmith et al. (2011). To our knowledge, the only paper specifically interested in the support of the coefficient function using a Bayesian approach is Grollemund et al. (2019). Moreover, in the last paper, the authors propose two estimators of the coefficient function with different properties. Our contribution is an extension of the Bliss method to other types of variables : scalar and categorical.

2 Model

We consider the functional parameters β_v as piecewise constant functions that can be described with a minimal number of intervals K_v :

$$\beta_v(t) = \sum_{k=1}^{K_v} \frac{b_{k,v}}{|\mathcal{I}_{k,v}|} \mathbb{1}_{\{t \in \mathcal{I}_{k,v}\}}, \qquad \text{for } v = 1, \cdots, q, \tag{2}$$

where $\mathcal{I}_{1,v}, \cdots, \mathcal{I}_{K_v,v}$ are intervals included in $\mathcal{C}$, $|\mathcal{I}_{k,v}|$ is the length of the interval and $b_{1,v}, \cdots, b_{K_v,v}$ are real parameters. The support is the union of all $\mathcal{I}_{k,v}$ if the coefficient b_k are not null. Thus, a period of time which does not influence the outcome will be outside the support. Replacing (2) in (1), the functional linear model becomes a linear model with

design depending on the intervals $\mathcal{I}_{k,v}$:

$$y^{(ij)} = \beta_0 + \alpha_i + \sum_{s=1}^{p} \gamma_s z_s^{(ij)} + \sum_{v=1}^{q} \sum_{k=1}^{K_v} b_{k,v} x_v^{(ij)}(\mathcal{I}_{k,v}) + \epsilon^{(ij)}, \tag{3}$$

with

$$x_v^{(ij)}(\mathcal{I}_{k,v}) = \frac{1}{|\mathcal{I}_{k,v}|} \int_{\mathcal{I}_{k,v}} x_v^{(ij)}(t)dt, \quad \text{for} \quad \mathcal{I}_{k,v} = [m_{1,v} \pm \ell_{1,v}, \cdots, m_{K_v,v} \pm \ell_{K_v,v}],$$

where for a fixed v, $m_{k,v}$ is the center and $\ell_{k,v}$ is the half length of the interval $\mathcal{I}_{k,v}$. Hence, let define the parameters set to model (1) of dimension $3K + r + p + 2$ by :

$$\theta \;=\; \left(\beta_0, (\alpha_1, \cdots, \alpha_r), (\gamma_1, \cdots, \gamma_p), (\theta_1, \cdots, \theta_q), \sigma^2\right),$$

where

$$\theta_v \;=\; ((b_{1,v}, \cdots, b_{K_v,v}), (m_{1,v}, \cdots, m_{K_v,v}), (\ell_{1,v}, \cdots, \ell_{K_v,v})).$$

For simplicity of notations, we focus on model (1) with only one scalar and one functional covariates. We assume that we observe $\{y^{(ij)},\ i = 1, \cdots, r$ and $j = 1, \cdots, n_i\}$ replicates of outcome, $(\alpha_1, \cdots, \alpha_r)$ levels of the categorical variable α and $\{z^{(ij)},\ i = 1, \cdots, r$ and $j = 1, \cdots, n_i\}$ are scalar observations and $\{x^{(ij)}(\mathcal{I}_k),\ i = 1, \cdots, r,\ j = 1, \cdots, n_i$ and $k = 1, \cdots, K\}$ are scalar observations. Notice that $n = \sum_{i=1}^{r} n_i$. The model (3) becomes :

$$y^{(ij)} | \beta_0, \alpha_i, \gamma, z^{(ij)}, (b_1, \cdots, b_K), x^{(ij)}(t) \rightsquigarrow \mathcal{N}\left(\beta_0 + \alpha_i + \gamma z^{(ij)} + \sum_{k=1}^{K} b_k x^{(ij)}(\mathcal{I}_k), \sigma^2\right). \tag{4}$$

We complete the Bayesian model (4) with the following *prior* distributions :

$$\beta_0 | \sigma^2 \rightsquigarrow \mathcal{N}\left(0, u_0\sigma^2\right), \qquad \alpha_i | \sigma^2 \rightsquigarrow \mathcal{N}\left(0, v_0\sigma^2\right), \qquad \gamma | \sigma^2 \rightsquigarrow \mathcal{N}\left(0, w_0\sigma^2\right),$$

and $\pi(\sigma^2) \propto 1/\sigma^2$. Furthermore, for $k = 1, \cdots, K$

$$\begin{aligned} b_k | \sigma^2, m_k, \ell_k &\rightsquigarrow \mathcal{N}_K\left(0, n\sigma^2(G + \nu\lambda_{max}(G) I_K)^{-1}\right), \\ m_k &\overset{i.i.d.}{\rightsquigarrow} \mathcal{U}(\mathcal{C}), \\ \ell_k &\overset{i.i.d.}{\rightsquigarrow} \exp\left(a \times |\mathcal{C}|\right), \end{aligned} \tag{5}$$

where G is the Gram matrix given by $G = x.(\mathcal{I}.)^t x.(\mathcal{I}.)$ with

$$x.(\mathcal{I}.) = \left\{x^{(ij)}(\mathcal{I}_k),\ \ i = 1, \cdots, r,\ j = 1, \cdots, n_i, \text{ and } k = 1, \cdots, K\right\}.$$

3 Implementation

The full posterior distribution can be written explicitly from the Bayesian model given in (5) by

$$\begin{aligned}
\beta|y,\sigma^2,m,\ell &\rightsquigarrow \mathcal{N}_K\left((\underline{x}^t\underline{x}+\underline{V})^{-1}\underline{x}^t y, \sigma^2(\underline{x}^t\underline{x}+\underline{V})^{-1}\right),\\
\sigma^2|y,\beta,m,\ell &\rightsquigarrow \Gamma^{-1}\left(\frac{n+K+r}{2}+1, \frac{1}{2}\left\{\mathrm{RSS}+\beta^t\underline{V}\beta\right\}\right),\\
f(m_k|y,\beta,\sigma^2,m_{-k},\ell) &\propto \exp\left\{-\mathrm{RSS}/2\sigma^2\right\}\times f(b|m_k,\ell_k,\sigma^2)\times f(m_k), \qquad k=1,\cdots,K\\
f(\ell_k|y,\beta,\sigma^2,\ell_{-k},m) &\propto \exp\left\{-\mathrm{RSS}/2\sigma^2\right\}\times f(b|m_k,\ell_k,\sigma^2)\times f(\ell_k), \qquad k=1,\cdots,K
\end{aligned}$$

where $\underline{x} = (1_n \mid A \mid z \mid x.(\mathcal{I}_.))$, $\beta^t = (\beta_0,(\alpha_1,\cdots,\alpha_r),\gamma,(b_1,\cdots,b_K))^t$ and $\mathrm{RSS} = ||y-\underline{x}\beta||^2$. Furthermore, let define

$$A = \begin{pmatrix}
1_{n_1} & 0_{n_1} & \ldots & \ldots & \ldots & \ldots & 0_{n_1}\\
0_{n_2} & 1_{n_2} & 0_{n_2} & \ldots & \ldots & \ldots & 0_{n_2}\\
\vdots & & & \ldots & & & \vdots\\
0_{n_i} & \ldots & 0_{n_i} & 1_{n_i} & 0_{n_i} & \ldots & 0_{n_i}\\
\vdots & & & & \vdots & & \\
0_{n_r} & \ldots & \ldots & \ldots & \ldots & 0_{n_r} & 1_{n_r}
\end{pmatrix} \qquad \underset{-}{V} = \begin{pmatrix}
v_0^{-1} & 0 & 0 & 0\\
0 & u_0^{-1} & 0 & 0\\
0 & 0 & w_0^{-1} & 0\\
0 & 0 & 0 & n^{-1}(G+\nu\lambda_{max}(G)\mathcal{I}_K)
\end{pmatrix}.$$

The full *posterior* conditional distributions of parameters m_k and ℓ_k are not know distributions. However, considering that m_k and ℓ_k evolve in a finite grid of their supports, it is possible to characterize these distributions by numerically computing their probability functions. Furthermore, the resulting Bayesian model depends on hyperparameters which are u_0, v_0, w_0, ν, a and K. It is possible to choose these hyperparameters so that the *prior* distributions are the less informative possible. For the parameter K, we choose it rather large which guarantees a low error insofar as the intervals can overlap. In Grollemund et al. (2019), the authors propose some default values for the hyperparameters. As usual with hierarchical models, sampling from the posterior distribution can be done with a Gibbs algorithm (see : Robert and Casella (2013), Chapter 7).

4 Applications

Numerical results based both on simulated data and real data from the vineyard will be presented.

Acknowledgments

The present work was supported by the LabEx NUMEV (ANR-10-LABX-0020) within the I-Site MUSE, and by the French National Research Agency under the Investments for the Future Program, referred as ANR-16-CONV-0004.

Bibliography

Grollemund, P.M., Abraham, C., Baragatti, M., Pudlo, P. (2019). Bayesian Functional Linear Regression with Sparse Step Functions, *Bayesian Analysis*, 14(1) :111–135.

Ramsay, J. and Silverman, B. (1997). Functional Data Analysis. *Springer-Verlag New-York*.

Ferraty, F. and Vieu, P. (2006). Nonparametric functional data analysis : theory and practice. *Springer Science & Business Media*.

Müller, H.-G. et Stadtm̈uller, U. (2005). Generalized functional linear models. *Annals of Statistics*, pages 774–805.

Yao, F., Müller, H.-G., Wang, J.-L., et al. (2005). Functional linear regression analysis for longitudinal data. *The Annals of Statistics*, 33(6) :2873–2903.

Wang, X., Ray, S., and Mallick, B. K. (2007). Bayesian curve classification using wavelets. *Journal of the American Statistical Association*, 102(479) :962–973.

Goldsmith, J., Wand, M. P., and Crainiceanu, C. (2011). Functional regression via variational Bayes. *Electronic journal of statistics*, 5 :572–602.

Robert, C.P. and Casella, G. (2013). Monte Carlo statistical methods. Springer-Verlag New York. 22, 72.

Applications : société & industrie

Modélisation de la Substitution entre Articles pour Optimiser leur Réapprovisionnement

Axel Potier[1], Christophe Biernacki[2], Julien Favre[3],
Matthieu Marbac[4] & Vincent Vandewalle[5]

[1] *Adeo service, Univ. Lille, CNRS, UMR 8524, Inria - Laboratoire Paul Painlevé, F-59000 Lille, France ; axel.potier@adeo.com*
[2] *Inria, Univ. Lille, CNRS, UMR 8524 - Laboratoire Paul Painlevé, F-59000 Lille, France ; christophe.biernacki@inria.fr*
[3] *Adeo service ; julien.favre@adeo.com*
[4] *Univ. Rennes, Ensai, CNRS, CREST - UMR 9194, F-35000 Rennes, France ; matthieu.marbac-lourdelle@ensai.fr*
[5] *Univ. Lille, CHU Lille, Inria - ULR 2694 METRICS, F-59000 Lille, France ; vincent.vandewalle@univ-lille.fr*

Résumé. Dans le commerce de détail, l'optimisation des stocks est un problème courant nécessitant un compromis entre le risque de rupture et les coûts générés par le sur-stock. Nous proposons une extension du modèle du vendeur de journaux (*Newsvendor problem*) prenant en compte le phénomène de substitution entre articles. Dans cette extension nous définissons un modèle de substitution probabiliste réaliste et générique, et nous montrons comment celui-ci intervient dans le calcul de la fonction de profit afin de déterminer les stocks optimaux. L'approche est illustrée sur des données simulées similaires aux données de vente rencontrées dans la société Adeo.

Mots-clés. Substitution, Problème du vendeur de journaux, Optimisation de réapprovisionnement

Abstract. In retail, inventory optimisation is a common problem requiring a trade-off between the risk of stock-out and the costs generated by overstocking. We propose an extension of the Newsvendor problem that takes into account the phenomenon of substitution between items. In this extension, we define a realistic and generic probabilistic substitution model, and we show how it is used in the calculation of the profit function to determine the optimal stocks. The approach is illustrated on simulated data similar to the sales data encountered at Adeo.

Keywords. Stock-out Substitution, Newsvendor Problem, Inventory Optimisation

1 Introduction

La prédiction des ventes est une étape essentielle dans la gestion d'un magasin, car elle permet de définir un réapprovisionnement évitant des sur-coûts induits par du sur-stockage

ou des ventes ratées du fait de produits en ruptures de stock (Hubert, 2013; Usunier and Bourbonnais, 2017). Lorsqu'un produit est indisponible, le client peut alors se reporter sur un produit similaire ou bien ne rien acheter. On parle ainsi de substitution entre produits (Shin et al., 2015). Dans ce travail, nous proposons d'incorporer le phénomène de substitution entre articles dans une stratégie de réapprovisionnement. Nous modélisons les ventes en deux étapes : 1. le choix d'un produit (demande primaire), 2. le report éventuel vers un autre article disponible en magasin en cas de demande initiale non satisfaite (demande secondaire) ; dans ce cas on parle de substitution. Nous proposons une stratégie de réapprovisionnement à partir de la modélisation précédente et d'une fonction objectif propre au magasin (dépendant des coûts d'achat et de stockage, des prix de ventes, du coût d'insatisfaction des clients lorsque la demande initiale n'est pas satisfaite, *etc.*) (Mahajan and Van Ryzin, 2001). Le modèle du vendeur de journaux (ou NVP pour "newsvendor problem", Choi (2012)) est un modèle classique pour le réapprovisionnement optimal. Dans ce modèle la fonction à optimiser est le profit. Celle-ci intègre la demande du client, des coûts tels que le prix de vente, d'achat, la valeur de récupération. Il existe de nombreuses extensions du NVP dans le cas où les produits sont substituables (Mcgillivray and Silver, 1978; Netessine and Rudi, 2003; Ma, 2016), mais celles-ci incorporent des hypothèses restrictives (demandes continues, non prise en compte de l'ordre dans lequel les produits entrent en rupture). D'autres modèles plus réalistes modélisent l'ordre d'arrivée des clients et intègrent les possibilités de report et de non achat (Mahajan and Van Ryzin, 2001). Ce problème du NVP avec substitution et l'estimation des lois associées à la demande et la substitution font l'objet de recherches actives (Yijie et al., 2022).

Dans la lignée de Mahajan and Van Ryzin (2001) nous proposons un modèle à la fois réaliste et générique qui cherche un compromis entre sur-stock et rupture pour limiter les coûts globaux. Sa flexibilité nous permet de proposer une variante dont l'objectif est l'application aux données réelles du groupe Adeo. Dans cet article nous n'adressons pas l'estimation des paramètres associés aux lois de ventes et aux probabilités de substitution qui font l'objet d'un travail en cours.

Dans la suite du document, nous présentons en partie 2 l'extension du modèle NVP. Puis en partie 3, nous illustrons par une application numérique l'intérêt de prendre en compte la substitution dans le réapprovisionnement. En partie 4, nous abordons les pistes retenues pour un travail ultérieur.

2 Modélisation des ventes de produits substituables

Au cours d'une période fixée, aucun réapprovisionnement n'est effectué et T clients arrivent, de manière séquentielle, avec une demande unitaire initiale $\boldsymbol{D}_t = (D_{t1}, \ldots, D_{tK})^\top$ pour un unique article, où $D_{tk} = 1$ si le client t a une demande (initiale) pour le produit k et $D_{tk} = 0$ sinon, avec $\sum_{k=1}^{K} D_{tk} = 1$. La demande initiale des clients n'étant pas toujours

satisfaite (*i.e.*, produit indisponible), un client peut se reporter sur un autre produit ou ne rien acheter, ce qui rend la "trajectoire" des demandes $\mathbf{D} = \{\boldsymbol{D}_t\}_{1\leq t\leq T}$. Il est important de noter que cette dernière est non observable en pratique. On note aussi $\mathbf{V} = \{\boldsymbol{V}_t\}_{1\leq t\leq T}$ la "trajectoire" des ventes où $\boldsymbol{V}_t = (V_{t0}, V_{t1}, \ldots, V_{tK})^\top$ avec $V_{t0} = 1$ si le client t n'achète aucun produit et $V_{tk} = 1$ si le client t achète le produit k, et $\sum_{k=0}^{K} V_{tk} = 1$. Les demandes initiales n'étant pas toujours satisfaites, on a généralement $\mathbf{V} \neq \mathbf{D}$. De plus, seules les évolutions de stocks sont observées. L'ordre des ventes effectuées pour $k \geq 1$ est observé au contraire de T ou encore des indices t des clients n'ayant rien acheté.

On considère ici que la demande initiale d'un client est insatisfaite uniquement si le produit est en rupture de stock. Considérant un stock initial $\boldsymbol{S} = (S_1, \ldots, S_K)^\top$ où $S_k \in \mathbb{N}$ est le stock du produit k au début de la période d'intérêt, le vecteur $\boldsymbol{I}_t = (I_{t1}, \ldots, I_{tK})^\top$ indique les produits encore accessibles pour le client t, avec $I_{tk} = 1$ si le produit k est accessible au client t (*i.e.,* $S_k - \sum_{t'=1}^{t-1} V_{t'k} \geq 1$, avec $\sum_{t=1}^{0} V_{tk} = 0$) et $I_{tk} = 0$ sinon (*i.e.,* $S_k = \sum_{t'=1}^{t-1} V_{t'k}$). Ainsi, pour le client t ayant une demande initiale k, si le produit est en stock le client l'achète (*i.e.,* $V_{tk} = 1$ si $D_{tk} = 1$ et $I_{tk} = 1$), alors que s'il n'est pas en stock un mécanisme de substitution dépendant de la demande initiale et des produits en stock se met en place (*i.e.,* la loi de $\boldsymbol{V}_t \mid \mathcal{F}_{t-1}, \boldsymbol{D}_t, \boldsymbol{I}_t$ est égale à $\boldsymbol{V}_t \mid \boldsymbol{D}_t, \boldsymbol{I}_t$, où $\mathcal{F}_{t-1}$ est la filtration naturelle). Ce mécanisme débouche sur la vente d'un autre produit (substitution) ou sur aucune vente (le client repart sans produit). Le client se reportera alors sur le produit $\ell \in [\![0, K]\!]$ (le cas $\ell = 0$ indique une vente perdue) avec une probabilité $\lambda_\ell(\boldsymbol{\pi}_k, \boldsymbol{I}_t)$, où $\forall k \in [\![1, K]\!]$, $\sum_{\ell=0}^{K} \lambda_\ell(\boldsymbol{\pi}_k, \boldsymbol{I}_t) = 1$ et par convention $\lambda_k(\boldsymbol{\pi}_k, \boldsymbol{I}_t) = 0$. Les probabilités de report de ventes $\lambda_\ell(\boldsymbol{\pi}_k, \boldsymbol{I}_t), \ldots, \lambda_\ell(\boldsymbol{\pi}_k, \boldsymbol{I}_t)$ dépendent de $\boldsymbol{\pi}_k \in \mathbb{R}^K$ définissant les probabilités de substitution du produit k par les autres produits inconditionnellement aux stocks et à la disponibilité des produits (définie par $\boldsymbol{I}_t$). Ainsi, la loi conditionnelle des ventes sachant la demande initiale et le stock est définie par

$$\forall k \in [\![1, K]\!],\ \mathbb{P}(V_{tk} = 1 \mid \boldsymbol{D}_t, \boldsymbol{I}_t) = I_{tk} D_{tk} + \sum_{\ell=1}^{K} (1 - I_{t\ell}) D_{t\ell} \lambda_k(\boldsymbol{\pi}_\ell, \boldsymbol{I}_t)$$

et

$$\mathbb{P}(V_{t0} = 1 \mid \boldsymbol{D}_t, \boldsymbol{I}_t) = \sum_{\ell=1}^{K} (1 - I_{t\ell}) D_{t\ell} \lambda_0(\boldsymbol{\pi}_\ell, \boldsymbol{I}_t).$$

Le profit occasionné sur la période de temps est déterminé par la fonction $\rho(\mathbf{D}, \mathbf{V}; \boldsymbol{\theta}, \boldsymbol{S})$ où $\boldsymbol{\theta}$ groupe les paramètres propres au magasin (coûts d'achat et de stockage, prix de ventes, coût d'insatisfaction des clients lorsque la demande initiale n'est pas satisfaite, *etc.* ; voir Mahajan and Van Ryzin (2001) pour des exemples). Notons que pour un stock initial fixé, le profit obtenu à la fin de la période considérée $\rho(\mathbf{D}, \mathbf{V}; \boldsymbol{\theta}, \boldsymbol{S})$ est une variable aléatoire. Dans un cadre de gestion des stocks, l'objectif est de déterminer le stock initial idéal $\boldsymbol{S}^\star_{\boldsymbol{\pi}}$ maximisant une certaine mesure statistique $m_{\eta,\boldsymbol{\pi}}$ (*e.g.,* espérance, q-quantile, *etc.*) de $\rho(\mathbf{D}, \mathbf{V}; \boldsymbol{\theta}, \boldsymbol{S})$ qui dépend de la loi des demandes initiales définie par η et des substituabilités entre produits définies par $\boldsymbol{\pi} = (\boldsymbol{\pi}_1^\top, \ldots, \boldsymbol{\pi}_K^\top)^\top$. Ainsi, on cherche

à déterminer

$$\boldsymbol{S}^{\star}_{\boldsymbol{\pi}} = \arg\max_{\boldsymbol{S}} m_{\eta,\boldsymbol{\pi}} \left[\rho(\mathbf{D}, \mathbf{V}; \boldsymbol{\theta}, \boldsymbol{S})\right].$$

3 Illustration numérique de l'intérêt du modèle

Instanciation du profit Dans le commerce de détail, il est classique (Van Ryzin 2001) de considérer des coûts très variés : coût d'achat, de vente, de récupération, par unité à la période de stockage et de rupture. Nous noterons ces différents coûts ω_k, η_k, ν_k, β_k et γ_k, respectivement, et on les regroupe dans le vecteur de paramètres $\boldsymbol{\theta} = (\omega_k, \eta_k, \beta_k, \nu_k, \gamma_k)_{k\in[K]}$. Ce paramètre va influer de la façon suivante sur le profit relatif au client t, noté pour l'occasion

$$\rho_t(\boldsymbol{D}_t, \boldsymbol{V}_t; \boldsymbol{\theta}) = \sum_{k=1}^{K} \mathbf{1}_{D_{tk}=1} \left[\sum_{\ell=1}^{K} \eta_\ell V_{t\ell} - \gamma_k V_{t0}\right],$$

prenant ainsi en considération que soit une vente est réalisée avec un gain associé, soit que la vente est perdue avec la pénalité financière qui en découle. Ce processus se répète pour chaque client t. À la fin il reste une certaine quantité de stock pour le produit k, notée S_k^{end} (cette quantité se déduit de l'ensemble du processus de vente), qui occasionne un coût de stockage auquel on retranche la valeur de récupération. En prenant aussi en compte le coût d'achat occasionné par le réapprovisionnement en début de période, on obtient alors le profit global sur tous les clients associé à la "trajectoire" $(\mathbf{D}, \mathbf{V})$ des demandes clients et des ventes effectives :

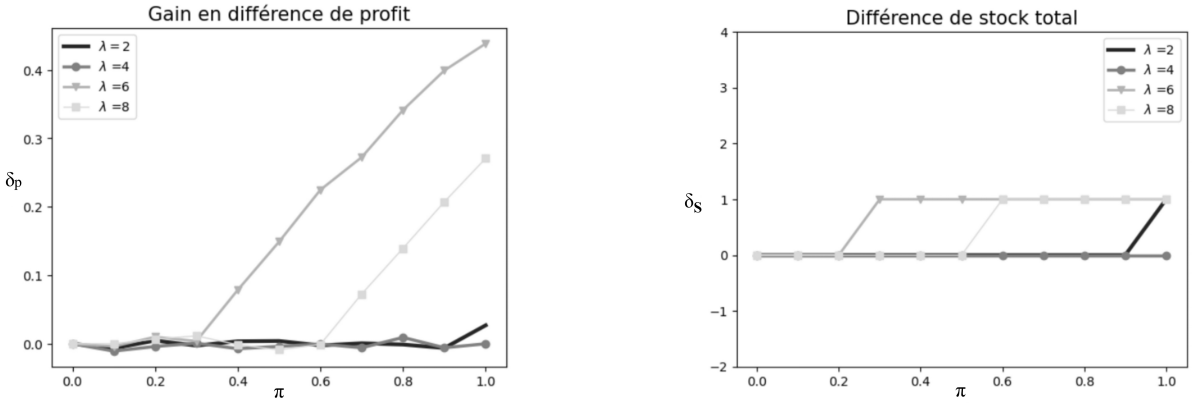
$$\rho(\mathbf{D}, \mathbf{V}; \boldsymbol{\theta}, \boldsymbol{S}) = \left\{\sum_{t=1}^{T} \rho_t(\boldsymbol{D}_t, \boldsymbol{V}_t; \boldsymbol{\theta}) - \sum_{k=1}^{K} (\beta_k - \nu_k) S_k^{\text{end}} - \sum_{k=1}^{K} \omega_k S_k\right\}.$$

L'intérêt du modèle de substitution proposé, en opposition à une méthode naïve qui ne tiendrait pas compte des substitutions alors qu'elles seraient présentes en réalité, pourra alors se mesurer au travers des deux quantités clés suivantes. D'une part le gain de profit, noté $\delta_\rho(\boldsymbol{\pi}) = m_{\eta,\boldsymbol{\pi}} \left[\rho(\mathbf{D}, \mathbf{V}; \boldsymbol{\theta}, \boldsymbol{S}^{\star}_{\boldsymbol{\pi}})\right] - m_{\eta,\boldsymbol{\pi}} \left[\rho(\mathbf{D}, \mathbf{V}; \boldsymbol{\theta}, \boldsymbol{S}^{\star}_{\mathbf{0}})\right]$, et d'autre part la différence d'approvisionnement initial, notée $\delta_{\boldsymbol{S}}(\boldsymbol{\pi}) = \sum_{k=1}^{K} \left[(\boldsymbol{S}^{\star}_{\boldsymbol{\pi}})_k - (\boldsymbol{S}^{\star}_{\mathbf{0}})_k\right]$.

À noter que dans la suite nous retiendrons la moyenne pour la mesure statistique $m_{\eta,\boldsymbol{\pi}}$. Par ailleurs, le calcul ainsi que l'optimisation du profit $m_{\eta,\boldsymbol{\pi}} \left[\rho(\mathbf{D}, \mathbf{V}; \boldsymbol{\theta}, \boldsymbol{S})\right]$ sont en général loin d'être triviaux, et feront partie des questions abordées dans le futur (voir la partie 4). Dans ce travail préliminaire, nous nous contenterons d'approcher cette quantité par une simple méthode de Monte Carlo (10^5 tirages), puis de l'optimiser par une procédure généraliste standard disponible sous R.

Scénario 1 : produits à faible dévaluation Dans le premier scénario, les valeurs (en €) ont été fixées de manière à représenter un scénario proche de ceux que l'on peut

trouver dans la compagnie Adeo, spécialisée dans le marché de l'habitat : $\eta_k = 15$, $\omega_k = 10$, $\nu_k = 9$, $\beta_k = 0.05$ et $\gamma_k = 1$. On constate ainsi que le coût de rupture est plus impactant que le coût de stockage. Les résultats de l'application numérique se trouvent dans la figure 1. La sous-figure de gauche illustre que le gain de profit δ_ρ, à λ fixé, augmente systématiquement quand la probabilité de substitution π s'accroît. La sous-figure de droite (liée à $\delta_{\boldsymbol{S}}$) indique que ce gain de profit est obtenu grâce à un léger sous-stockage.



(a) Gain de profit.

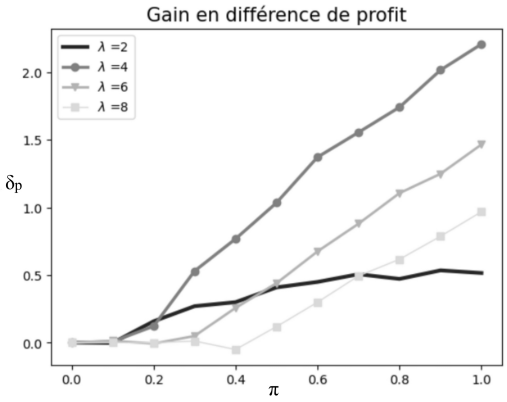
(b) Différence d'approvisionnement.

Figure 1: Cas de la valeur de récupération élevée (scénario 1).

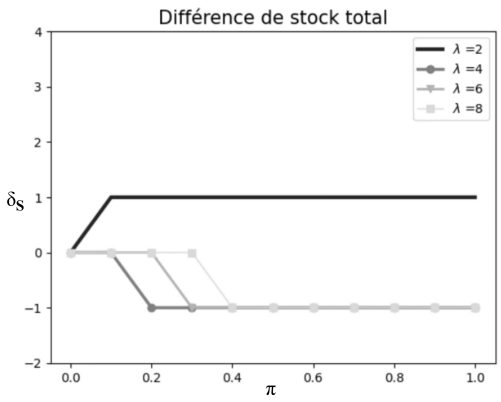
Scénario 2 : produits à forte dévaluation Le second scénario est associé à des produits de denrées périssables, donc éloigné du marché de l'habitat décrit par le scénario 1. On a ainsi une forte dévaluation des produits au cours de la période, ce qui se traduit par une forte perte de la valeur du stock. Afin de matérialiser cette situation, la seule différence avec le scénario précédent correspond à $\nu_k = 0$. Les résultats de cette nouvelle application numérique se trouvent dans la figure 2. Comme l'expérience précédente, à λ fixé on observe dans la sous-figure de gauche une augmentation du gain en fonction de π. La sous-figure de droite indique que ce gain est obtenu cette fois en concédant soit un peu plus de stock, soit un peu moins.

4 Questions ouvertes pour les travaux futurs

L'illustration précédente montre l'intérêt d'intégrer le mécanisme de substitution pour l'optimisation des coûts. Ces travaux préliminaires seront suivis par les questions d'inférence des lois d'arrivée des demandes et des probabilités de substitution. Pour cela nous nous attellerons aux questions d'identifiabilité des paramètres du modèle et à leur estimation.



(a) Gain de profit.



(b) Différence d'approvisionnement.

Figure 2: Cas de la forte dévaluation (scénario 2).

Nous nous intéresserons également aux problématiques d'évaluation et d'optimisation du profit, suite logique de la phase d'estimation.

References

Choi, Tsan-Ming (2012). *Handbook of Newsvendor problems: Models, extensions and applications*. Vol. 176. Springer.

Hubert, Thibault (Jan. 2013). "Prévision de la demande et pilotage des flux en approvisionnement lointain". Theses. Ecole Centrale Paris. URL: `https://tel.archives-ouvertes.fr/tel-00879853`.

Ma, Shouyu (2016). "Modélisation et Analyse de Nouvelles Extensions pour le Problème du Vendeur de Journaux". PhD thesis. Université Paris Saclay (COmUE).

Mahajan, S. and G. Van Ryzin (2001). "Stocking retail assortments under dynamic consumer substitution". In: *Operations Research* 49(3), pp. 334–351. ISSN: 0030364X.

Mcgillivray and Silver (1978). " Some Concepts For Inventory Control Under Substitutable Demand ". In: *Information Systems and Operational Research* 5986.

Netessine, Serguei and Nils Rudi (2003). "Centralized and Competitive Inventory Models with Demand Substitution". In: *SSRN Electronic Journal*.

Shin, Hojung et al. (2015). "A classification of the literature on the planning of substitutable". In: *European Journal of Operational Research* 246(3), pp. 686–699.

Usunier, Jean-Claude and Régis Bourbonnais (2017). *Prévision des Ventes : Théorie et pratique*. Economica. URL: `https://hal.archives-ouvertes.fr/hal-01507734`.

Yijie, Zheng et al. (2022). "Multi-Product Newsvendor with Stockout-Based Substitution: A Non-Parametric Approach". In: *Available at SSRN 4008913*.

Number of zero velocity points: a critical parameter for handwriting model estimation towards dysgraphia diagnosis assistance

Yunjiao Lu[1] & Jérôme Boutet[2] & Vincent Brault[1] & Caroline Jolly[3] & Etienne Labyt[2,4] & Raphaël Lambert[2,3] & Jean-Charles Quinton[1]

[1] *Univ. Grenoble Alpes, CNRS, Grenoble INP*, LJK, 38000 Grenoble, France*
[2] *CEA, University Grenoble Alpes, Leti, 38000 Grenoble, France*
[3] *CNRS, University Grenoble Alpes, University Savoie Mont Blanc, LPNC, 38000 Grenoble, France*
[4] *Mag4Health, 9 Avenue Paul Verlaine, 38000 Grenoble*
yunjiao.lu@univ-grenoble-alpes.fr

Résumé. Une nouvelle voie pour l'analyse de l'écriture manuscrite afin d'aider à diagnostiquer la dysgraphie chez les enfants, en utilisant des modèles oscillatoires dynamiques, a été étudiée. Le *Parsimonious Oscillatory Model of Handwriting* (POMH) s'est avéré capable de reconstruire la trace de l'écriture manuscrite à condition qu'elle soit suffisamment lisse. Pour les enfants souffrant de dysgraphie, POMH conduit à une plus grande déviation lors de la tentative de reconstruction du signal. Cette différence de performance de POMH est censée servir de discriminant pour la détection automatique des enfants dysgraphiques. Un paramètre critique pour POMH est le nombre de points de vélocité nulle dans l'écriture d'un symbole. Les relations entre l'âge, le nombre de points de vélocité nulle et la performance de POMH sont étudiées comme travail préliminaire pour établir un algorithme de classification pour l'aide au diagnostic de la dysgraphie.

Mots-clés. Dysgraphie, Modèle d'écriture, Construction de feature, Classification, Traitement du signal, Morphologie mathématique

Abstract. A new avenue for handwriting analysis to help diagnose dysgraphia in children, using dynamic oscillatory models, has been investigated. The Parsimonious Oscillatory Model of Handwriting (POMH) is shown to be able to reconstruct the handwriting trace provided that it is sufficiently smooth. For children suffering from dysgraphia, POMH leads to greater deviation when trying to reconstruct the signal. This difference in POMH performance is thought to serve as a discriminant for automatically detecting children with dysgraphia. A critical parameter for POMH is the number of zero velocity points when writing a symbol. The relationships between age, number of zero velocity points and performance of POMH are investigated as premilinary work to establish a classification algorithm for dysgraphia diagnosis assistance.

*Institute of Engineering Univ. Grenoble Alpes

Keywords. Dysgraphia, Handwriting model, Feature construction, Classification, Signal processing, Mathematical morphology

1 Introduction

Handwriting is crucial for successful integration into society, yet hinges on a long learning process. Handwriting disorders —called dysgraphias— may therefore yield severe consequences spanning from infancy to adulthood. For French children, dysgraphia is usually diagnosed using the BHK test, which consists in copying a text for 5 minutes, with the handwritten production then rated on 13 criteria by psychomotricity specialists (adapted in French by Charles et al. (2004) based on Hamstra-Bletz et al. (1987)). Given the difficulty of an evaluation solely based on visual traces, using tablets to record and exploit the handwriting dynamics has been proposed (Asselborn et al. (2018); Deschamps et al. (2019); Devillaine et al. (2020)), for instance considering the time to make certain strokes to better diagnose dysgraphia.

To better understand and exploit its dynamics, mathematical models of nonpathological handwriting have been proposed, assuming oscillatory and fluid generation of strokes (Hollerbach (1984); André et al. (2014)). The purpose of this work is to study how such models behave when applied to dysgraphic children handwriting, and whether a lack of fit may help diagnosis. To this end, the Parsimonious Oscillatory Model of Handwriting (POMH; André et al. (2014)) and the associated method for parameter estimation are first presented. Then, a first study on the differences in model estimation between diagnosed and non-diagnosed children is presented. A discussion on the future use of our results concludes this work.

2 Parsimonious Oscillatory Model of Handwriting

POMH assumes handwriting results from two orthogonal oscillators, corresponding to different degrees of freedom in hand movements. The model is defined as follows:

$$\begin{cases} \dot{x}(t) = a_x(t)\sin(\omega_x(t) \times t + \phi_x(t)) \\ \dot{y}(t) = a_y(t)\sin(\omega_y(t) \times t + \phi_y(t)) \end{cases}$$

where $\dot{x}$ (resp. $\dot{y}$) is the velocity on the x-axis (resp. y) and a_x, ω_x and ϕ_x (resp. a_y, ω_y, ϕ_y) are piecewise constant functions. The associated break points of the model $t_{x,1}, \cdots, t_{x,N_x}$ correspond to the moments of zero velocity along the x-axis (with the equiva-

lent parametrization for the y-axis), such that for all $i \in \{1, \ldots, N_x - 1\}$ and $t \in [t_{x,i}; t_{x,i+1}[$:

$$\omega_x(t) = \frac{\pi}{t_{x,i+1} - t_{x,i}}, \quad \phi_x(t) = -\frac{\pi t_{x,i}}{t_{x,i+1} - t_{x,i}}$$

$$\text{and} \qquad a_x(t) = \frac{\pi}{2\,(t_{x,i+1} - t_{x,i})} \sum_{j | t_j \in [t_{x,i}; t_{x,i+1}[} \frac{x_{j+1} - x_j}{t_{j+1} - t_j}$$

with (t_j, x_j, y_j) the recorded timestamps and coordinates of the pen point from the tablet. The correct estimation of the moments of zero velocity is therefore crucial for the reconstruction of the letters.

3 Number of zero velocity points: a critical parameter for the estimation of POMH

The moments of zero velocity must be obtained from the recorded handwriting data, here comprised of isolated symbols (letter or number). However, the latter are noisy and must be filtered to find a meaningful number of zeros which occurred when the participant wrote the symbol. To do so, the closing morphological operator (acting as a low-pass filter) is applied on a logical variable indicating whether the speed is cancelled for (x_t) and (y_t) discretized signals, respectively.

On the one hand, given that the temporal sampling rate of the tablet is stable ($\simeq$ 200Hz), the size w of the structuring element of the closing operator logically has a direct impact on the number of extracted zero velocity moments, with larger sizes filtering out more zeros. For instance, the first row in Fig. 1 corresponds to a symbol "a" written by a child in second grade (CE1 in the French education system); with w set to 3 (equivalent to a time window of roughly 15ms), 20 zeros are found in x and 26 in y, but with $w = 17$ ($\simeq 85ms$), the number of zeros respectively drops to 8 and 9. On the other hand, for each symbol, we expect a theoretical fixed number of zeros according to how handwriting is taught in school. However, more zeros do not necessarily mean that the child is dysgraphic, especially for young-aged children who are still in the learning stage. Nevertheless, when children advance through the learning process, the number of zeros should decrease and converge to a fixed number for a given symbol (details in the next section). This tendency is expected by handwriting professionals, even though at the end of the learning process, children usually personalize their handwriting and switch from script to cursive script (attaching letters for flow and efficiency).

Obviously, the more zeros are exploited, the more break points in POMH, and the better the fit to the data. Yet, when imposing a low number of zeros for a given symbol, POMH can successfully reconstruct non-pathological handwriting but has more difficulty with dysgraphic children. For instance on Fig. 1, the panels on the left emphasize the writing process w.r.t time; the child on top row wrote the letter "a" with a single stroke

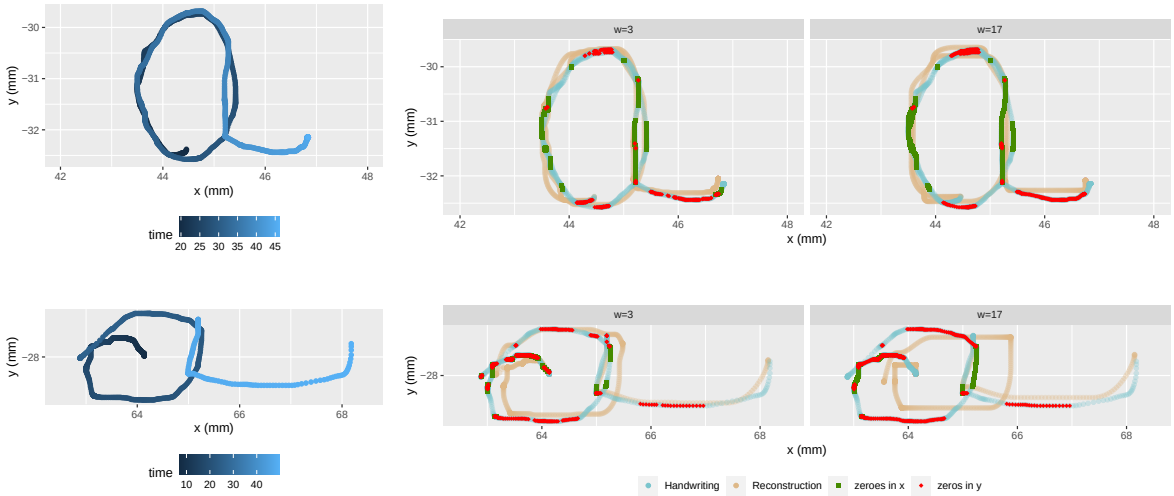


Figure 1: **Letter "a" written by a child in second grade without notable writing trouble (a) and a child diagnosed with dysgraphia (b).** Left: the recorded handwriting where color represents time; Middle (resp. Right): the original signal (light blue), the reconstructed signal (light orange) and moments of zero velocity in x (green square) and in y (red cross) when the bandwidth of closing operator is $w = 3$ (resp. $w = 17$)

(starting from the bottom, one and a half circle in clockwise direction before drawing the tail), while the (diagnosed) child on bottom row wrote the letter in a more classical way (with two strokes). The other panels overlay the reconstructed trace (light orange) on top of the original trace (light blue), with zeros displayed for both x (green square) and y (red cross). In the middle panels, with $w = 3$, the numbers of zeros are 20 in x, 26 in y for the first child, 23 and 26 for the second. With $w = 17$ in the right panel, these numbers decrease to 8 and 9 for the first child, 7 and 9 for the second. Although the numbers of zeros are similar given the structuring element size, it can be noticed that the reconstruction with a lower number of zeros moves away drastically from the original trace for the dysgraphic child (Fig. 1 bottom right) and remains correctly estimated for the non-diagnosed child (Fig. 1 top right). The average distance (in millimeters) between the original and reconstructed traces increases from 0.1294 to 0.1509 (+16%) for the non-diagnosed child and from 0.2006 to 0.5006 (+149%) for the dysgraphic child.

4 Number of zeros in x and y decreases when age of participants increases

In this section, we study the relationship between the age of participants and the number of zeros. To study performance at age $a \in [6.5, 12]$ years (first to sixth grade, CP to sixième in the French education system) and consider a sufficient amount of data, children with age $a \pm 3$ months are considered. Then, as we expect the number of zeros to decrease and plateau when size w increases, we simply aggregate across children the number of zeros obtained for different values of w in a wide enough range (in $\{3, 5, \ldots, 23\}$) for a given symbol and age range. Numbers of zeros corresponding to the aforementioned plateaus should be over-represented in the resulting distribution. Given that most of the dataset corresponds to non-dysgraphic children, this approach allows to parsimoniously extract the most representative —and possibly meaningful— number of zeros.

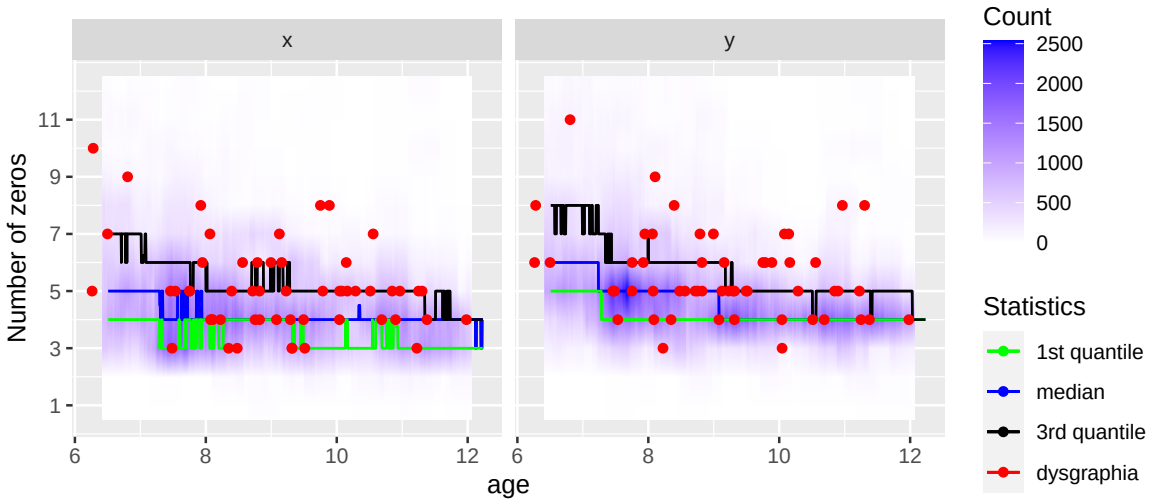


Figure 2: **Density and statistics of the number of zeros on x and y-axis against age, for letter "a".** For non-diagnosed children, density (blue shade), $Q1$ (green), median (blue line) and $Q3$ (black) of the number of zeros in x (left) and in y (right), aggregated across values of w; For individuals with dysgraphia, median (red points).

If we take the symbol "a" as an example (Fig. 2), the three quartiles for the number of zeros decreases as the age of participants increases. Dispersion also decreases with age, as reflected by the density. Each red point represents the median of the numbers of zeroes (aggregated across sizes w) for each dysgraphic child, when writing letter "a". Compared to non-diagnosed children, it can be observed that there are more red points above the blue line (median for non-diagnosed children) than below. The associated children therefore stop or oscillate more than the average non-diagnosed child of the same age.

5 Discussion

Preliminary results presented in the previous two sections encourage us to develop a classification algorithm for dysgraphia detection based on POMH. Once adequate zero velocity moments are chosen, POMH can reconstruct the handwriting trace for a non-diagnosed participant but deviates from the original signal for individuals diagnosed with dysgraphia (Fig. 1). Therefore, we believe that the reconstruction error made by POMH can help discriminate children with dysgraphia.

The number of zero velocity moments therefore constitutes a critical parameter to exploit POMH reconstruction error, and both depend on the symbol written and age of the participant to classify. For a given symbol in our dataset, we demonstrated that the number of zeros decreases and converges when age increases. One way of including age into the classification algorithm is to adjust the number of zeros as a function of age (e.g., through a regression model).

Once the number of zeros given the symbol and age of the participant is set, the corresponding moments of zero velocity must be determined before applying POMH. We plan to create an inverse function to find the size w from the number of zeros and then apply the mathematical morphology parameterized by w on the recorded handwriting trace to find the moments; or we may try to directly maximize POMH fit by adjusting the positions of the fixed number of zeros (e.g., using dynamic programming).

Bibliographie

André, G., Kostrubiec, V., Buisson, JC. et al. A parsimonious oscillatory model of handwriting. *Biol Cybern* 108, 321–336 (2014). https://doi.org/10.1007/s00422-014-0600-z

Asselborn, T., Gargot, T., Kidziński, Ł. et al. Automated human-level diagnosis of dysgraphia using a consumer tablet. *npj Digital Med* 1, 42 (2018). https://doi.org/10.1038/s41746-018-0049-x

M. Charles, R. Soppelsa, et J.-M. Albaret. Bhk : échelle d'évaluation rapide de l'écriture chez l'enfant. *Ecpa*, 2004.

Deschamps, L., Gaffet, C., Aloui, S. et al. Methodological issues in the creation of a diagnosis tool for dysgraphia. *npj Digit. Med.* 2, 36 (2019). https://doi.org/10.1038/s41746-019-0114-0

Devillaine, L.; Lambert, R.; Boutet, J.; Aloui, S.; Brault, V.; Jolly, C.; Labyt, E. Analysis of Graphomotor Tests with Machine Learning Algorithms for an Early and Universal Pre-Diagnosis of Dysgraphia. *Sensors* 2021, 21, 7026. https://doi.org/10.3390/s21217026

Hamstra-Bletz, L., J. DeBie, and B. P. L. M. Den Brinker. Concise evaluation scale for children's handwriting. *Lisse: Swets*, 1 (1987): 623-662.

Hollerbach, J. M. "An oscillation theory of handwriting." *Biological cybernetics* 39.2 (1981): 139-156.

Prévision probabiliste fondée sur l'échangeabilité d'un modèle d'ensemble en environnement.

Éric PARENT [1] & Jacques BERNIER [2]

[1] *UMR MIA-Paris, AgroParisTech, INRA, Université Paris-Saclay, 75005, Paris*
eric.parent@agroparistech.fr
[2] *Ingénieur EDF à la retraite*
jacques.bernier2@orange.fr

Résumé. La prévision *d'ensemble* se fonde sur des scénarios obtenus en modifiant les conditions et les paramètres initiaux d'un code déterministe dynamique décrivant un mécanisme physique (modèles météorologiques régionaux, modèles climatiques, modèles de précipitation et de transport solide ...). Ces scénarios variés traduisent l'espoir du physiciens de représenter l'incertitude due à l'état initial du système (par exemple la température de l'atmosphère, l'humidité du sol, etc.) ou découlant de la connaissance incomplète des paramètres. En propageant les perturbations des conditions limites à travers le code numérique intégro-différentiel complexe simulant le comportement du système, on obtient un ensemble de trajectoires hypothétiques de la variable à prédire (par exemple, la température de l'aéroport, la pluie, le débit des cours d'eau...) également connues sous le nom de *membres*. Dans quelle mesure les informations véhiculées par les différents membres de l'ensemble peuvent-elles aider à construire une prévision probabiliste de la quantité à prévoir? La littérature statistique fourmille de méthodes prêtes à l'emploi, connues sous le nom de techniques de "post-traitement" des ensembles, mais peu d'entre elles traitent formellement des spécificités des ensembles : (1) la trajectoire de la Nature peut ne pas être exactement celle générée par le code déterministe, (2) les membres de l'ensemble proviennent de processus physiques, ce qui devrait permettre un certain apprentissage statistique, (3) les membres de chaque ensemble ont tendance à être, à cause du principe même de leur construction, l' échantillonnage d'un modèle de type *échangeable*.

En statistique, deFinetti et ses successeurs ont démontré que la propriété d'échangeabilité conduit à un modèle à effet aléatoire. L'effet aléatoire résume parcimonieusement l'information véhiculée par les membres de l'ensemble vis à vis la variable à prédire. Nous construisons un modèle normal échangeable pour les températures et suggérons d'autres développements bayésiens hiérarchiques *ad hoc* : données environnementales avec valeurs nulles (précipitations) ou structures conjointes de séries hydro-météorologiques (pluies-débits). Les performances des modèles proposés sont vérifiées et comparées à celles d'autres techniques de post-traitement sur une longue série d'enregistrements de températures. Les résultats montrent la robustesse des structures prédictives parcimonieuses construites, ce qui plaide en faveur de l'échangeabilité et de la cohérence probabiliste lors de la modélisation des membres d'un ensemble environnemental.

Mots-clés. Modèle bayésien hiérarchique, Prévision Probabiliste, Post-traitement d'ensemble environmental, échangeabilité ...

Abstract. *Ensemble* forecast is based on scenarios obtained by modifying the initial conditions and parameters of a dynamic deterministic code describing a physical mechanism (regional weather models, climate models, precipitation and solid transport models ...). These various scenarios reflect the physicists' hope to represent the uncertainty due to the poorly detremined initial state of the system (e.g. atmospheric temperature, soil moisture, etc.) or resulting from incomplete knowledge of the parameters. By propagating the boundary condition perturbations through the complex integro-differential numerical code mimicking the system behaviour, a bunch of hypothetical trajectories of the variable to be predicted (e.g. airport temperature, rainfall, river flow...) also known as *members of the ensemble* is obtained. To what extent can the information conveyed by the different members of the ensemble help to construct a probabilistic forecast of the quantity to be predicted? Statistical literature offers a wealth of off-the-shelf methods known as "post-processing" techniques for ensembles, but few of them formally address the specific features of an ensemble: (1) the trajectory of Nature may not be exactly the one generated by the deterministic code, (2) the members of the set come from physical processes, which should allow some statistical learning, (3) the members of each ensemble tend to be, because of the very principle of their construction, the sampling of an *exchangeable* model.

In statistics, deFinetti and his successors have shown that the property of exchangeability leads to a random-effect model. The random effect parsimoniously sums up the information conveyed by the members of the ensemble with regard to the variable to be predicted. We construct a normal exchangeable model for temperatures and suggest other *ad hoc* hierarchical Bayesian developments: environmental data with zero values (precipitation) or joint structures of hydro-meteorological series (rain-flow). The performances of the proposed models are checked and compared with those of other post-processing techniques over a long series of temperature records. The results show the robustness of such statistical structures, which argues in favour of exchangeability and probabilistic consistency when modelling the members of an environmental ensemble.

Keywords. Hierarchical Bayesian model, Probabilistic forecasting, Ensemble post-processing techniques, Exchangeability ...

1 Construction d'une prévision probabiliste fondée sur un ensemble environnemental

Dans le contexte de la prévision d'ensemble, le problème posé, au départ à tout opérateur prévisionniste est la modélisation probabiliste adaptée à la structure des S membres de

l'ensemble $X = \{x_1, x_2, ...x_S\}$ et le passage à la distribution conditionnelle de la cible Y.

Comme dans Courbariaux *et al* (2019), nous travaillons sur les $S = 50$ membres de températures de l'ensemble issu du Centre Européen pour les Prévisions Météorologiques à Moyen Terme pour la station de Vouglans, dans l'Ain. Commençons par distinguer trois types d'échantillons porteurs d'informations:

1. l'échantillon *climatique* (souvent long) où on dispose d'observations de la cible Y, la grandeur température journalière moyenne observée (c'est un long historique journalier (1953 à 2015), d'où son importante pour la calibration).

2. l'échantillon de *calage* où les modèles marginaux des ensembles X puis le modèle prédictif de $Y|X$ sont estimés (on observe conjointement X et Y de 2005 à 2008, puis de 2011 à 2015),

3. l'échantillon de *vérification- validation* des résultats de la prévision (qui pourra être constitué de sous périodes à extraire de la période conjointe, le reliquat restant pour le calage).

2 Modélisation marginales de l'ensemble X en calage

Notre hypothèse de travail minimale s'appuie sur une certaine symétrie des membres; l'échangeabilité traduira l'invariance de la probabilité de l'ensemble par rapport à toute permutation des S membres. La symétrie acceptée impose, bien sûr, le respect de la stationnarité des séries à traiter, éventuellement obtenue préalablement par une procédure *ad hoc* de stationnarisation.

Le théorème de deFinetti (1939), dit de représentation après son extension en 1955 par Savage et Hewitt, stipule que toute séquence *infinie* est échangeable **si et seulement si** il existe "une variable aléatoire Z" telle que l'on a en probabilité, en utilisant la notation *crochet* de Gelfand et Smith (1939) pour les densités :

$$[x_1, ..., x_S] = \int_Z \prod_{s=1}^{S} [x_s|Z][Z]dZ$$

On notera que la réciproque, *seulement si*, s'applique immédiatement même pour tout échantillon fini : tout échantillon d'un *mélange* est échangeable.

Précisons que Z n'est pas n'importe quelle variable aléatoire, elle est étroitement liée à l'échangeabilité des membres, exprimant leur structure particulière et dépend de l'ensemble. Reprenons Bernardo (1996) : *Le théorème de représentation, résultat de pure théorie de probabilité, prouve que, si les observations sont jugées échangeables, alors elles doivent vraiment être un échantillon aléatoire d'un modèle et une distribution a priori sur le paramètre du modèle* (le Z), (imposant de fait une approche bayesienne). Il faut cependant noter que le théorème de représentation est un théorème d'existence: en

géneral il ne spécifie pas le modèle et ne spécifie jamais la distribution a priori requise. Les hypothèses additionnelles qui sont habituellement nécessaires pour spécifier un modèle particulier sont décrites dans les applications. Ici, comme souvent pour les températures, nous invoquons la normalité et proposons le modèle à effets aléatoires construit sur 2 pivôt réels Z_{1t} et Z_{2t}:

$$\begin{aligned} X_{ts} &= \alpha + \beta Z_{1t} + \lambda Z_{2t}^{-0.5} \times \epsilon_{ts} \\ \epsilon_{ts} &\sim N(0,1) \end{aligned} \tag{1}$$

On notera que α, β, λ sont les paramètres de niveau 2 de la structure hiérarchique (cf Bernardo), il faut leur joindre les paramètres des modèles latents des pivôts. Pour des raisons de simplicité statistique, on adoptera l'hypothèse de conjuguées naturelles (Raiffa et Schlaifer, 1961); soit , en notant G la loi gamma:

$$Z_{2t} \sim G(g,1) \tag{2}$$

$$Z_{1t}|Z_{2t} \sim N(0, Z_{2t}^{-0.5}) \tag{3}$$

Ces trois équations définissent un modèles marginal d'ensemble échangeable compatible avec la représentation issue du théorème de deFinetti.

Si, en première analyse on prend pour exemple la période de calage des dennées de température journalières désaisonnalisées (mai-novembre) à Vouglans, on a un échantillon de taille $n > 850$, ce qui nous conduit, d'une part, à pouvoir utiliser des priors non informatifs, et d'autre part, à négliger les incertitudes a posteriori sur les paramètres en les posant égaux à leurs moyennes a posteriori dans les inferences de modèles.

Comme il se doit pour ce modèle normal, la loi de $Z = (Z_1, Z_2)$ conditionnelle à X possède une statistique exhaustive $T(X_t)$ à t fixé : le couple $T(x_t = (\bar{x}_t, s_t^2)$ (moyenne et variance empirique), observable sur chaque ensemble et qui les remplace comme pivôt $Z = (Z_1, Z_2)$ dans toute équation de prévision. Lauritzen (2007) les appelent *résumantes des latentes Z* dans l'expression du théorème de deFinetti. On en déduit les lois marginales conditionnelles de type Fisher (F) et Student (T) pour les composantes du résumé de l'ensemble :

$$\left(\frac{S.s_t^2}{\lambda^2}\right) \sim F(S-1, 2g) \tag{4}$$

$$\left(\frac{\bar{x}_t - \alpha}{k \times s_t}\right) \sim T(S-1) \tag{5}$$

$$\text{avec } k = \left(\sqrt{\frac{\beta^2 + S\lambda^2}{\lambda^2}}\right)$$

3 Modélisation du lien entre la cible Y et l'ensemble X en calage

Pour la phase prédictive, il nous faudra expliciter $[Y|X] = \int [Y|Z][Z|X]dZ$ où l'on peut remplacer les pivôts Z par les statistiques résumantes $T(X)$. On suppose bien sûr que pour chaque temps t, X_t, (qu'on pourrait qualifier d'exhaustif vis à vis de Y) est censé contenir toutes les informations nécessaires à la prévision.

On cherche donc à construire un modèle bivariable $[Y, T(X)]$ qui doit encapsuler les propriétés marginales des résumantes d'ensemble d'un côté et des propriétés climatiques de Y (issues de la calibration) de l'autre côté.

On suggère d'étendre ici une méthode de *copule* normale inspirée de Krzysztofowicz (1997) , en supposant que le résumé statistique s_t^2, la variance de l' ensemble, est un prédicteur plus ou moins relié à la variance de la prédictive. Nous proposons un modèle sur les grandeurs transformées U, V, W à deux paramètres θ et ψ à estimer sur la période de calage, sous la forme:

$$\begin{aligned} V &= \frac{Y - \mathbb{E}(Y)}{\sqrt{\mathbb{V}(Y)}} \sim N(0,1) \\ U &= N\left(T_{S-1}^{-1}\left(\frac{\bar{x}_t - \alpha}{k \times s_t}\right)\right) \\ W &= G_{g,1}\left(F_{S-1,2g}^{-1}\left(\frac{S.s_t^2}{\lambda^2}\right)\right) \\ V_t &= \theta \times U_t + \psi \times \sqrt{W_t} \times \epsilon_t \\ \epsilon_t &\sim N(0,1) \end{aligned} \tag{6}$$

Pour W_t, on pourrait également utiliser une autre transformation d'un prédicteur positif de variance d'erreur de prévision, par exemple une loi normale tronquée sur la demi-droite $0 < W < +\infty$

4 Conclusion

Comparé aux approches classiques en matière de prévisions d'ensemble de Jolliffe *et al* (2012) , le modèle échangeable pour les ensembles environnementaux repose sur une propriété statistique constructive dont le réalisme opérationnel est facilement défendable : l'échangeabilité des trajectoires possibles des grandeurs météorologiques est justement la propriété que souhaitent les physiciens qui établissent les membres de l'ensemble! Quoique son adaptation à certains types de données reste un challenge (le traitement séparé des valeurs nulles des précipitations pose un délicat problème d'introduction de variables latentes, par exemple), les premiers essais de ce modèle présentent des aspects prometteurs

de robustesse et de parcimonie. Il peut sans doute d'ailleurs être perfectionnés de multiples façons: introduction de mémoire autorégressive marginale des latentes d'ensemble, descripteurs d'information supplémentaires au niveau prédictif, adaptabilité par apprentissage séquentiel régulièrement réactualisé, etc.

Remerciements

Ce travail fait suite à la thèse de Marie Courbariaux financée par EDF et Hydro-Québec (contrat AgroParisTech/Innovation 694R) et soutenue en 2016. Les réflexions méthodologiques de cette communication ont été initiées par les échanges soutenus durant cette période et par la suite avec les cadres des services de recherche développement de ces deux sociétés productrices d'hydro-électricité: Rémy Garcon, Luc Perreault et Joël Gailhard.

Bibliographie

Bernardo, J. M. (1996). The concept of Exchangeability and its applications. *Far East Journal of Mathematical Sciences*, 4:111–122.

Courbariaux, M., Barbillon, P., Perreault, L., et Parent, É. (2019). Post-processing multiensemble temperature and precipitation forecasts through an exchangeable normal-gamma model and its tobit extension. *Journal of Agricultural, Biological and Environmental Statistics*, 24(2):309–345.

deFinetti, B. (1937). La prévision: ses lois logiques, ses sources subjectives. In *Annales de l'institut Henri Poincaré*, volume 7, pages 1–68.

Gelfand, A. E. et Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410):398–409.

Jolliffe, I. T. et Stephenson, D. B. (2012). *Forecast verification: a practitioner's guide in atmospheric science*. John Wiley & Sons.

Hewitt E et Savage L. J. (1955). Symetric measures on cartesian masures in Trans. Amer. Math. Soc. 80.

Krzysztofowicz, R. (1997). Transformation and normalization of variates with specified distributions. *Journal of Hydrology*, 197(1):286–292.

Lauritzen S. (2007) Exchangeability and deFinetti's Theorem. http://www. stats. ox. ac. uk/˜ steffen/teaching/grad/definetti.pdf.

Schlaifer, R. et Raiffa, H. (1961). *Applied statistical Decision Theory*. Division of Research, Harvard Business School, Boston, MA.

Apprentissage non supervisé

Clustering et détection d'anomalies dans les données fonctionnelles

Martial AMOVIN-ASSAGBA[1,2], Irène GANNAZ[3] & Julien JACQUES[1].

[1] *Univ Lyon, Univ Lyon 2, ERIC EA3083, Lyon, France, julien.jacques@univ-lyon2.fr*
[2] *Arpege Master K, Saint-Priest, France, martial.amovin@masterk.com.*
[3] *Univ Lyon, INSA de Lyon, CNRS UMR 5208, Institut Camille Jordan, F-69621 Villeurbanne, France, irene.gannaz@insa-lyon.fr*

Résumé. Ce travail propose une méthode permettant simultanément de grouper en clusters et de détecter des anomalies dans des données fonctionnelles multivariées. Les données fonctionnelles sont décomposées dans une base de fonctions de dimension finie. La méthode repose ensuite sur des modèles de mélanges gaussiens contaminés parcimonieux sur cette décomposition. Un algorithme ECM est utilisé pour l'inférence du modèle. La performance du modèle est illustrée sur des données simulées.

Mots-clés. Détection d'anomalies, clustering, données fonctionnelles contaminées.

Abstract. This work proposes a method to group simultaneously into clusters and detect anomalies in multivariate functional data. Functional data are decomposed into a finite-dimensional basis functions. The method is then based on parcimonious contaminated Gaussian mixtures models on this decomposition. An ECM algorithm is used for model inference. The performance of the model is illustrated on simulated data.

Keywords. Anomaly detection, clustering, contaminated functional data.

1 Introduction

Les nouvelles méthodes d'acquisition de données favorisent la collecte massive des données avec une fréquence très élevée. Une manière de gérer ces amas de données est de les considérer comme des données fonctionnelles (Ramsay et Silverman, 2005), les observations étant assimilées à des courbes. Un des défis majeurs dans l'analyse de données fonctionnelles est la détection d'anomalies. Un certain nombre de travaux ont été proposés pour détecter des anomalies dans les données fonctionnelles univariées et multivariées. Certains travaux sont basés sur des approches de type kmeans robustes (Garcia-Escudero et Gordaliza, 2005), sur des fonctions de profondeur (Hubert et al. 2017), ou sur des techniques d'isolement (Staerman et al. 2019). Un point commun de ces méthodes est qu'elles requièrent toutes une connaissance a priori de la proportion des données anormales.

La détection de données anormales peut être grandement affectée par la présence de plusieurs clusters distincts au sein des données. Une approche intéressante est alors de

réaliser simultanément un clustering des données et une détection de données anormales. Dans cette optique, (Punzo et McNicholas, 2016) proposent, dans un cadre non fonctionnel, un modèle de mélange gaussien contaminé dans lequel chaque composante du mélange contient des données normales et/ou anormales.

Dans cet article, en s'inspirant de (Punzo et McNicholas, 2016) et sous l'hypothèse que les données fonctionnelles peuvent être décomposées dans une base de fonctions de dimension finie, nous proposons un modèle de mélange gaussien contaminé pour les données fonctionnelles multivariées. En plus de grouper les données en différents clusters, cette méthode permet de détecter des anomalies dans les données fonctionnelles. Dans les sections suivantes, nous présentons le modèle mathématique, son inférence, et des résultats obtenus sur simulations.

2 Modélisation

Soit $X = X(t), t \in [0, T]$, une variable aléatoire fonctionnelle de dimension p. Soient $X_i(t)_{\{1 \leq i \leq n\}}$, n réplications indépendantes de la variable aléatoire fonctionnelle X et $x_i(t) = (x_i^1(t), x_i^2(t), \ldots, x_i^p(t))$, une réalisation de $X_i(t)$, $1 \leq i \leq n$. L'objectif est à la fois de grouper ces données en différents clusters et de détecter des données anormales.

2.1 Reconstruction fonctionnelle des données et description du modèle

En pratique, les expressions fonctionnelles des courbes observées ne sont pas connues et les données à disposition sont des observations discrètes enregistrées sur un nombre fini de points. Il est donc nécessaire de reconstruire la forme fonctionnelle des données. Pour ce faire, nous décomposons les courbes x_i^j dans une base de fonctions $\psi_m^j \in L_2, 1 \leq m \leq M_j$:

$$x_i^j(t) = \sum_{m=1}^{M_j} c_{im}^j \psi_m^j(t),$$

où M_j est le nombre de fonctions de base par composante et $\psi_m^j, 1 \leq m \leq M_j$ est la m-ème composante de la base de fonctions ψ^j. Le choix de la base de fonctions ainsi que du nombre de fonctions dépendent de la nature des données. Posons c_i le vecteur concaténant tous les coefficients de l'observation i obtenus pour chaque composante : $c_i = (c_i^1, \ldots, c_i^p)$ avec $c_i^j = (c_{i1}^j, \ldots, c_{iM_j}^j)$. Chaque donnée x_i peut être écrite sous la forme :

$$x_i(t) = \Psi(t) c_i^T \text{ avec } \Psi(t) = \begin{pmatrix} \psi_1^1(t) & \cdots & \psi_{M_1}^1(t) & 0_{M_2} & \ldots & 0_{M_p} \\ 0_{M_1} & & \psi_1^2(t) & \cdots \psi_{M_2}^2(t) & \ldots & 0_{M_p} \\ \vdots & & \vdots & & \ddots & \vdots \\ 0_{M_1} & & 0_{M_2} & & \ldots & \psi_1^p(t) \cdots \psi_{M_p}^p(t) \end{pmatrix}.$$

Les modèles de mélanges gaussiens contaminés (Punzo et McNicholas, 2016) supposent que la densité de chaque cluster est un mélange de deux distributions gaussiennes multivariées dont l'une représente les observations normales et l'autre les données anormales. Cependant, de manière générale, la notion de densité de probabilité n'existe pas pour les variables aléatoires fonctionnelles. (Delaigle et Hall, 2010) ont montré que cette densité peut être approchée par la densité de probabilité des scores de l'Analyse en Composantes Principales Fonctionnelle (ACPF). Plusieurs travaux définissent alors des algorithmes de clustering probabilistes en considérant des modèles de mélange gaussiens sur ces scores (Schmutz et al. 2018). A noter que cette hypothèse est équivalente à supposer un modèle de mélange gaussien sur les coefficients c_i. De ce fait, la densité du mélange peut être écrite sous la forme :

$$h = \sum_{k=1}^{K} \pi_k \left[w_k f(c_i, \mu_k, \Sigma_k) + (1 - w_k) f(c_i, \mu_k, \eta_k \Sigma_k) \right] \tag{1}$$

où K est le nombre total de clusters, π_k est la probabilité d'appartenir au cluster k, w_k est la proportion de données normales du cluster k, μ_k et Σ_k sont respectivement la moyenne et la covariance du cluster k, η_k le facteur d'inflation de la variance pour les données anormales, et $f(\cdot, \mu_k, \Sigma_k)$ la densité gaussienne.

L'irrégularité des courbes observées peut entraîner l'utilisation de bases de fonctions ψ^j de relativement grande dimension, et ainsi une augmentation du nombre de paramètres à estimer, notamment du fait des matrices de covariance Σ_k. Pour contrer le fléau de la dimension, nous proposons d'utiliser la technique de réduction de dimension de (Schmutz et al. 2020), qui proposent de projeter les données de chaque cluster dans un sous-espace latent fonctionnel de faible dimension $d_k < M = \sum_{j=1}^{p} M_j$. À l'aide d'une ACPF, l'expression des fonctions propres de chaque cluster peut être obtenu à partir des fonctions ψ_m :

$$\varphi_{km}(t) = \sum_{l=1}^{M} s_{kml} \psi_l(t)$$

où s_{kml} sont les coefficients des fonctions propres. Les d_k premières composantes de la variance du cluster k, dans l'espace engendré par les fonctions propres, sont modélisées de manière fine par les d_k premières valeurs propres. Les $M - d_k$ dernières composantes sont modélisées de façon parcimonieuse par un seul paramètre b_k. La variance du cluster k dans l'espace propre s'écrit alors $\text{diag}(a_{k1}, \ldots, a_{kd_k}, b_k, \ldots, b_k)$.

2.2 Inférence du modèle

Un moyen usuel pour estimer les paramètres d'un modèle de mélange est l'algorithme EM. C'est un algorithme itératif qui alterne deux étapes: l'étape E (*Expectation*) qui calcule l'espérance de la log-vraisemblance complétée par les variables latentes, et l'étape M (*Maximization*) qui maximise cette espérance.

Deux variables latentes sont introduites dans le modèle. La première est la variable z_{ik}. $z_{ik} = 1$ si la donnée multivariée x_i appartient au cluster k et 0 sinon. La seconde variable introduite notée v_{ik} est en lien avec la présence de données anormales, $v_{ik} = 1$ si la donnée x_i du cluster k est une donnée normale et 0 sinon. L'expression de la log-vraisemblance complétée est alors :

$$\begin{aligned}\ell_c(\Phi) &= \sum_{i=1}^{n}\sum_{k=1}^{K}\{z_{ik}\log\pi_k + z_{ik}[v_{ik}\log w_k + (1-v_{ik})\log(1-w_k)]\} \\ &= -\frac{1}{2}\sum_{k=1}^{K}\left\{M(n_k - n_{k_n})\log\eta_k + n_k\left(\sum_{l=1}^{d_k}\log(a_{kl}) + \sum_{l=d_k+1}^{M}\log(b_k)\right)\right. \\ &\quad \left. + n_k tr\left[\sum_{l=1}^{d_k}\frac{s'_{kl}W^{\frac{1}{2}}R_kW^{\frac{1}{2}}s_{kl}}{a_{kl}} + \sum_{l=d_k+1}^{B}\frac{s'_{kl}W^{\frac{1}{2}}R_kW^{\frac{1}{2}}s_{kl}}{b_k}\right]\right\} - \frac{nM}{2}\log(2\pi)\end{aligned}$$

où $n_k = \sum_{i=1}^{n} z_{ik}$ et $n_{k_n} = \sum_{i=1}^{n} z_{ik}v_{ik}$ sont respectivement le nombre de données et le nombre de données normales du cluster k, $R_k = \frac{1}{n_k}\left(\sum_{i=1}^{n} z_{ik}\left(v_{ik} + \frac{1-v_{ik}}{\eta_k}\right)(c_i - \mu_k)'(c_i - \mu_k)\right)$ est la matrice de covariance empirique du cluster k, W est la matrice du produit scalaire entre les fonctions de base Ψ et $\Phi = \{\Phi_1, \Phi_2\}$ avec $\Phi_1 = \{\pi_k, w_k, \mu_k, a_k, b_k, S_k\}_{k=1}^{K}$ et $\Phi_2 = \{\eta_k\}_{k=1}^{K}$. S_k est la matrice dont la $l-$ème colonne est s_{kl}.

En maximisant $\ell_c(\Phi)$, l'expression de l'estimation de Φ_2 est fonction de Φ_1 et vice versa. Nous proposons alors d'utiliser l'algorithme ECM (Expectation Conditional Maximization) où chaque étape M de EM est remplacée par deux étapes M. À la première étape M, nous fixons Φ_2, nous réalisons une ACPF par cluster et nous estimons les paramètres Φ_1; à la seconde étape M, nous fixons Φ_1 et nous estimons Φ_2.

3 Expériences

Dans cette section, nous présentons les données simulées et les résultats des expériences que nous avons réalisées sur ces données.

3.1 Données

Un jeu de données bivarié composé de 3 clusters a été simulé. Les proportions de mélange sont identiques $\pi_1 = \pi_2 = \pi_3 = \frac{1}{3}$. Pour simplifier, nous considérons que pour chaque classe, la variance est commune pour toutes les dimensions intrinsèques d_k, c'est à dire $a_{kj} = a_k$ pour $k = 1, \ldots, K$ et $j = 1, \ldots, d_k$. L'objectif de ces expériences est d'une part d'évaluer l'efficacité de l'algorithme ECM en retrouvant les paramètres a_k, b_k et d_k, et d'autre part d'évaluer la performance de notre modèle du point de vue du clustering et de la détection d'anomalies. Nous fixons $a_1 = 150, a_2 = 100, a_3 = 50$, $b_1 = 5, b_2 =$

$3, b_3 = 10$ $\eta_1 = 30, \eta_2 = 50$ et $\eta_3 = 1$. Les dimensions intrinsèques de chaque cluster sont $d_1 = 5, d_2 = 20, d_3 = 10$. $\mu_1 = (1, 0, 50, 100, 0, \ldots, 0), \mu_2 = (0, 0, 80, 0, 40, 2, 0, \ldots, 0), \mu_3 = (0, \ldots 0, 20, 0, 80, 0, 0, 100)$. La proportion des données anormales est 5‰. La forme fonctionnelle des données est reconstruite à partir de 35 fonctions de Fourier. La Figure 1 présente les données simulées.

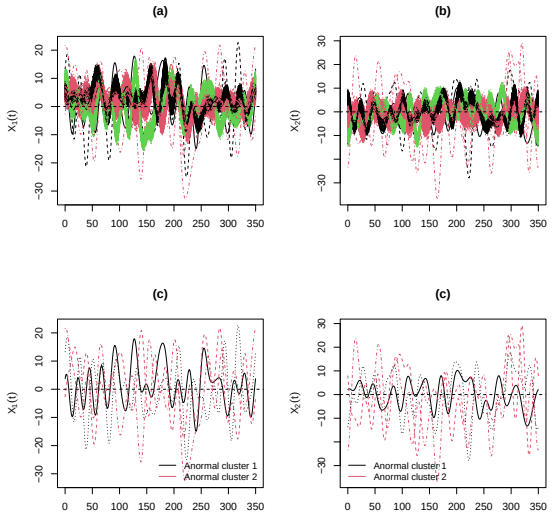


Figure 1: En haut : Données simulées. La première composante est à gauche et la seconde à droite. Chaque courbe est colorée selon son groupe. En bas : Courbes anormales. Les couleurs des courbes anormales diffèrent selon leur cluster d'attachement : noire pour les courbes anormales du groupe 1 et rouge pour celles du groupe 2.

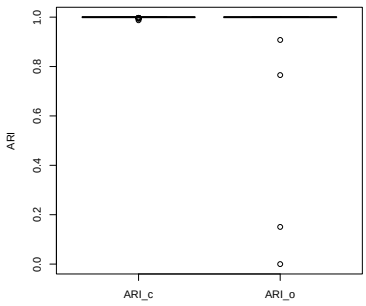


Figure 2: Performance du clustering (ARI_c) et de la détection d'anomalies (ARI_o).

3.2 Résultats

Nous fixons $K = 3$ mais notons que K peut être déterminé en considérant plusieurs valeurs et en choisissant le meilleur par BIC (Amovin et al. 2020). Les dimensions intrinsèques des clusters sont obtenues par le test de Cattell avec un seuil fixé à 0.2. Nous simulons 50 jeux de données de taille 1000. La Table 1 donne les moyennes et écarts-types des estimations des paramètres du modèle. La performance du modèle en terme de clustering et de détection d'anomalie est évaluée par l'indice de Rand (*Adjusted Rand index*, noté ARI) (Figure 2). Les résultats sont très bons, l'algorithme détectant toutes les données anormales pour 92% des simulations, sans faux positifs.

Paramètres	Groupe 1	Groupe 2	Groupe 3
a_k	150.45 (5.72)	100.22 (2.23)	52.13 (1.27)
b_k	4.91 (0.06)	2.80 (0.03)	9.59 (0.09)
d_k	5 (0)	20 (0)	10 (0)

Table 1: Moyenne (et écart type) des estimations des paramètres du modèle sur 50 simulations.

4 Conclusion

Dans ce travail, nous proposons une technique de clustering pour détecter des anomalies dans les données fonctionnelles multivariées. Les résultats obtenus à travers ce modèle sur données simulées présentent de bonnes performances. Ces résultats sont confirmés sur données réelles. Pour une mise en oeuvre opérationnelle d'un point de vue industriel de cet algorithme, nous pensons étendre ce modèle à une version *online* permettant la mise à jour des paramètres au fur et à mesure que de nouvelles données arrivent.

Bibliographie

Amovin-Assagba, M., Gannaz, I., & Jacques, J. (2021). Outlier detection in multivariate functional data through a contaminated mixture model. *arXiv preprint arXiv:2106.07222*.

Garcia-Escudero, L. A., & Gordaliza, A. (2005). A proposal for robust curve clustering. Journal of classification, 22(2), 185-201.

Hubert, M., Rousseeuw, P., & Segaert, P. (2017). Multivariate and functional classification using depth and distance. Advances in Data Analysis and Classification, 11(3), 445-466.

Punzo, A., & McNicholas, P. D. (2016). Parsimonious mixtures of multivariate contaminated normal distributions. *Biometrical Journal*, 58(6), 1506-1537.

Ramsay, J. O., & Silverman, B. W. (2005). Functional data analysis. Springer series in statistics.

Schmutz, A., Jacques, J., Bouveyron, C., Cheze, L., & Martin, P. (2020). Clustering multivariate functional data in group-specific functional subspaces. *Computational Statistics*, 35(3), 1101-1131.

Staerman, G., Mozharovskyi, P., Clémençon, S., & d'Alché-Buc, F. (2019, October). Functional isolation forest. In Asian Conference on Machine Learning (pp. 332-347). PMLR.

Clustering spectral adaptatif

Ilaria Giulini [1], Aurélie Fischer [1], Mathilde Mougeot [2]

[1] *LPSM, Université de Paris, 8 place Aurélie Nemours, 75013 Paris, aurelie.fischer@lpsm.paris, ilaria.giulini@lpsm.paris*
[2] *Centre Borelli, ENS Paris-Saclay, 4 Av. des Sciences, 91190 Gif-sur-Yvette, mathilde.mougeot@ens-paris-saclay.fr*

Résumé. Nous nous intéressons au clustering spectral pour des réalisations i.i.d. d'une loi de probabilité inconnue dont le support est une union de composantes compactes connexes, dans un espace de Hilbert. Nous décrirons un algorithme, basé sur la modification de la procédure par Ng, Jordan et Weiss (2001) proposée dans Giulini (2016). Un grand avantage de la méthode est qu'elle permet d'estimer automatiquement le nombre de classes. Concernant les garanties théoriques, comme démontré dans Giulini (2016), la convergence de la procédure peut être caractérisée en termes de convergence d'opérateurs de Gram. Les performances de cette méthode adaptative seront illustrées par des expériences numériques approfondies sur des données simulées et des données réelles se présentant sous forme d'images.

Mots-clés. Clustering spectral, espace de Hilbert à noyau auto-reproduisant, nombre de classes adaptatif.

Abstract. We consider spectral clustering for i.i.d. realizations from an unknown probability distribution whose support is a union of compact connected components, in a Hilbert space. We will describe an algorithm, based on the modification of the procedure by Ng, Jordan and Weiss (2001) proposed in Giulini (2016). A particularly interesting feature is that the method allows to automatically estimate the number of clusters. From the theoretical point of view, as shown in Giulini (2016), the convergence of the procedure may be characterized in terms of convergence of Gram operators. The performance of the adaptive method will be illustrated through extensive numerical experiments on simulated data as well as real image data.

Keywords. Spectral clustering, reproducing kernel Hilbert space, adaptive number of clusters.

1 Introduction

La classification non supervisée ou partitionnement (*clustering*) consiste à identifier une structure de groupes au sein d'un ensemble d'observations (Duda, Hart, Stork, 2001). Contrairement à la classification supervisée, il n'existe pas d'étiquettes définies à l'avance

qu'on chercherait à affecter aux données. L'objectif est de regrouper des observations qui sont proches au sens d'une certaine notion de distance. Les groupes formés doivent être aussi homogènes et séparés que possible. Un enjeu important de la classification non supervisée consiste à sélectionner de façon appropriée le bon nombre de classes. Parmi les différentes méthodes de partitionnement, on trouve le clustering spectral, qui exploite les éléments propres d'une matrice laplacienne construite à partir d'une matrice de distances (voir, par exemple, von Luxburg, 2007). C'est ce type de méthode qui nous intéresse ici.

2 Méthodologie et résultats théoriques

Le point de départ dans l'article Giulini (2016) est l'algorithme de Ng, Jordan, Weiss (2001), interprété dans un espace de dimension infinie. Plus précisement, le clustering spectral est vu comme un changement de représentation dans un espace de Hilbert à noyau reproduisant. L'idée dans Giulini (2016) est de remplacer la projection sur l'espace engendré par les plus grands vecteurs propres de la matrice laplacienne, proposée par Ng, Jordan, Weiss (2001), par une certaine puissance de l'opérateur laplacien. En effet, une telle itération effectue une sorte de troncature lisse des valeurs propres et amène donc à une réduction naturelle de la dimension. Cela permet de présenter un algorithme qui estime automatiquement le nombre de classes.

Pour choisir automatiquement le nombre m d'itérations, il faut fixer un nombre de classes maximal, que l'on note M. En d'autres termes, on suppose une borne supérieure sur le nombre de classes. Ce choix de M est robuste, comme on peut le constater dans les simulations : M peut être surestimé sans que cela affecte les performances de l'algorithme. Une fois le bon nombre m d'itérations obtenu, les points sont regroupés selon la représentation induite par la puissance m-ième de la matrice laplacienne.

3 Algorithme

L'algorithme de Clustering Spectral Adaptatif implémenté en pratique, noté AKSC (*Adaptive Kernel Spectral Clustering* en anglais), est un algorithme récursif. A chaque itération, cet algorithme calcule dans un premier temps la représentation induite par la puissance m-ième de la matrice laplacienne compte tenu de la valeur supposée pour le nombre maximal de classes M puis, dans un deuxième temps, le partitionnement induit par cette représentation ainsi que le nombre de groupes associé c_m en fonction d'un seuil de proximité, S. La valeur supposée du nombre maximal de classes M est enfin mise à jour en utilisant la valeur c_m calculée précédemment. L'algorithme converge lorsque la valeur de c_m devient supérieure à M ou que l'hypothèse sur le nombre maximal de classes devient inférieure ou égale à deux. On peut montrer qu' à chaque itération la valeur de M décroit et que l'algorithme converge ainsi en un nombre fini d'itérations.

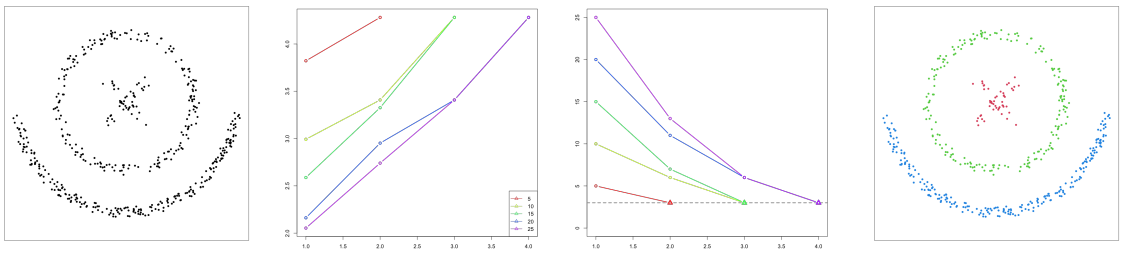
Algorithme AKSC

/* Initialisation des paramètres*/
$M \leftarrow 25$; $\xi \leftarrow 0.1$; $S \leftarrow 0.01$;
$A = \text{dist}(X)^2$; $\beta = \text{Trace.newton}(A)$; $A \leftarrow \exp(-\beta * A)$;
$A = U\Sigma U^t$;
$c \leftarrow M - 1$;
while $(c \leq M)$ & $(M \geq 2)$ **do**
  $m = \lfloor -\log \xi / \log(\lambda_2/\lambda_{M+1}) \rfloor$;
  $A = U\Sigma^m U^t$;
  $A \leftarrow \text{diag}(A^{-2})\ A\ \text{diag}(A^{-2})$;
  $(c_m, G_m) \leftarrow \text{Groupe}(A, S)$;
  $M \leftarrow c_m$
end
return c_m, G_m

La fonction *Groupe()* de l'algorithme calcule le nombre de groupes et fournit également les labels des groupes pour les observations (G) à un paramètre d'échelle près (S). La valeur du seuil S est directement liée à l'échelle souhaitée de détermination des groupes.

4 Résultats numériques

Nous présentons tout d'abord un exemple simple, créé à partir de données simulées, pour lequel le nombre de classes ainsi que le partitionnement est défini sans ambiguïté. La Figure 1 illustre les capacités de l'algorithme AKSC à converger vers la solution souhaitée (Fig 1.d), selon la représentation induite par la puissance m-ième de la matrice laplacienne (Fig 1.b) compte tenu de plusieurs hypothèses initiales sur le nombre maximal de groupes contenus dans les données (Fig 1.c).



(a) (b) (c) (d)

Figure 1: Données simulées (a). Évolution logarithmique du nombre d'itérations (b), et du nombre de classes (c) pendant le processus de convergence. Résultat de classification obtenu par l'algorithme AKSC (d).

Il faut souligner qu'ici le nombre de groupes attendus dans les données ne nécessite pas d'être instancié contrairement au cas de nombreux algorithmes de clustering. C'est par exemple le cas a priori pour l'algorithme K-means, le clustering basé sur les modèles de mélange gaussiens, la classification hiérarchique ascendante, même s'il existe pour ces algorithmes plusieurs stratégies permettant de calculer le nombre de classes.

En analyse d'images, les besoins en classification automatique d'objets sont nombreux d'un point de vue opérationnel, notamment pour répertorier et localiser les objets dans des images. Nous avons appliqué l'algorithme AKSC sur un ensemble d'images en noir et blanc numérisées correspondant à des planches d'un catalogue d'objets. Pour chaque image, le clustering est réalisé sur la position absolue dans l'image des pixels de couleur noire.

La Figure 2 présente les résultats obtenus en utilisant l'algorithme AKSC et plusieurs autres algorithmes de clustering (K-means, clustering spectral, clustering basé sur les modèles de mélange gaussiens, clustering hiérachique ascendant) pour lesquels le nombre exact de groupes recherchés a été donné à l'algorithme.

Les résultats obtenus illustrent la capacité de l'algorithme AKSC à répertorier automatiquement des objets même ceux avec des formes variées et ajourées. Comparé aux résultats obtenus à l'aide d'autres algorithmes de clustering, même quand le nombre d'objects contenus dans la planche est indiqué, l'algorithme AKSC apporte la solution attendue sans aucune hypothèse sur le nombre de classes.

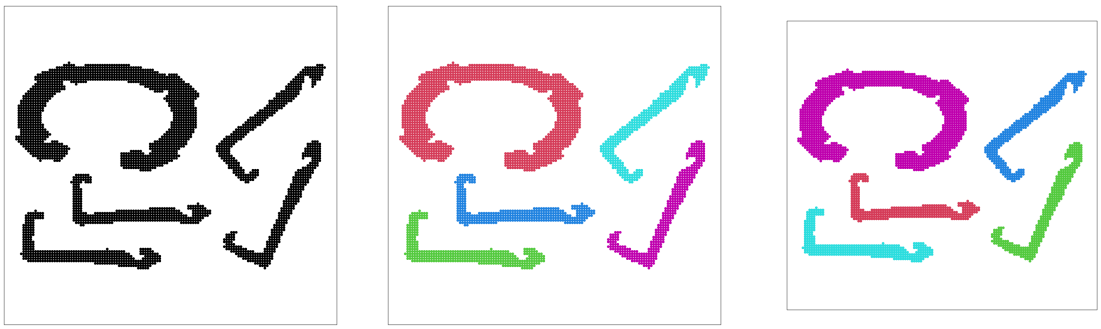
Bibliographie

Giulini, I. (2016). Kernel Spectral Clustering. arxiv1606.06519.

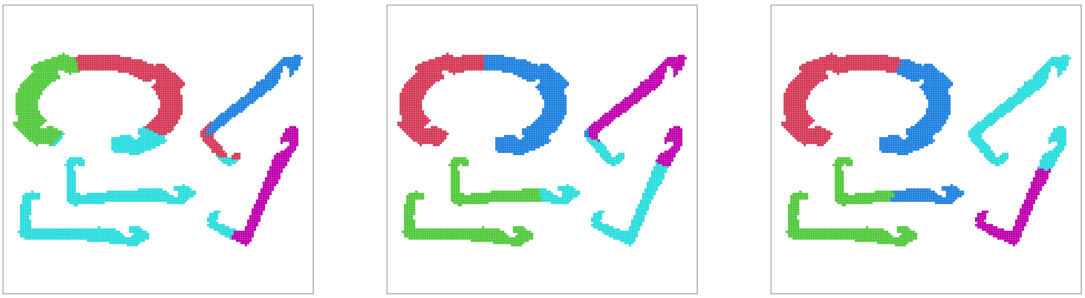
von Luxburg, U. (2007), A tutorial on spectral clustering. *Statistics and Computing*, 17, 395-–416.

Ng, A.Y. and Jordan, M.I. and Weiss, Y. (2001). On spectral clustering: Analysis and an algorithm, In *Advances in Neural Information Processing Systems*, 849—856.

Stork, D. G and Duda, R. O and Hart, P. E. (2001). Pattern classification, *A Wiley-Interscience Publication.*



(a) (b) (c)



(d) (d) (f)

Figure 2: Clustering de pièces manufacturées. Données initiales (a). Résultats obtenus par l'algorithme AKFC (b). En spécifiant le nombre de groupes attendus (K=5) en paramètre des algorithmes, Clustering spectral (c), Mélanges gaussiens (d), Kmeans (e), Clustering hiérarchique ascendant (f).

Sparse and group-sparse clustering for mixed data An illustration of the `vimpclust` package

Marie Chavent [1] & Alex Mourer [2] & Madalina Olteanu [3]

[1] *Univ. Bordeaux, CNRS, INRIA, Bordeaux INP, IMB, UMR 5251, F-33400 Talence, marie.chavent@u-bordeaux.fr*

[2] *SAMM EA 4543, Université Panthéon Sorbonne, moureralex@gmail.com*

[3] *CEREMADE, UMR 7534, Université Dauphine PSL, olteanu@ceremade.dauphine.fr*

Résumé. Les données en grande dimension contiennent souvent un mélange de variables numériques et catégorielles. De plus, dans de nombreux cas, les variables sont disponibles sous la forme de groupes définis a priori (mesures répétées, catégories d'attributs, ...). En pratique, classifier ces données pose plusieurs problèmes : comment utiliser les données mixtes pour construire des clusters pertinents? comment sélectionner les variables ou les groupes de variables les plus pertinents pour le *clustering*? Avec une approche *k-means*, il est possible d'utiliser une version pénalisée de la variance inter-classes, et de trouver à la fois la meilleure partition des données, et les variables ou groupes de variables les plus informatifs. Le présent manuscrit décrit les méthodes des *sparse k-means* et des *group-sparse k-means* pour des données mixtes, en utilisant le package `R` `vimpclust`. L'exemple d'un petit jeu de données réelles illustre comment la sélection de variables peut être directement combinée avec le *clustering*, en fournissant à la fois une classification pertinente et en préservant la qualité du *clustering*.

Mots-clés. clustering, k-means parcimonieux, pénalités L_1 et L_1-groupe, données mixtes, packages `R`.

Abstract. High-dimensional data may often contain both numerical and categorical features, and in some cases features may be available as natural groups (repeated measurements, categories of features, ...). Clustering this kind of data raises several issues: how to simultaneously deal with numerical and categorical features? how to build meaningful clusters of the input entities? how to select the most informative features or groups of features for the clustering? In the k-means framework, one may rely on a penalised version of the between-cluster variance, and find both the best partitioning of the data, and the most informative features or groups of features. The present manuscript illustrates sparse k-means and group-sparse k-means for mixed data, using the `vimpclust` package. The example provided on a small real-life dataset shows how feature selection may be directly combined with clustering, and provide a meaningful selection while preserving the quality of the clustering.

Keywords. clustering, sparse k-means, L_1 and group-L_1 penalties, mixed data, `R` packages.

1 Introduction

While feature selection received a great deal of attention in the supervised learning framework during the past twenty years, it is only much later and relatively recently that it effectively emerged in the unsupervised context. At the same time, with the dimensionality of the collected data constantly increasing, feature selection for clustering becomes a crucial issue, since the presence of many uninformative features could potentially damage the clustering, introduce noise and favour unstable results. Furthermore, features may often be available as priorly known groups (repeated measurements, categories of attributes, ...), and imply that group-wise selection should be considered instead of feature-wise one.

Since model-based approaches are naturally suited for including L_1 and related penalties, several contributions were proposed in the literature, and are summarised, for instance, in [1], which contains a detailed review on model-based clustering for high-dimensional data. In the k-means context, one may cite the sparse k-means procedure, introduced in [9] and based on a L_1-penalised version of the between-class variance, or more recent developments in [8], [2] or [6].

That being said, all methods cited above are essentially designed for numerical data, while real data is often made of numerical and categorical features. Some of the authors above touch upon the question of categorical features, by mentioning the possibility of making them numerical using a transformation through dummy variables. However, this processing step is not that immediate, since the Euclidean distance on zero-one vectors is not particularly suited for being mixed with Euclidean distances on numerical variables. Other authors implicitly suggest that the proposed algorithms may be written in terms of distances or dissimilarities between input data only, and hence it suffices to use an appropriate distance for categorical features. Nevertheless, the distance-based approaches may rapidly translate into an increased complexity if the size of the data becomes large.

In a recent contribution [5], we introduced an explicit method for feature selection in a mixed-data framework, by building on a penalized group-L_1 criterion. An R package, called `vimpclust`[1] and available on CRAN, has been implemented and allows one to train sparse k-means procedures on numerical or mixed features, or on groups of numerical features. In the present manuscript, we extend the above to groups of mixed features and add a new functionality to the package. In the rest of the paper, we briefly recall how to train sparse k-means for groups of numerical features, for mixed features, and for groups of mixed features, and illustrate these methods using the `vimpclust` package and a small real-life dataset.

[1] https://cran.r-project.org/web/packages/vimpclust/index.html

2 Sparse clustering for groups of numerical features

Group-sparse clustering may be of interest in many real-life situations, where a natural structure of groups is available for the features. In other practical applications, groups of features may be the output of a clustering applied to the features.

In the following, suppose that the p numerical features are divided into L priorly known groups, such that $\mathbf{X} = \left[\mathbf{X}^1|\ldots|\mathbf{X}^L\right] \in \mathbb{R}^{n\times p}$ is the matrix of data, where $\ell = 1,\ldots,L$, and $\mathbf{X}^\ell \in \mathbb{R}^{n\times p_\ell}$ represents the features in group ℓ, and $p_1 + ... + p_\ell = p$.

Group sparse k-means searches both a partitioning $C_1, ..., C_K$, and a set of weights $\mathbf{w}^T = (\mathbf{w}^1, \ldots, \mathbf{w}^L) \in \mathbb{R}^p$, maximising the between-class variance penalised by the L_1-group penalty:

$$\max_{C_1,...,C_K,\mathbf{w}\in\mathbb{R}^p} \mathbf{w}^T\mathbf{b} - \lambda \sum_{\ell=1}^{L} \sqrt{p_\ell}\|\mathbf{w}^\ell\|_2, \tag{1}$$

where $\|\mathbf{w}\|_2 \leq 1$ and $\mathbf{w} \geq 0$, and where $\lambda > 0$ is a tuning hyperparameter.

In the notation above, $\mathbf{b}^T = (b_1, ..., b_p)$ is the vector containing the between-class variance computed per each feature,

$$b_j = b_j(\mathbf{X}, C_1, ..., C_K) = \sum_{k=1}^{K} \frac{n_k}{n}(\overline{X}_{j,k} - \overline{X}_j)^2, \tag{2}$$

with $\overline{X}_j$ being the average of feature j over the entire data set, $\overline{X}_{j,k}$ the average of feature j in cluster k, and n_k the size of cluster C_k. One should also remark here that each component $\mathbf{w}_\ell$ or $\mathbf{b}_\ell$ in $\mathbf{w}$ and, respectively, $\mathbf{b}$, are actually vectors of size p_ℓ.

For a fixed number of clusters K and a fixed tuning parameter λ, the optimisation problem above may be solved using an iterative algorithm, which alternates two steps:

1. For a fixed vector of weights $\mathbf{w}$, find the partitioning $C_1, ..., C_K$ maximising $\mathbf{w}^T\mathbf{b}$. This is actually equivalent to training an usual k-means algorithm on the features scaled by the weights $\tilde{X}_j = \sqrt{w_j}X_j$, $j = 1, ..., p$.

2. For a fixed partitioning $C_1, ..., C_K$, find the vector of weights $\mathbf{w}$ maximising the penalised between-class variance, and such that $\|\mathbf{w}\|_2 \leq 1$. The solution may be found analytically as a function of the group soft-thresholding operator, and is described in detail in [5].

From a practical point of view, the group soft-thresholding operator will drop all groups of features which between-class variance norm $\|\mathbf{b}_\ell\|_2$ is smaller than $\sqrt{p_\ell}\lambda$, and shrink by $\sqrt{p_\ell}\lambda$ the between-class variance norm of the remaining groups of variables.

3 Sparse clustering for mixed data

In the context of mixed data, the group-sparse principle above may be used for building a procedure for variable selection. Indeed, each categorical feature may be transformed into a group of dummy ones, thus producing a natural structure of groups on the data. If the input data $\mathbf{X}$ is described by d_1 categorical features and d_2 numerical ones, such that each of the categorical features has p_j possible values, $j = 1, ..., d_1$, she may transform each categorical feature X_j into p_j dummy variables $\tilde{\mathbf{X}}_j = (\tilde{X}_j^1, ..., \tilde{X}_j^{p_j}) \in \{0,1\}^{n \times p_j}$ and thus define a natural group structure on the transformed data $\mathbf{Y} = [\mathbf{Y}_1|...|\mathbf{Y}_{d_1+d_2}]$, where $\mathbf{Y}_\ell = \tilde{\mathbf{X}}_\ell$ for $\ell = 1, ..., d_1$, $\mathbf{Y}_\ell = X_\ell$ for $\ell = d_1 + 1, ..., d_1 + d_2$, with respective group sizes $\mathbf{p}^T = (p_1, ..., p_{d_1}, 1, ..., 1) \in \mathbb{R}^{d_1+d_2}$.

When training group-sparse k-means on the group structure above, this amounts to performing variable selection in a mixed-data context. Before applying the algorithm described in the previous section, $\mathbf{Y}$ must be properly preprocessed. This prior step is described for example in [4]: numerical variables are scaled to zero mean and unit variance, while the dummy variables are centered and normalized by $1/\sqrt{\frac{n}{n_{j,s}}}$, where $n_{j,s}$ is the number of input data taking the s-th value of the jth categorical feature, or equivalently the sum over $\tilde{X}_j^s$. The scaling applied to the dummy variables actually leads to using a χ^2 distance on the categorical features, while the numerical features, after scaling, are compared through the usual Euclidean distance.

4 Sparse clustering for groups of mixed features

Suppose now that the $p = d_1 + d_2$ categorical and numerical features are divided into L priorly known groups. The sparse clustering procedures described above (for groups of numerical features and for non grouped mixed data) may be combined to select groups of mixed features. Indeed, the input data $\mathbf{X}$ may be transformed into a numerical matrix $\mathbf{Y}$ with $p_1 + \ldots + p_{d_1}$ dummy variables replacing the d_1 categorical features, this matrix $\mathbf{Y}$ being pre-processed as described section (3). The p_j dummy variables of each categorical features X_j may then be assigned to the group ℓ of X_j, thus defining L groups of the $p_1 + \ldots + p_{d_1} + d_2$ transformed features. The sparse clustering method for groups of numerical variables described in Section 2 may then be applied to the pre-processed numerical matrix $\mathbf{Y}$ and these new groups. Because the levels of a categorical feature belong by construction to the same group, if this group is selected, the feature is selected.

5 An illustration of `vimpclust`

The `vimpclust` package has been implemented as a general tool for sparse k-means clustering. It handles sparse k-means for numerical features as introduced in [9], but also

group-sparse k-means for numerical features, sparse k-means for mixed features, and group sparse k-means for mixed features. The package is available on CRAN, and contains several detailed vignettes with use cases.

As an example, sparse k-means and group sparse k-means for mixed data are illustrated on the `wine` dataset [7]. The `wine` data contains a description of 21 wine varieties, using 31 features, including soil type and wine label (categorical with respectively five and three categories) and different sensory descriptors (numerical). The dimensionality of the data is larger than the number of inputs, and feature selection through sparse clustering is crucial for meaningful interpretation.

First, the data is clustered using the sparse k-means procedure for mixed data, using the `sparsewkm` function in `vimpclust`. The number of clusters is fixed equal to four.

```
res <- sparsewkm(X = wine, centers = 4)
plot(res, what="weights.features")
plot(res, what="expl.var")
```

The regularisation paths for the weights of each feature, but also the regularisation path of the ratio of explained variance are illustrated in Figure 1. The hyper-parameter λ varies between 0 and 1, as the initial features were all scaled to unit variance. If one decides to keep the last configuration before the ratio of explained variance drops significantly, ten features out of thirty are eventually selected: *Aroma.quality.before.shaking, Quality.of.odour, Balance, Intensity, Overall.quality, Surface.feeling, Aroma.quality, Smooth, Harmony, Typical*.

What about if one is interested in selecting groups of significant features for the clustering? In order to do so, and since one has no prior knowledge on the groups of features for the `wine` dataset, we decide to build groups of features using the `ClustOfVar` procedure as described in [3]. The resulting 7 groups of features are illustrated in Figure 2.

Clustering is achieved using the `groupsparsewkm` function in `vimpclust`, and by providing a supplementary argument `index`, which contains groups labels for the features. Here again, four clusters for the input data were searched for.

```
res <- groupsparsewkm(X, centers = 4, index = groupes)
plot(res, what = "weights.features")
plot(res, what="expl.var")
```

The regularisation paths both in terms of features, and in terms of groups of features, are illustrated in Figure 3.

After selecting λ such that the best trade-off between the ratio of explained variance and the number of selected groups is achieved, only three groups of features are kept. Their composition in detailed in Figure 4. As one may easily see, all features selected by the sparse k-means procedure for mixed data also appear in the selected groups of

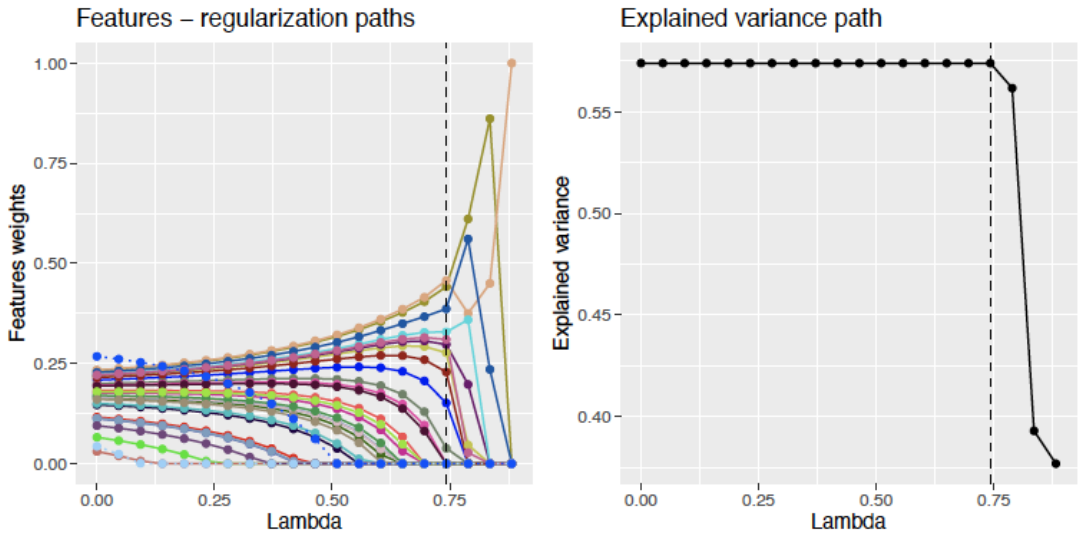


Figure 1: Regularisation paths for different values of λ. Left: features paths. Right: ratio of explained variance path.

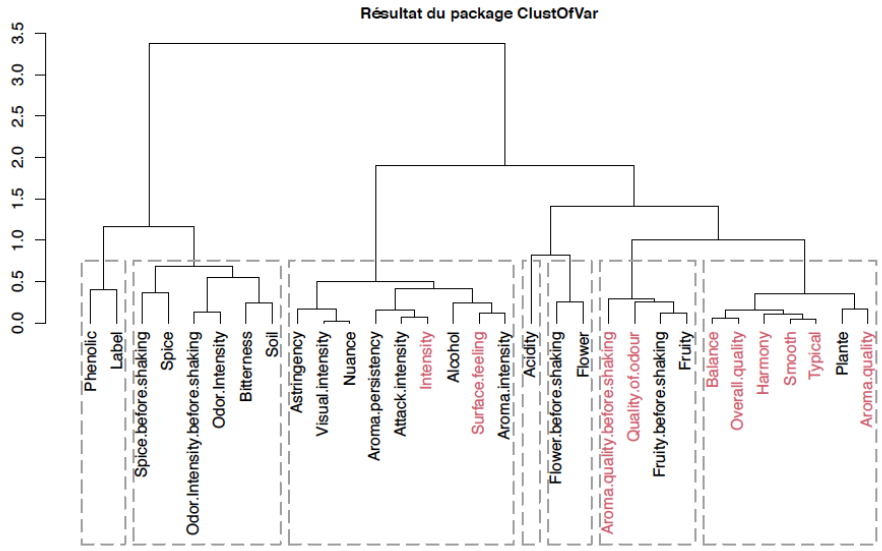


Figure 2: Resulting groups of features after having trained `vimpclust`

features. The results of the two procedures are thus coherent. Furthermore, groups of features are interesting to consider since they provide better insights on the available information in the data.

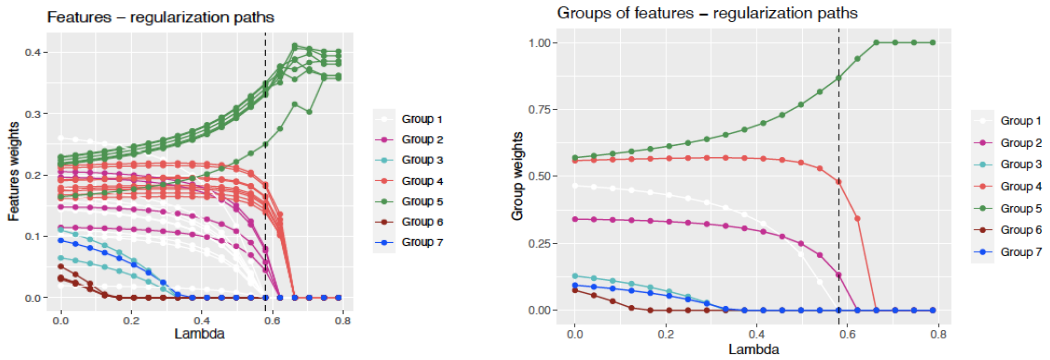


Figure 3: Regularisation paths for different values of λ. Left: features paths. Right: groups of features paths.

Group 2	Group 4	Group 5
Aroma.quality.before.shaking	Visual.intensity	Plante
Fruity.before.shaking	Nuance	Aroma.quality
Quality.of.odour	Surface.feeling	Balance
Fruity	Aroma.intensity	Smooth
	Aroma.persistency	Harmony
	Attack.intensity	Overall.quality
	Astringency	Typical
	Alcohol	
	Intensity	

Figure 4: Selected groups of features by the group-sparse k-means. Red: features selected individually by the sparse k-means procedure.

6 Conclusion

Several k-means generalisations, allowing to train sparse clusterings on mixed data, and to select important features or groups of important features were described in the above sections. These variants of k-means are implemented in a recent package available on CRAN, `vimpclust`. A detailed example on a small real-life data set allowed to illustrate how the proposed methods produce both meaningful clusterings and feature selection, the latter being a crucial issue for high-dimensional data.

References

[1] Ch. Bouveyron and C. Brunet-Saumard. Model-based clustering of high-dimensional data: A review. *Computational Statistics & Data Analysis*, 71:52–78, 2014.

[2] S. Chakraborty and S. Das. A strongly consistent sparse k-means clustering with direct l_1 penalization on variable weights. *arXiv preprint arXiv:1903.10039*, 2019.

[3] M. Chavent, B. Kuentz-Simonet, V.and Liquet, and J. Saracco. Clustofvar: An r package for the clustering of variables. *Journal of Statistical Software*, 50:116, 2012.

[4] M. Chavent, V. Kuentz-Simonet, A. Labenne, and J. Saracco. Multivariate analysis of mixed data: The pcamixdata r package. *arXiv preprint arXiv:1411.4911*, 2014.

[5] M. Chavent, J. Lacaille, A. Mourer, and M. Olteanu. Sparse k-means for mixed data via group-sparse clustering. In *ESANN 2020 - 28th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, volume 978-2-87587-074-2, Bruges / Virtual, Belgium, October 2020.

[6] Z. Huo and G. Tseng. Integrative sparse k-means with overlapping group lasso in genomic applications for disease subtype discovery. *Ann. Appl. Stat.*, 11(2):1011–1039, 06 2017.

[7] S. Le, J. Josse, and F. Husson. Factominer: An r package for multivariate analysis. *Journal of Statistical Software*, 25(1):1–18, 2008.

[8] W. Sun, J. Wang, Y. Fang, et al. Regularized k-means clustering of high-dimensional data and its asymptotic consistency. *Electronic Journal of Statistics*, 6:148–167, 2012.

[9] D.M. Witten and R. Tibshirani. A framework for feature selection in clustering. *Journal of the American Statistical Association*, 105(490):713–726, 2010. PMID: 20811510.

Cartes auto-organisatrices pour l'exploration des données partiellement observées et l'imputation des données manquantes

Sara Rejeb [1,2] & Catherine Duveau [2] & Tabea Rebafka [1]

[1] *LPSM, Sorbonne Université, Université de Paris & CNRS, 4, Place Jussieu, 75005 Paris, France*
sara.rejeb@sorbonne-universite.fr
tabea.rebafka@sorbonne-universite.fr
[2] *Safran Aircraft Engines, Réau, 77550 Moissy-Cramayel, France*
catherine.duveau@safrangroup.com

Résumé. Les cartes auto-organisatrices de Kohonen sont des réseaux de neurones qui accomplissent deux tâches simultanément : visualiser des données multidimensionnelles par projection sur une carte bidimensionnelle et discrétiser l'espace de façon ordonnée. L'algorithme classique de Kohonen développé pour optimiser ces cartes est conçu uniquement pour des données complètes. Cependant, dans de nombreuses applications, les données contiennent naturellement de nombreuses données manquantes. Dans ce travail, nous proposons alors une extension des cartes auto-organisatrices pour des données partiellement observées. Notre approche consiste à traiter simultanément les deux problèmes de l'apprentissage d'une carte auto-organisatrice et de l'imputation des données manquantes. Nous définissons une nouvelle fonction de perte, qui vise à optimiser simultanément la carte et également l'imputation des valeurs manquantes. Les simulations numériques illustrent de bonne performance en termes de qualité de la carte et d'imputation des données manquantes comparé à l'état de l'art. Une implémentation de la méthode en R est disponible sur le CRAN dans le package `missSOM`.

Mots-clés. Cartes auto-organisatrices, données incomplètes, imputation des données manquantes.

Abstract. Kohonen self-organizing maps are neural networks that perform two tasks simultaneously: visualizing multidimensional data by projection onto a two-dimensional map and discretizing the space in an ordered way. The classical Kohonen algorithm developed to optimise maps is conceived only for complete data. However, in many applications, the data naturally contains many missing data. In this work, we propose an extension of self-organising maps to partially observed data. Our approach is to deal simultaneously with the two problems of learning a self-organizing map and imputing missing data. We define a new loss function, which aims at simultaneously optimizing the map and also the imputation of missing values. Numerical simulations illustrate good performance in terms of map quality and missing data imputation compared to the state of the art. An implementation of the method in R is available on CRAN in the package `missSOM`

Keywords. Self-organizing maps, incomplete data, missing data imputation.

1 Introduction

Les tâches courantes de l'exploration des données sont la visualisation et le clustering de données. Bien qu'il existe de nombreuses méthodes permettant d'accomplir l'une ou l'autre de ces tâches, les cartes auto-organisatrices fournissent simultanément une représentation visuelle des données en faible dimension sous forme de carte ordonnée qui discrétise les données. Introduite par Kohonen [1995], cette technique d'apprentissage non supervisé est devenue très populaire dans de nombreuses applications réelles pour la visualisation, l'analyse et la modélisation de données, car elle fournit des résultats facilement interprétables avec une vision globale des données. Par ailleurs, de nombreuses variantes ont été développées comme des extensions à des types de données plus complexes (mixtes, textuelles, etc.) [Kohonen and Somervuo, 2002, Lebbah et al., 2005]. Ceci montre tout l'intérêt et l'importance des cartes de Kohonen jusqu'à aujourd'hui.

Les valeurs manquantes constituent un problème commun aux ensembles de données dans la plupart des domaines d'application et sont à l'origine de diverses raisons. Dans l'industrie, les instruments de mesure peuvent présenter des dysfonctionnements donnant lieu à des entrées erronées et manquantes [Lakshminarayan et al., 2004]. Dans le domaine de la recherche médicale, des données manquantes peuvent apparaître dans les essais cliniques lorsque des patients abandonnent ou cessent de prendre un traitement pendant une certaine période [Shih, 2002].

Les données manquantes représentent une perte d'informations et leur impact sur les résultats statistiques peut être conséquent, entraînant notamment des estimations biaisées.

Dans la suite, nous nous intéressons aux cartes auto-organisatrices en présence de données manquantes.

Dans ce travail, nous avons l'ambition de combiner les tâches d'imputation de données manquantes et d'apprentissage d'une carte auto-organisatrice par une approche fondée sur des principes. Notre motivation est le fait que toute méthode d'imputation non triviale est basée sur un certain modèle de données, il est donc naturel d'utiliser la carte auto-organisatrice pour l'imputation. Inversement, une meilleure carte peut être apprise lorsque les données sont complètes plutôt que incomplètes. Ainsi, traiter les deux tâches simultanément peut être bénéfique pour les deux.

2 Adaptation des cartes auto-organisatrices aux données incomplètes

Dans cette section, nous introduisons quelques notations et rappelons le principe des cartes auto-organisatrices de Kohonen.

2.1 Algorithme classique de Kohonen

Soit $X = [x_1, \ldots, x_n] \in \mathbb{R}^{n \times p}$ une matrice de données contenant n observations $x_i \in \mathbb{R}^p$. Une carte auto-organisatrice est généralement une carte bidimensionnelle représentée comme une grille régulière composée de K neurones fixes et disposés de manière ordonnée. Une fonction de voisinage $V_\lambda : \{1, \ldots, K\}^2 \mapsto \mathbb{R}_+$ définie la relation d'ordre entre les neurones. Le rayon de voisinage $\lambda > 0$ décrit la zone d'influence autour d'un neurone.

Chaque neurone k est associé à un vecteur prototype $w_k \in \mathbb{R}^p$, également appelé vecteur code. Idéalement, les vecteurs prototypes des neurones voisins sur la grille sont proches les uns des autres, de sorte que les points de données qui sont proches dans l'espace d'entrée le sont également sur la carte. Pour ce faire, on considère la fonction de perte F définie par

$$F(W) = \frac{1}{2n} \sum_{i=1}^{n} \sum_{k=1}^{K} V_\lambda(k, h(x_i, W)) \|x_i - w_k\|^2, \tag{1}$$

où $W = [w_1, \ldots, w_K] \in \mathbb{R}^{p \times K}$ est la matrice des K vecteurs prototypes et $h : \mathbb{R}^p \times \mathbb{R}^{p \times K} \mapsto \{1, \ldots, K\}$ représente la fonction d'affectation, attribuant l'observation x à un des neurones de la carte, celui associé au vecteur prototype le plus proche de l'observation dans l'espace d'entrées, définie par

$$h(x, W) = \arg \min_{1 \leq k \leq K} \|x - w_k\|. \tag{2}$$

La fonction de perte F prend en compte toutes les distances entre chaque observation et tous les vecteurs codes, pondérées par la fonction de voisinage évaluée sur les neurones correspondants. Notons que dans le cas particulier où la fonction de voisinage satisfait $V_\lambda(k, \ell) = 0$ pour tout $k \neq \ell$, la fonction de perte F est le critère minimisé par l'algorithme k-means.

Pour minimiser la fonction perte F, Ritter et al. [1992] a montré que, dans un même cadre donné, un algorithme de descente de gradient peut être utilisé. Cet algorithme est appelé l'algorithme stochastique de Kohonen. Dans cette procédure itérative, une observation x_i prise au hasard est traitée à chaque itération. D'abord, le neurone gagnant $\ell = h(x_i, W^{(t)})$ est déterminé, puis tous les vecteurs prototypes sont mis à jour. A l'itération $t + 1$, les nouveaux vecteurs prototypes $w_k^{(t+1)}$ pour $k = 1, \ldots, K$ sont obtenus par

$$w_k^{(t+1)} = w_k^{(t)} + \varepsilon_t V_\lambda(k, \ell)(x_i - w_k^{(t)}), \tag{3}$$

où $(\varepsilon_t)_{t\geq 0}$ est une suite décroissante de pas d'apprentissage. Cette mise à jour attire tous les prototypes vers l'observation x_i.
La procédure est répétée jusqu'à convergence vers un minimum local.

2.2 Adaptation aux données incomplètes

Considérons maintenant une matrice de données incomplète $n \times p$ contenant des valeurs manquantes. Soit la matrice binaire d'indicateurs de réponses $M = (m_{i,j})_{i,j}$ dans $\{0,1\}^{n\times p}$ qui indique où sont situées les entrées manquantes dans la matrice de données et est définie par

$$m_{i,j} = \begin{cases} 1 & \text{si } x_{i,j} \text{ est observé} \\ 0 & \text{si } x_{i,j} \text{ est manquant} \end{cases}$$

Nous désignons par X^{obs} l'ensemble des valeurs de données observées et par X^{miss} l'ensemble des entrées de données non observées cachées par la matrice M. Les données complètes sont désignées par $X^{\text{compl}} = (X^{\text{obs}}, X^{\text{miss}})$. De même, pour le vecteur d'observation x_i, nous désignons par x_i^{obs} et x_i^{miss} les entrées observées et manquantes respectivement. Avec un certain abus de notation, $x_i^{\text{compl}} = (x_i^{\text{obs}}, x_i^{\text{miss}})$ est le vecteur d'observation complet correspondant aussi à la i-ième ligne de X^{compl}.

2.2.1 Nouvelle fonction de perte

Nous introduisons une nouvelle fonction de perte qui recherche à optimiser simultanément la carte auto-organisatrice et l'imputation des données manquantes. Autrement dit, nous recherchons à la fois les meilleurs vecteurs prototypes $W \in \mathbb{R}^{p\times K}$ de la carte ainsi que les meilleures valeurs pour les données manquantes désignées par X^* choisies dans l'ensemble des valeurs possibles pour les entrées manquantes $\mathcal{X}^{\text{miss}}$.

Pour définir ce nouveau critère, une adaptation de la définition du neurone gagnant est donnée. En présence de valeurs manquantes, il est naturel de restreindre la distance euclidienne dans (2) aux seules entrées observées. Plus précisément, pour tout vecteur $x^{\text{obs}} \in \mathbb{R}^{p'}$, $m \in \{0,1\}^p$ avec

$$h^{\text{miss}}(x^{\text{obs}}, m, W) = \arg\min_{1\leq k\leq K} \|x^{\text{obs}} - w_k \odot m\|^2,$$

où $w \odot m$ représente un vecteur de dimension p' composé des éléments w_i de w tels que $m_i = 1$. Nous définissons maintenant la nouvelle fonction de perte F_{missom} associée par

$$F_{\text{missom}}(W, X^*) = \frac{1}{2n}\sum_{i=1}^{n}\sum_{k=1}^{K} V_\lambda\left(k, h^{\text{miss}}(x_i^{\text{obs}}, m_i, W)\right)\left\|(x_i^{\text{obs}}, x_i^*) - w_k\right\|^2,$$

où $m_i \in \mathbb{R}^p$ est la i-ème ligne de la matrice M et $(x_i^{\text{obs}}, x_i^*)$ désigne le i-ième vecteur d'observation complété par x_i^*. Alors que

$$\left\|(x_i^{\text{obs}}, x_i^*) - w_k\right\|^2 = \left\|x_i^{\text{obs}} - w_k \odot m_i\right\|^2 + \left\|x_i^* - w_k \odot (\mathbf{1}_p - m_i)\right\|^2,$$

où $\mathbf{1}_p = (1, \ldots, 1)^T \in \mathbb{R}^p$, le critère F_{missom} peut être décomposé en deux termes, en fonction des entrées observées et manquantes, comme suit

$$F_{\text{missom}}(W, X^*) = F_{\text{obs}}(W) + F_{\text{miss}}(W, X^*),$$

où $F_{\text{obs}}(W)$ est la partie de la fonction perte sur les entrées observées, donnée par

$$F_{\text{obs}}(W) = \frac{1}{2n} \sum_{i=1}^{n} \sum_{k=1}^{K} V_\lambda \left(k, h^{\text{miss}}(x_i^{\text{obs}}, m_i, W)\right) \left\| x_i^{\text{obs}} - w_k \odot m_i \right\|^2,$$

et $F_{\text{miss}}(W, X^*)$ est la contribution des valeurs imputées sur la fonction perte, définie par

$$F_{\text{miss}}(W, X^*) = \frac{1}{2n} \sum_{i=1}^{n} \sum_{k=1}^{K} V_\lambda \left(k, h^{\text{miss}}(x_i^{\text{obs}}, m_i, W)\right) \left\| x_i^* - w_k \odot (\mathbf{1}_p - m_i) \right\|^2.$$

Remarquons que dans le cas des données complètes ($M = \mathbf{1}_{n \times p}$), $F_{\text{miss}}(W, X^*) = 0$ pour tout W et X^*, auquel cas le critère F_{missom} correspond à celui des cartes de Kohonen classique, soit $F_{\text{missom}}(W, X^*) = F(W)$.

2.2.2 Algorithme de minimisation

Pour la minimisation de $(W, X^*) \mapsto F_{\text{miss}}(W, X^*)$ sur $\mathbb{R}^{K \times p} \times \mathcal{X}^{\text{miss}}$ nous proposons d'alterner la minimisation de W et de X^* en gardant l'un des deux arguments fixe.

Pour X^* fixe, la fonction $W \mapsto F_{\text{missom}}(W, X^*)$ est similaire à la fonction objective F dans (1) dans le cas de données complètes appliquées aux données augmentées $X^{\text{aug}} = (X^{\text{obs}}, X^*)$. La seule différence réside dans la définition des neurones gagnants par h^{miss} qui apparaissent dans la fonction de voisinage V_λ. Ainsi, un algorithme de Kohonen appliqué à X^{aug} peut être utilisé pour trouver les meilleurs vecteurs W.

À son tour, lorsque W est fixé, la minimisation de $X^* \mapsto F_{\text{missom}}(W, X^*)$ se résume à minimiser $X^* \mapsto F_{\text{missom}}(W, X^*)$. Ce problème a une solution unique et explicite, donnée pour tous les $1 \leq i \leq n$ et j tels que $m_{i,j} = 0$ par

$$x_{i,j}^* = \frac{\sum_{k=1}^{K} V_\lambda(k, h^{\text{miss}}(x_i^{\text{obs}}, m_i, W)) w_{k,j}}{\sum_{k=1}^{K} V_\lambda(k, h^{\text{miss}}(x_i^{\text{obs}}, m_i, W))}. \tag{4}$$

Autrement dit, les valeurs imputées sont une moyenne pondérée des vecteurs prototypes pondérés selon la fonction de voisinage.

En résumé, l'algorithme met à jour les valeurs imputées pour les données manquantes et applique l'algorithme classique de Kohonen avec une fonction d'affectation au neurone gagnant ajustée h^{miss} pour apprendre la carte. Cette opération est répétée jusqu'à convergence ou jusqu'à atteindre un nombre maximal d'itérations choisi par l'utilisateur. Comme valeurs initiales pour les valeurs imputées X^*, on peut simplement imputer par la moyenne ou la médiane des variables obtenues sur les entrées observées.

2.2.3 Version accélérée

L'algorithme peut s'avérer en revanche coûteux en termes de temps de calcul, en particulier lorsque le nombre d'itérations est important, puisque l'intégralité de l'algorithme standard de Kohonen est exécutée à chaque itération. Nous avons alors proposé une version accélérée en entrelaçant les mises à jour des données manquantes et les itérations de l'algorithme classique de Kohonen qui est donc exécuté qu'une seule fois. Ainsi, le temps de calcul de la version accélérée de missSOM est comparable à celui de l'algorithme de Kohonen standard, puisque la mise à jour des valeurs imputées est rapide. Une étude numérique a été réalisée pour comparer les performances des deux versions et ont montré des cartes similaires avec un gain significatif en temps de calculs pour la version accélérée.

3 Simulations numériques

Nous avons évalué et comparé les performances de notre méthode à l'état de l'art sur données simulées et sur un cas de données réel. Les performances des méthodes ont été évaluées en termes de qualité de la carte auto-organisatrice, de qualité de l'imputation des données manquantes et en temps de calculs.

Cette étude numérique montre de bonnes performances de notre méthode. De plus, l'imputation des données manquantes optimisé simultanément avec la carte permet notamment de ne pas dégrader la qualité de la carte.

References

Teuvo Kohonen. *Self-organizing maps*. Springer, Berlin, 1995.

Teuvo Kohonen and Panu Somervuo. How to make large self-organizing maps for nonvectorial data. *Neural Networks*, 15(8):945–952, 2002. ISSN 0893-6080.

K. Lakshminarayan, S. Harp, and T. Samad. Imputation of missing data in industrial databases. *Applied Intelligence*, 11:259–275, 2004.

Mustapha Lebbah, Aymeric Chazottes, Fouad Badran, and Sylvie Thiria. Mixed topological map. In *Proc. of ESANN*. ACM, 2005.

Helge Ritter, Thomas Martinetz, and Klaus Schulten. *Neural Computation and Self-Organizing Maps: An Introduction*. Addison-Wesley Longman Publishing Co., Inc., USA, 1992.

W. Shih. Problems in dealing with missing data and informative censoring in clinical trials. *Current Controlled Trials in Cardiovascular Medicine*, 3:4 – 4, 2002.

Détection d'atypiques pour données de composition avec la méthode ICS

Christine Thomas-Agnan [1] & Anne Ruiz-Gazen [2] & Thibault Laurent [3] & Camille Mondon [4]

[1] *Toulouse School of Economics, France; christine.thomas@tse-fr.eu*
[2] *Toulouse School of Economics, France; anne.ruiz-gazen@tse-fr.eu*
[3] *Toulouse School of Economics, CNRS, France; thibault.laurent@tse-fr.eu*
[4] *DMA, Ecole Normale Supérieure, PSL, France; camille.mondon@ens.psl.eu*

Résumé. La méthode ICS (Invariant Coordinate Selection) est une méthode statistique introduite par Tyler et al. (2009) basée sur la diagonalisation jointe de deux matrices de dispersion. Une approche de cette méthode basée sur un modèle et appelée ICA (Invariant Coordinate Analysis) a déja été adaptée aux données de composition par Muehlmann et al. (2021). Dans une approche sans modèle, ICS est aussi efficace pour détecter des valeurs atypiques, voir Nordhausen et Ruiz-Gazen (2022). Nous proposons de développer une version d'ICS pour détecter des atypiques dans les données de composition. Cette méthode est d'abord introduite dans un espace de coordonnées ilr associé à une matrice de contraste et suit la procédure de Archimbaud et al. (2018b). Nous montrons ensuite que la technique est indépendante du choix de matrice de contraste et peut être entièrement définie dans le simplexe. Pour cela nous établissons d'abord un ensemble de propriétés des matrices satisfaisant la propriété "somme-zero" et introduisons une définition de la distance de Mahalanobis adaptée au simplexe ainsi que la classe des M-estimateurs de dispersion en une étape. Nous définissons également la famille de distributions elliptiques dans le simplexe. Nous montrons ensuite comment interpréter les résultats directement dans le simplexe pour deux jeux de données simulées ainsi qu'un jeu de données de parts de marché pour le marché automobile.

Mots-clés. données de composition, ICS, détection d'atypiques, matrices de dispersion.

Abstract. Invariant Coordinate Selection (ICS) is a multivariate statistical method introduced by Tyler et al. (2009) and based on the simultaneous diagonalization of two scatter matrices. A model based approach of ICS, called Invariant Coordinate Analysis, has already been adapted for compositional data in Muehlmann et al. (2021). In a model free context, ICS is also helpful at identifying outliers Nordhausen and Ruiz-Gazen (2022). We propose to develop a version of ICS for outlier detection in compositional data. This version is first introduced in coordinate space for a specific choice of ilr coordinate system associated to a contrast matrix and follows the outlier detection procedure proposed by Archimbaud et al. (2018b). We then show that the procedure is independent of the choice of contrast matrix and can be defined directly in the simplex. To do so, we first establish

some properties of the set of matrices satisfying the zero-sum property and introduce a simplex definition of the Mahalanobis distance and the one-step M-estimators class of scatter matrices. We also need to define the family of elliptical distributions in the simplex. We then show how to interpret the results directly in the simplex using two artificial datasets and a real dataset of market shares in the automobile industry.

Keywords. compositional data, invariant coordinate selection, outlier detection, scatter matrices.

1 Objectif

Les données de composition sont par nature multivariées. Ce sont des vecteurs à composantes positives pour lesquels on s'intéresse aux parts relatives: ils peuvent être représentés par un vecteur du simplexe en divisant chaque composante par leur somme. Les méthodes classiques de statistique ne peuvent pas s'appliquer directement à ces données en raison de la présence des contraintes de positivité et de somme égale à un. En utilisant des transformations, on peut envoyer ces données dans un espace de dimension inférieure dans lequel les contraintes ont disparu et alors on peut procéder comme d'habitude.

Nous adaptons dans ce travail la méthode ICS pour la détection de valeurs atypiques dans des données de composition. La méthode ICS (Invariant Coordinate Selection) est une méthode statistique introduite par Tyler et al. (2009) basée sur la diagonalisation jointe de deux matrices de dispersion: par exemple la matrice de variance-covariance et la matrice basée sur les moments d'ordre 4. Pour une distribution elliptique, ces matrices sont toutes proportionnelles à la matrice de variance-covariance. Mais lorsque l'on s'éloigne du contexte elliptique, les deux matrices vont estimer des quantités différentes et leur comparaison va permettre de révèler de potentielles structures d'intérêt comme des observations atypiques ou des groupes. Dans cette présentation, nous nous intéressons à la détection d'anomalies tel que détaillé dans Archimbaud et al. (2018b).

2 Rappels sur la méthode ICS

Pour deux matrices de dispersion $\mathbf{S}_1(\mathbf{X})$ et $\mathbf{S}_2(\mathbf{X})$, ICS consiste à faire la diagonalisation jointe de $\mathbf{S}_1(\mathbf{X})$ et $\mathbf{S}_2(\mathbf{X})$. La méthode trouve une matrice $p \times p$ $\mathbf{H}(\mathbf{X})$ et une matrice diagonale $\mathbf{\Lambda}(\mathbf{X})$ telles que:

$$\mathbf{H}(\mathbf{X})^T \mathbf{S}_1(\mathbf{X}) \mathbf{H}(\mathbf{X}) = \mathbf{I}_p \quad \text{and} \quad \mathbf{H}(\mathbf{X})^T \mathbf{S}_2(\mathbf{X}) \mathbf{H}(\mathbf{X}) = \mathbf{\Lambda}(\mathbf{X}), \tag{1}$$

où $\mathbf{I}_p$ est matrice identité de dimension p. La matrice $\mathbf{\Lambda}(\mathbf{X})$ contient les valeurs propres de $\mathbf{S}_1(\mathbf{X})^{-1}\mathbf{S}_2(\mathbf{X})$ dans l'ordre décroissant, et les colonnes de la matrice $\mathbf{H}(\mathbf{X}) = (\mathbf{h}_1, \ldots, \mathbf{h}_p)$

contiennent les vecteurs propres correspondants:

$$\mathbf{S}_2(\mathbf{X})\mathbf{H}(\mathbf{X}) = \mathbf{S}_1(\mathbf{X})\mathbf{H}(\mathbf{X})\mathbf{\Lambda}(\mathbf{X})$$

$$\text{ou} \quad \mathbf{S}_1(\mathbf{X})^{-1}\mathbf{S}_2(\mathbf{X})\mathbf{H}(\mathbf{X}) = \mathbf{H}(\mathbf{X})\mathbf{\Lambda}(\mathbf{X}).$$

En utilisant un estimateur de localisation affine-équivariant $\mathbf{T}(\mathbf{X})$, les scores ICS sont donnés par:

$$\mathbf{Z} = \mathbf{H}(\mathbf{X})^T(\mathbf{X} - \mathbf{T}(\mathbf{X})), \tag{2}$$

et définissent les coordonnées invariantes.

Lorsqu'il y a une faible proportion d'atypiques, Archimbaud et al. (2018b) montrent qu'on peut se concentrer sur les composantes indépendantes correspondant aux plus fortes valeurs propres. Les atypiques sont alors les points pour lesquels la distance ICS calculée avec ces composantes est supérieure à un seuil choisi.

3 La méthode ICS pour données de composition

En passant par la transformation clr, on peut définir pour un vecteur de composition aléatoire, une espérance, une matrice de variance-covariance et une matrice basée sur les moments d'ordre 4. Nous introduisons aussi la distance de Mahalanobis pour un tel vecteur. Pour adapter une méthode aux données de composition, on peut, comme cela est classique dans ce domaine, commencer par la définir dans un espace de coordonnées. Plusieurs transformations sont possibles, dont la transformation ilr qui est paramétrée par une matrice dite de contraste $\mathbf{V}$ (Pawlowsky-Glahn et al., 2015). Il faut ensuite démontrer que la procédure ne dépends pas de ce choix arbitraire de $\mathbf{V}$. Pour cela il est d'abord nécéssaire de démontrer certains résultats d'algèbre. Auparavant, nous montrons que si une transformée ilr d'un tel vecteur suit une loi elliptique, alors c'est le cas aussi pour toute autre transformée ilr de ce vecteur et il est facile d'établir le lien entre les paramètres de ces distributions. Les endomorphismes du simplexe sont associés à une famille de matrices $\mathcal{A}$ telles que la somme par ligne et par colonne est nulle. Cet ensemble est en fait une algèbre pour les opérations de perturbation et d'exponentiation qui font du simplexe un espace vectoriel. Nous définissons la transformation clr des matrices et la transformation ilr des matrices de $\mathcal{A}$ et montrons leurs propriétés. Notamment, dans le cas d'une matrice symmétrique et donc diagonalisable, nous établissons le lien entre les éléments propres de cette matrice au sens usuel, et ceux de ses transformées ilr et clr.

Grace à ces résultats, nous pouvons montrer que la procédure proposée ne dépend pas du choix de la tranformation ilr et peut s'exprimer entièrement dans le simplexe. Nous montrons aussi comment reconstruire les données à partir des axes ICS et des scores ICS directement dans le simplexe. Enfin nous montrons comment interpréter les axes ICS dans le simplexe, et comment choisir des représentations appropriées pour visualiser les atypiques avec des triangles ternaires.

4 Exemple d'application

Voici un exemple d'application sur un jeu de données simulées de petite dimension. On simule $n = 100$ vecteurs de composition avec $D = 3$ composantes, selon un mélange de deux groupes: 90% des observations sont simulées dans un espace ilr selon une loi gaussienne multivariée de paramètres $\mu_1^* = (0, 0)$ and $\Sigma_1^* = \begin{pmatrix} 0.04 & 0.02 \\ 0.02 & 0.04 \end{pmatrix}$ et les 10% restants sont simulés selon une loi gaussienne multivariée de paramètres $\mu_2^* = (\frac{2}{\sqrt{2}} \log 2, \frac{-1}{\sqrt{6}} \log 2)$ and $\Sigma_2^* = \begin{pmatrix} 0.05 & 0 \\ 0 & 0.05 \end{pmatrix}$. La figure 1 montre sur la gauche le jeu de données dans le simplexe $\mathcal{S}^3$ et sur la droite les points transformés dans un espace de coordonnées ilr. Les points de couleur turquoise (resp. magenta) appartiennent au premier groupe (resp. le groupe 2) et on peut voir que la composante x_2 a des valeurs plus élevées dans le groupe 2 que dans le groupe 1. Nous appliquons la méthode ICS dans espace de coordonnées ilr avec le package **ICSOutlier** d' Archimbaud et al. (2018a). Les valeurs propres (Figure 2 sur la gauche) sont 1.57 et 0.81 ce qui suggère de ne conserver qu'un seul axe ICS. La figure 2 à droite montre les distances ICS calculées avec les coordonnées ilr. La droite horizontale représente le seuil basé sur des simulations de Monte Carlo et un quantile d'ordre 90%. La figure 1 sur la droite permet d'indentifier les atypiques représentés par un cercle plein. La méthode identifie correctement les dix observations du groupe 2 comme atypiques alors q'une seule observation de l'autre groupe est incorrectement identifiée. Les deux premiers axes ICS sont $a_1^* = (0.31, -0.1)$ et $a_2^* = (0.12, 0.22)$ dans l'espace ilr et $a_1 = (0.27, 0.43, 0.30)$ et $a_2 = (0.28, 0.33, 0.39)$ dans le simplexe. La figure 3 montre les projections des observations sur le premier axe ICS à gauche et sur le second à droite, dans l'espace ilr en haut et dans le simplexe en dessous. On voit que le premier axe discrimine bien les deux groupes. D'autres exemples seront également présentés dont un avec un plus grand nombre de composantes et un jeu de données réelles.

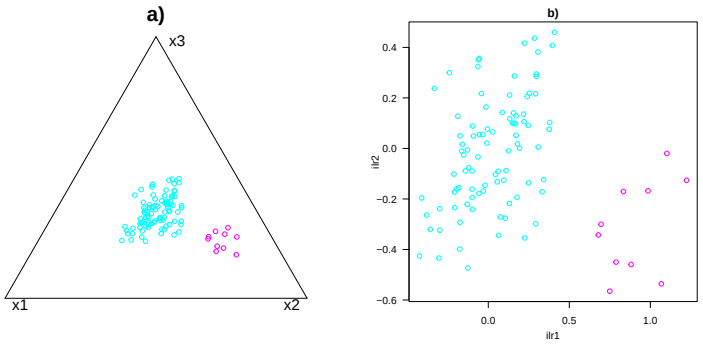


Figure 1: De gauche à droite: le jeu de données dans le simplexe et dans l'espace des coordonnées ilr.

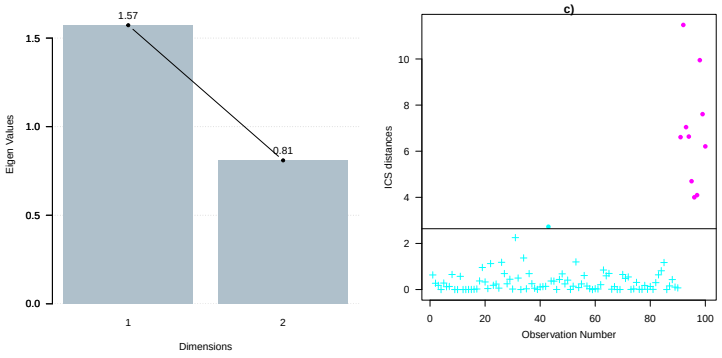


Figure 2: De gauche à droite: les valeurs propres des atypiques avec ICS.

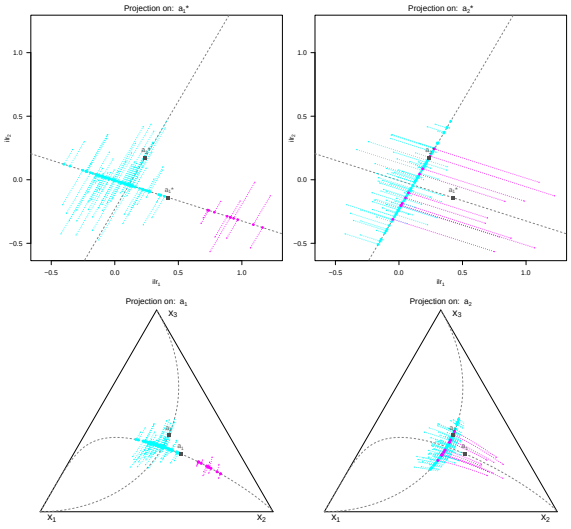


Figure 3: Projections des données sur les axes ICS (ICS 1 à gauche et ICS 2 à droite) dans l'espace ilr (en haut) et dans le simplex (en bas) .

Bibliographie

Archimbaud, A., Nordhausen, K., et Ruiz-Gazen, A. (2018a). ICS for multivariate outlier detection with application to quality control. Computational Statistics & Data Analysis, 128, 184-199.
Archimbaud A., Nordhausen K. et Ruiz-Gazen A., ICSOutlier: Unsupervised Outlier Detection for Low-Dimensional Contamination Structure, The R Journal (2018b) 10:1, pages 234-250.
Muehlmann, C., Fačevicová, K., Gardlo, A., Janečková, H., et Nordhausen, K. (2021). Independent component analysis for compositional data. In Advances in Contemporary Statistics and Econometrics (pp. 525-545). Springer, Cham.
Nordhausen, K., et Ruiz-Gazen, A. (2022). On the usage of joint diagonalization in multivariate statistics. Journal of Multivariate Analysis, 188, 104844.
Pawlowsky-Glahn, V., Egozcue, J. J., et Tolosana-Delgado, R. (2015). Modeling and analysis of compositional data. John Wiley & Sons.
Tyler, D. E., Critchley, F., Dümbgen, L., et Oja, H. (2009). Invariant co-ordinate selection. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 71(3), 549-592.

Extrêmes

On the estimation of extreme quantiles with neural networks

Michaël Allouche [1] , Stéphane Girard [2] & Emmanuel Gobet [3]

[1] *Centre de Mathématiques Appliquées (CMAP), Ecole Polytechnique and CNRS, Université Paris-Saclay, Route de Saclay, 91128 Palaiseau Cedex, France; michael.allouche@polytechnique.edu.*

[2] *Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France; stephane.girard@inria.fr.*

[3] *Centre de Mathématiques Appliquées (CMAP), Ecole Polytechnique and CNRS, Université Paris-Saclay, Route de Saclay, 91128 Palaiseau Cedex, France; emmanuel.gobet@polytechnique.edu.*

Résumé. Dans cette communication, nous proposons une nouvelle paramétrisation des réseaux de neurones à une couche cachée capable d'estimer des quantiles extrêmes à partir de données issues d'une loi à queue lourde. Nous apportons une analyse de l'erreur d'approximation en norme uniforme entre le log-quantile extrême et son approximation par le réseau de neurones. Des expériences numériques sont réalisées sur des données simulées, où les performances de notre méthode sont comparées avec d'autres estimateurs issus de la littérature.

Mots-clés. Statistique des valeurs extrêmes, estimation de quantile, lois à queue lourde, réseau de neurones.

Abstract. In this communication, we propose a new parametrization for one-hidden layer neural networks able to estimate extreme quantiles, starting from data from heavy-tailed distributions. We provide an analysis of the uniform error between the extreme log-quantile and its neural network approximation. Numerical experiments are conducted on simulated data to compare the performance of our method with other estimators from the literature.

Keywords. Extreme value theory, quantile estimation, heavy-tailed distribution, neural networks.

1 Introduction

Let $X_1, \ldots, X_n$ be an i.i.d sample from a cumulative distribution function (c.d.f) F. In this communication, we are interested in the estimation of the quantile function defined by $q(\cdot) := F^{\leftarrow}(\cdot) = \inf\{x \in \mathbb{R} : F(x) \geq \cdot\}$, at the extreme level $1 - \alpha_n$ *i.e.* such that $n\alpha_n \to 0$ as $n \to \infty$. This task is challenging since, first the c.d.f F is unknown,

and second, $q(1-\alpha_n)$ is almost surely asymptotically larger than the sample maxima. Focusing on distributions in the Maximum domain of attraction of Fréchet, it is known from [1, Theorem 1.2.1] and [1, Proposition B.1.9.9] that the tail quantile function $U(t) := q(1-1/t)$ defined for all $t > 1$, is regularly varying with index $\gamma > 0$, *i.e.*

$$U(t) = t^{\gamma} L(t), \tag{1}$$

where γ is the so-called tail-index and L is a slowly varying function at infinity. Consider the anchor level β_n such that $k_n := \lfloor n\beta_n \rfloor \to \infty$ as $n \to \infty$. The idea is to estimate the extreme quantile $q(1-\alpha_n) = U(1/\alpha_n)$ basing on an intermediate quantile $q(1-\beta_n) = U(1/\beta_n)$ and on the (log) extrapolation factor $p_n = \log(\beta_n/\alpha_n) > 0$. Equivalently, letting $z_n = \log(1/\beta_n)$ yields $\alpha_n = \exp(-z_n - p_n)$, and we define the log-spacing function as

$$f(p_n, z_n) = \log U(\exp(z_n + p_n)) - \log U(\exp(z_n)) = \gamma p_n + \varphi(p_n, z_n), \tag{2}$$

with

$$\varphi(p_n, z_n) := \log\left(\frac{L(\exp(p_n + z_n))}{L(\exp(z_n))}\right). \tag{3}$$

The above expression (3) can be evaluated using the well-known second-order condition of the tail quantile function which states that there exist $\gamma > 0, \rho_2 \leq 0$ and a function A_2 positive or negative with $A_2(t) \to 0$ as $t \to \infty$ such that for all $y \geq 1$ [3, Equation (13)]

$$\log U(yt) - \log U(t) = \gamma \log y + A_2(t) \int_1^y y_2^{\rho_2 - 1} \, \mathrm{d}y_2 + o(A_2(t)), \quad \forall y \geq 1 \text{ as } t \to \infty.$$

Equivalently $|A_2|$ is regularly varying with index ρ_2 which drives the bias of most extreme quantile estimators: the larger ρ_2 is, the larger the asymptotic bias (see Figure 1 for an illustration). More generally, one can consider [4] the J-th order condition defined for all $J \geq 2$ as

$$\log U(ty) - \log U(t) = \gamma \log y + \sum_{j=2}^{J} \prod_{\ell=2}^{j} A_\ell(t) R_j(y) + o\left(\prod_{j=2}^{J} A_j(t)\right), \tag{4}$$

as $t \to \infty$ for all $y > 0$, and where for all $j = 2, \ldots, J$:

- $\rho_j \leq 0$,
- A_j is a positive or negative function such that $A_j(t) \to 0$ as $t \to \infty$ and $|A_j|$ is regularly-varying with index ρ_j at infinity,
- for all $y \geq 1$,

$$R_j(y) = \int_1^y y_2^{\rho_2 - 1} \int_1^{y_2} y_3^{\rho_3 - 1} \cdots \int_1^{y_{j-1}} y_j^{\rho_j - 1} \, \mathrm{d}y_j \ldots \mathrm{d}y_3 \, \mathrm{d}y_2. \tag{5}$$

2 Main result

Our objective is to approximate the unknown log-spacing function f in (2) by a neural network (NN) $f_{J,\phi}$ parametrized by ϕ and depending on J such that the NN extreme quantile approximation can be defined as

$$\begin{aligned} q_{J,\phi}(1-\alpha_n \mid 1-\beta_n) &:= q(1-\beta_n)\exp\Big(f_{J,\phi}\big(\log(\beta_n/\alpha_n), \log(1/\beta_n)\big)\Big) \\ &= q(1-\beta_n)(\beta_n/\alpha_n)^{\gamma}\exp\Big(\varphi_{J,\theta}\big(\log(\beta_n/\alpha_n), \log(1/\beta_n)\big)\Big), \end{aligned} \tag{6}$$

where $\phi = (\theta, \gamma)$ is the full set of parameters to be estimated by the NN. Note that assuming $\varphi(\cdot,\cdot) = 0$ in (2) yields back Weissman approximation [5]:

$$q_W(1-\alpha_n|1-\beta_n) = q(1-\beta_n)(\beta_n/\alpha_n)^{\gamma}.$$

Therefore (6) can be seen as an extension of the Weissman estimator where $\varphi_{J,\theta}(\cdot,\cdot)$ is interpreted as a bias-correcting term. Combining (2) and (4), we get

$$\varphi(p_n, z_n) = \sum_{j=2}^{J}\prod_{\ell=2}^{j} A_\ell(\exp(z_n))R_j(\exp(p_n)) + o\left(\prod_{j=2}^{J} A_j(\exp(z_n))\right),$$

and we thus propose to build an approximation $\varphi_{J,\theta}(\cdot,\cdot)$ of $\varphi(\cdot,\cdot)$ by first, discarding the $o(\cdot)$ term, and second building an appropriate approximation of the main term by a one hidden layer NN. More details on the NN construction will be provided during the talk. We establish in the next result the convergence rate of the one hidden layer NN approximation of the extreme quantile in a log-scale.

Theorem 1 *Assume the J-th order condition (4) holds together with a technical assumption on A_j with $j = 2, \ldots, J$ for some $J \geq 2$. Then, there exists a one hidden-layer feedforward neural network approximation $q_{J,\phi}(1-\alpha_n \mid 1-\beta_n)$ of the extreme quantile $q(1-\alpha_n \mid 1-\beta_n)$ with $J(J-1)/2$ neurons and $\mathcal{O}(J^2)$ parameters such that*

$$\alpha_n^{\bar{\rho}_J} \inf_{\phi} \left|\log q(1-\alpha_n \mid 1-\beta_n) - \log q_{J,\phi}(1-\alpha_n \mid 1-\beta_n)\right| \to 0, \tag{7}$$

as $\alpha_n \to 0$, $\beta_n \to 0$, $\beta_n/\alpha_n \to \infty$ and where $\bar{\rho}_J = \rho_2 + \cdots + \rho_J$.

As expected, the rate of convergence of the approximation error is driven by $(\rho_2, \ldots, \rho_J)$ which is coherent with usual extreme-value results based on the second-order condition, see [1, Chapter 4] for examples.

3 Numerical results

The proposed NN estimator built with $J \in \{2, \ldots, 5\}$, is compared on simulated data to the original Weissman estimator (W) and to six other bias-reduced estimators (CW, CH, CH_p, PRB_p, $\text{CH}_{p^\star}$, $\text{PRB}_{p^\star}$) described in [2] of the extreme quantile at level $1-\alpha_n = 1-1/(2n)$ using the performance criterion:

$$\text{RMedSE}\left(\check{q}_n, \alpha_n\right) = \underset{i\in\{1,\ldots,N\}}{\text{median}} \left(\frac{\check{q}_n^{(i)}(1-\alpha_n \mid 1-\beta^\star(i))}{q_n(1-\alpha_n)} - 1 \right)^2,$$

where $\check{q}_n^{(i)}(1-\alpha_n|1-\beta^\star(i))$ denotes the estimator (either the NN estimator or another competitor) computed on a selected anchor level $\beta^\star(i)$ associated with the i-th replication, $i \in \{1, \ldots, N\}$. While we present on Table 1 some results on the Burr distribution, the full study was achieved on six heavy-tailed distributions with $N = n = 500$. It appears that our neural network estimator is an efficient tool for estimating extreme quantiles in difficult heavy-tailed situations where other estimators almost all fail. In particular, in the situation where ρ is large ($\rho = -1/8$), the NN estimator is most of the time the only one to provide realistic values (RMedSE smaller than 1). Note that in other easier situations, our model still produces good estimations. Indeed, it appears on Figure 1 that both on an easy situation ($\rho = -2$) and on a difficult one ($\rho = -1/4$), the NN estimator features a nice stability in terms of bias and RMedSE for a wide range of k_n values. Therefore, even though our proposed method is numerically more expensive than the other ones, it is a very effective estimator for all heavy-tailed situations. Finally, note that, in the above simulations, the NNs are built with 1 to 10 neurons (recall that $J \in \{1, \ldots, 5\}$), which remains acceptable from the computational cost point of view.

Burr	NN	W	CW	CH	CH_p	PRB_p	$\text{CH}_{p^\star}$	$\text{PRB}_{p^\star}$
$\gamma = 1/8$								
$\rho = -1/8$	**0.0392**	-	-	0.5375	0.2713	0.3745	0.3578	0.1203
$\gamma = 1/4$								
$\rho = -1/8$	**0.1567**	-	-	-	-	-	-	0.6357
$\gamma = 1/2$								
$\rho = -1/8$	**0.2847**	-	-	-	-	-	-	-
$\gamma = 1$								
$\rho = -1/8$	**0.3133**	-	-	-	-	-	-	-
$\rho = -1/4$	**0.1962**	-	-	-	-	-	-	0.6617
$\rho = -1/2$	0.2142	-	-	0.0949	0.1021	0.1488	**0.0874**	0.1185
$\rho = -1$	0.1877	-	**0.1289**	0.4120	0.3737	0.3761	0.3658	0.4261
$\rho = -2$	**0.1432**	0.2065	0.2115	0.3394	0.3384	0.2893	0.2933	0.3058

Table 1: RMedSE associated with eight estimators of the extreme quantile at level $1-\alpha_n = 1 - 1/(2n)$ on eight Burr distributions. The best result is emphasized in bold. Results larger than 1 are not displayed.

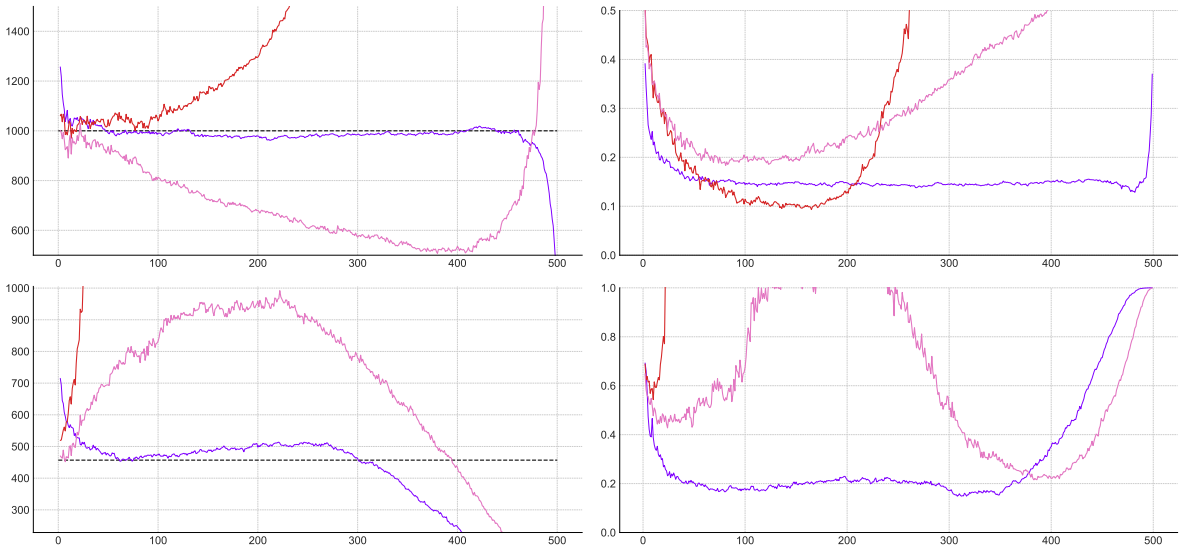


Figure 1: Illustration on simulated data sets of size $n = 500$ from a Burr distribution with $\gamma = 1$ and $\rho \in \{-2, -1/4\}$ (from top to bottom). Median of the estimators (left panel) of the extreme quantile (black dashed line) at level $1 - \alpha_n = 1 - 1/(2n)$ and RMedSE (right panel), as functions of $k_n \in \{2, \ldots, n-1\}$, computed on $N = 500$ replications, associated with W (red), $\text{PRB}_{p^\star}$ (pink), NN (purple).

References

[1] L. de Haan and A. Ferreira. *Extreme value theory*. Springer Series in Operations Research and Financial Engineering. Springer, New York, 2006.

[2] J. El Methni and S. Girard. A refined Weissman estimator for extreme quantiles. `https://hal.inria.fr/hal-03266676`, 2021.

[3] M.I. Gomes, L. de Haan, and L. Peng. Semi-parametric estimation of the second order parameter in statistics of extremes. *Extremes*, 5(4):387–414, 2002.

[4] X.Q. Wang and S.H. Cheng. General regular variation of the n-th order and 2nd order Edgeworth expansions of the extreme value distribution. II. *Acta Mathematica Sinica (English Series)*, 22(1):27–40, 2006.

[5] I. Weissman. Estimation of parameters and large quantiles based on the k largest observations. *Journal of the American Statistical Association*, 73(364):812–815, 1978.

Théorème de Donsker pour des mesures empiriques locales sur des régions aléatoires

Benjamin Bobbia [1] & Clément Dombry [2] & Davit Varron [3]

[1] *Léonard de Vinci Pôle Universitaire, Research Center, 92 916 Paris La Défense, France.*
benjamin.bobbia@devinci.fr
[2] *Université de Franche-Comté, 16 route de Gray 25 000 Besançon*
clement.dombry@univ-fcomte.fr
[3] *Université de Franche-Comté, 16 route de Gray 25 000 Besançon*
davit.varron@univ-fcomte.fr

Résumé. Les processus empiriques sont aujourd'hui des objets bien connus. Une des raisons qui a poussée au développement de l'étude des processus empiriques est qu'il est possible, dans de nombreuses modélisations, d'écrire les estimateurs comme images de mesures empiriques. Dans ce travail nous regardons le cas particulier des mesures empiriques locales, c'est-à-dire, la mesure empirique construite sur un sous-échantillon d'observations conditionnées à un être dans une certaine partie de l'espace. De nombreux résultats existent pour ce type de mesure, mais que peut-on-dire si la partie de l'espace en question dépend des données ? Il est possible de s'en sortir au prix d'une grande technicité et de conditions de régularité supplémentaires, le propos de ce travail et de présenter un cadre général pour l'étude de ces mesures empiriques particulières permettant d'obtenir des résultats asymptotiques à moindre coût (technique et conditions). Les résultats sont présentés ici au travers du prisme de la théorie de valeurs extrêmes et nous proposons un exemple qui met en lumière la pertinence de cette approche notamment lorsqu'elle est couplée avec des méthodes issues du transport optimal.

Mots-clés. Mesure empirique locale, Extrêmes, théorème de Donsker.

Abstract. Nowadays, empirical processes are well known objects. A reason that push forward theirs studies is that, in many models, we can write the estimators as images of empirical measures. In this work we investigate the case of local empirical measures built over a sub-sample with data conditioned to be in a certain area. There exists numerous results about such question, but what can we say if the area of interest is data driven ? In the present work we present a general framework which allows to derive asymptotic results for this particular empirical measures (with "low" cost in terms of technicality and assumptions). This approach is mainly presented through the prism of extreme values theory. And we present an example illustrating how this empirical conditional measure assumption work well with optimal transport methods.

Keywords. Local empirical measures, Extremes, Donsker's theorem.

Cette communication présente certains résultats issus de l'article en prépublication *A Donsker and Glivencko-Cantelli theorem for random measures linked to extreme value theory* qu'il est possible de trouver à l'adresse
https://hal.archives-ouvertes.fr/hal-03402380/document.

1 La mesure empirique locale et motivations

Considérons P_0 une mesure sur un espace $\mathcal{X}$ ainsi qu'un échantillon $(X_1, \dots, X_n)$ i.i.d de loi P_0. On définit alors la mesure empirique, avec pour but d'estimer P_0, par

$$\mathbb{P}_n := \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}.$$

Ainsi, pour une fonction φ suffisamment régulière, $\varphi(\mathbb{P}_n)$ est censée être proche de $\varphi(P_0)$ au sens de la convergence en probabilité. Il apparaît alors naturel de s'intéresser à cette quantité dans le cadre de l'estimation. Cependant dans certains contextes comme l'estimation par plus proches voisins ou la théorie des valeurs extrêmes, l'estimation met en jeu uniquement un sous-échantillon de $(X_1, \dots, X_n)$, on construit alors une mesure empirique, dite locale, sur ce sous-échantillon d'intérêt. Cette mesure est définie comme suit. Soit $(A_n)_{n\geq 1}$ une suite d'ensembles tel que $p_n := P_0(A_n)$ tend vers 0, on définit alors

$$\mathbb{P}_n^{loc} := \frac{1}{card(\{i \text{ tel que } X_i \in A_n\})} \sum_{i=1}^{n} \mathbb{1}_{A_n}(X_i)\delta_{X_i}.$$

L'utilisation d'une telle mesure apparaît notamment dans l'approche peak-over-thresholds introduite par Balkema et de Haan (1974) en théorie des valeurs extrêmes. Cette approche consiste à utiliser uniquement les données supérieures à un certain seuil y_n (tendant vers l'infini) pour obtenir des informations sur la queue de distribution. Dans ce cas notre mesure empirique locale s'écrit

$$\mathbb{P}_n^{ext} := \frac{1}{card(\{i \text{ tel que } X_i > y_n\})} \sum_{X_i > x_n} \delta_{X_i/x_n}.$$

Dans cette approche, il est souvent préférable d'utiliser comme seuil un point de l'échantillon, en particulier le k-ième plus grand que l'on note $X_{n-k:n}$ (avec $k = k_n$ convergeant vers l'infini mais moins vite que n). Ce qui revient à considérer une mesure empirique locale avec un ensemble A_n dépendant des données, objet centrale de ce travail.

2 Résultats

Un des objectifs du présent travail est de fournir un cadre général dans lequel il sera pertinent d'étudier les mesures empiriques locales introduites précédemment. L'hypothèse

centrale de cette approche est l'**hypothèse de mesure empirique conditionnelle**. Nous la présentons ici ainsi qu'un résultat de type Donsker s'appuyant sur cette dernière. La partie suivante présente un exemple dans lequel nous pouvons vérifier que cette hypothèse est satisfaite puis un exemple concernant une application plus général d'une telle hypothèse.

Hypothèse de mesure empirique conditionnelle. *Soit $(P_n)_{n\geq 1}$ une suite de mesures aléatoires sur $\mathcal{X}$. On dit que P_n satisfait la condition de mesure empirique conditionnelle s'il existe une variable aléatoire U_n prenant ses valeurs dans un espace métrique (E, d), muni de sa tribu borélienne et d'une mesure μ, et une famille de probabilités $(P_u)_{u\in E}$ sur $\mathcal{X}$, telles que pour presque tout $u \in E$ on a, sachant que $U_n = u$,*

$$P_n \overset{\mathcal{L}}{=} \frac{1}{k_n} \sum_{i=1}^{k_n} \delta_{X_i^{(u)}}$$

avec k_n un entier et $(X_1^{(u)}, \ldots, X_{k_n}^{(u)})$ i.i.d de loi P_u.

Dans le contexte des extrêmes, il est possible de montrer que la suite de mesures empiriques locales $(\mathbb{P}_n^{ext})_{n\geq 1}$ satisfait cette hypothèse de mesure empirique conditionnelle avec $U_n = X_{n-k_n:n}$ et $P_u(x) = P_0(x/u|X > u)$. Plus de détail sur les applications sont données dans la partie suivante. De plus, des mesures empiriques satisfaisant cette hypothèse peuvent être rencontrées en abordant des problèmes liés au plus proches voisins ou aux variations régulières.
Sous cette hypothèse sur la famille $(P_n)_{n\geq 1}$ et lorsque la suite U_n converge en loi vers un certain U dans E, il est possible d'établir des résultats de convergence pour le processus empirique indexé par une classe de fonctions $\mathcal{F}$

$$\mathbb{G}_n(f) := \sqrt{k_n}(P_n(f) - P_U(f)), \quad f \in \mathcal{F},$$

où $P(f)$ désigne $\int f \mathrm{d}P$.

Théorème 1. *Sous des hypothèses classiques sur la classe $\mathcal{F}$ et de régularité sur l'application $u \mapsto \{f \mapsto P_u(f)\}$ on a*

$$\mathbb{G}_n \overset{\mathcal{L}}{\to} \mathcal{W}, \qquad \text{dans } l^\infty(\mathcal{F}),$$

avec $\mathcal{W}$ le P_U-Pont Brownien indexé par $\mathcal{F}$. C'est-à-dire $\mathcal{W}$ est un processus Gaussien centré dont la covariance est donnée par $Cov(\mathcal{W}(f_1), \mathcal{W}(f_2)) = P_U(f_1 f_2) - P_U(f_1)P_U(f_2)$ pour tout f_1 et $f_2 \in \mathcal{F}$. Nous renvoyons le lecteur vers la monographie de A. W. van der Vaart et J. A. Wellner pour plus d'informations sur de tels processus.

3 Exemples de domaines d'applications

Cette section détaille deux applications parmi celles envisagées dans ces travaux. La première concerne l'estimation de l'indice des valeurs extrêmes dans le cas d'observations

réelles à queues lourdes. La seconde application met en lumière le fait que l'**hypothèse de mesure empirique conditionnelle** permet d'appliquer aux mesures empiriques locales des résultats de couplage normalement réservés aux mesures sur un échantillon d'observations indépendantes.

3.1 Théorie des valeurs extrêmes

Cette exemple est important, non pas au sens de l'apport de nouvelles connaissances en tant que telles, mais au sens où c'est celui-ci qui a initié ce travail. Il a aussi l'avantage de mettre en évidence la pertinence de l'**hypothèse de mesure empirique conditionnelle** dans un contexte connu, ce qui la rend plus facilement appréhendable dans une première approche.
Comme nous l'avons évoqué précédemment, $(\mathbb{P}_n^{ext})_{n\geq 1}$ vérifie l'**hypothèse de mesure empirique conditionnelle**, c'est-à-dire :

Lemme 1 (B., Dombry, Varron (2021))**.** *Supposons que les X_i ont une distribution commune continue.* ***Conditionnellement*** *à $X_{n-k:n} = x$, on a l'égalité suivante en loi*

$$\mathbb{P}_n^{ext} \overset{\mathcal{L}}{=} \frac{1}{k}\sum_{i=1}^{k} \delta_{X_i^{(x)}}$$

où k est un entier et $(X_1^{(x)}, \dots, X_k^{(x)})$ est un échantillon i.i.d de distribution commune P_x définie pour tout borélien B par $P_x(B) = \mathbb{P}(X \in B \mid X > x)$.

De plus on sait que, sous des hypothèses de variations régulières, P_x converge vers une distribution de Pareto lorsque x tend vers l'infini (Balkema et de Haan, 1975). Ainsi, le théorème 1 permet d'affirmer que pour une classe de fonctions $\mathcal{F}$ appropriée (Une classe VC de fonctions L^2 par exemple, voir la monographie de A. W. van der Vaart et J. A. Wellner pour plus de détails), on a

$$\sqrt{k}(\mathbb{P}_n^{ext} - P_0) \xrightarrow{\mathcal{L}} \mathcal{W}_{P_0} \quad \text{dans } l^\infty(\mathcal{F}),$$

avec P_0 la distribution de Pareto et $\mathcal{W}_{P_0}$ le P_0-Pont Brownien indexé par $\mathcal{F}$.
Ce résultat nous permet l'estimation du bien nommé indice des valeurs extrêmes γ qui renseigne sur le comportement de la queue de distribution de P_0 (plus γ est élevé, plus la queue de P_0 est lourde). En effet, pour les distributions à queues lourdes γ peut être estimé par le célèbre estimateur de Hill (1975)

$$\hat{\gamma}_n := \frac{1}{k}\sum_{Xi > X_{n-k:n}} \log\left(\frac{X_i}{X_{n-k:n}}\right) = \varphi(\mathbb{P}_n^{ext}),$$

avec $\varphi : P \mapsto \int \log \mathrm{d}P$.
Donc la normalité asymptotique de l'estimateur de Hill est une conséquence du théorème 1 suivit de la delta-méthode.

3.2 Transport optimal

Dans cette application, nous montrons comment l'**hypothèse de mesure empirique conditionnelle** permet d'appliquer aux mesures empiriques locales des résultats de couplage normalement réservés aux mesures sur un échantillon indépendant. Ce résultat de couplage met en jeu la distance de Wasserstein. On rappel que la distance de Wasserstein d'ordre $1 \leq p < \infty$ entre deux mesures de probabilités P et P^* (sur un espace mesuré $(\mathcal{X}, d)$) est défini comme suit

$$W_p(P, P^*) = \inf\{\mathbb{E}(d(X, X^*)^p)^{1/p}, X \sim P, X^* \sim P^*\}.$$

Nous considérons le cas particulier de la distance de Wasserstein entre deux mesures Π et Π^* sur $\mathcal{M}(\mathcal{X})$ (i.e l'ensemble des mesures sur $\mathcal{X}$) muni de la distance de Wasserstein. Cette distance est appelée distance de Wasserstein d'ordre deux et est définie comme suit

$$W_p^2(\Pi, \Pi^*) = \inf\{\mathbb{E}(d(P, P^*)^p)^{1/p}, P \sim \Pi^*, P^* \sim \Pi^*\}.$$

Muni de ces distances, il est possible de démontrer un résultat de couplage entre mesures empiriques. Plus précisément, soit $X_1, ..., X_n$ et $X_1^*, ..., X_n^*$ des échantillons i.i.d de distributions P and P^*. Alors, les mesures empiriques

$$\Pi_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i} \quad \text{et} \quad \Pi_n^* = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i^*}$$

satisfont le théorème suivant.

Théorème 2 (B, Dombry, Varron. (2021))**.** *Soit $p \in [1, \infty)$. Alors,*

$$W_p^{(2)}(P_{\Pi_n}, P_{\Pi_n^*}) \leq W_p(P, P^*),$$

avec P_{Π_n} et $P_{\Pi_n^*}$ les lois respectives de Π_n et Π_n^*.

En choisissant judicieusement les mesures à coupler, ce résultat peut par exemple permette de démontrer la normalité asymptotique d'estimateurs. Cependant, il ne peut pas s'appliquer directement à une mesure empirique locale. En effet ce résultat de couplage met en jeu des mesures empiriques sur des échantillons i.i.d. Or ce n'est pas le cas lorsque l'on considère des mesures empiriques locales dont le choix du sous échantillon dépend des données. Néanmoins, si la suite de mesures empiriques $(\mathbb{P}_n^{loc})_{n\geq 1}$ satisfait l'**hypothèse de mesure empirique conditionnelle**, on a alors conditionnellement à $A_n = A$ avec $A \subset \mathcal{X}$,

$$\mathbb{P}_n^{loc} \overset{\mathcal{L}}{=} \frac{1}{k}\sum_{i=1}^{k} \delta_{X_i^{(A)}} = \Pi_A$$

où k est un entier et $(X_1^{(A)}, \ldots, X_k^{(A)})$ est un échantillon i.i.d de distribution commune P_A définie pour tout borélien B par $P_A(B) = \mathbb{P}(X \in B \mid X \in A)$.

Nous faisons ici un abus de notation afin de rendre les résultats plus lisible. Ici, l'ensemble A_n est paramétré par une partie des données et la suite $(\mathbb{P}_n^{loc})_{n\geq 1}$ vérifie l'**hypothèse de mesure empirique conditionnelle** avec U_n le vecteur des variables qui paramétrisent A_n et P_U la loi jointe correspondante. Dans un soucis de clarté nous identifions (en termes de notation) A_n et U_n. Le conditionnement précédent est alors un conditionnement sur les valeurs des variables qui servent à définir l'ensemble A_n. Ainsi, conditionner par $A_n = A$ signifie conditionner les variables servant à paramétrer A_n à être égales aux valeurs paramétrant l'ensemble A, ce dernier n'étant pas aléatoire.

Donc conditionnellement à $A_n = A$, on a

$$W_p^{(2)}(P_{\Pi_n}, P_{\Pi_n^*}) = W_p^{(2)}\left(P_{\Pi_A}, P_{\Pi_n^*}\right).$$

On en déduit alors une version du théorème 2 pour les mesures empiriques locales.

Théorème 3. *Soit $p \in [1, \infty)$ et P^* une mesure de probabilité. Si la suite de mesures $(\mathbb{P}_n^{loc})_{n\geq 1}$ satisfait l'**hypothèse de mesure empirique conditionnelle** pour une mesure P_A, alors*

$$\mathbb{E}\left(W_p^{(2)}(P_{\Pi_n^{loc}}, P_{\Pi_n^*}) \mid A_n = A\right) \leq W_p(P_A, P^*).$$

Où Π_n^{loc} désigne la loi de $\mathbb{P}_n^{loc}$.

Bibliographie

Balkema, A. A. and de Haan, L. (1974). Residual Life Time at Great Age. *The Annals of Probability*, 2(5):792-804.

Hill, B. M. (1975). A Simple General Approach to Inference About the Tail of a Distribution. *The Annals of Statistics*, 3(5):119-131.

Bobbia, B. and Dombry, C. and Varron, D. (2021). The coupling method in extreme value theory. *Bernoulli*, 27(3):1824-1850.

van der Vaart, A. W. and Wellner, J.A. (1996). *Weak convergence and empirical processes.* Springer Series in Statistics. Springer-Verlag, New York.

Non-parametric estimator of a multivariate madogram for missing-data and extreme value framework

Alexis Boulin [1] & Elena Di Bernardino [1] & Thomas Laloë [1] & Gwladys Toulemonde [2]

[1] *Laboratoire J.A. Dieudonné, UMR CNRS 7351, Université Côte d'Azur, Nice, France*
[2] *Univ Montpellier, CNRS, Inria, Montpellier, France*

Résumé. Modéliser la dépendance entre maxima est un sujet d'intérêt dans les domaines d'application d'analyse du risque. Dans cet objectif, la copule de valeurs extrêmes, caractérisée par le madogramme, peut être utilisée comme une description de la structure de dépendance. Concrètement, la famille des distributions à valeurs extrêmes est très riche et survient naturellement comme la limite composante par composante des maxima préalablement normalisés. Dans cette présentation, nous étudions l'estimation non paramétrique du madogramme lorsque les données sont absentes complètement au hasard. Nous fournissons un théorème de la limite centrale fonctionnelle pour l'estimateur considéré du madogramme, correctement normalisé, vers un processus Gaussien tendu pour lequel la fonction de covariance dépend des probabilités de perte de la donnée. L'expression explicite de la variance asymptotique est aussi donnée. Nos résultats sont illustrés dans une étude numérique lorsque la taille d'échantillon est finie.

Mots-clés. Madogramme, Copule de valeurs extrêmes, Absence complètement au hasard, Estimation non paramétrique.

Abstract. The modeling of dependence between maxima is an important subject in several applications in risk analysis. To this aim, the extreme value copula function, characterised via the madogram, can be used as a margin-free description of the dependence structure. From a practical point of view, the family of extreme value distributions is very rich and arises naturally as the limiting distribution of properly normalised componentwise maxima. In this talk, we investigate the nonparametric estimation of the madogram where data are completely missing at random. We provide the functional central limit theorem for the considered multivariate madrogram correctly normalized, towards a tight Gaussian process for which the covariance function depends on the probabilities of missing. Explicit formula for the asymptotic variance is also given. Our results are illustrated in a finite sample setting with a simulation study.

Keywords. Madogram, Extreme value copula, Missing Completely At Random (MCAR), Nonparametric estimation.

1 Introduction

Soit $(\Omega, \mathcal{A}, \mathbb{P})$ un espace probabilisé où vit $\mathbf{X} = (X_1, \ldots, X_d)$ un vecteur aléatoire à valeurs dans l'espace mesurable $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ avec $d \geq 2$. Ce vecteur aléatoire est distribué selon F et ses marges sont notées $F_j(x) = \mathbb{P}\{X_j \leq x\}$ pour tout $x \in \mathbb{R}$ et $j \in \{1, \ldots, d\}$. Une fonction $C : [0,1]^d \to [0,1]$ est une copule s'il s'agit de la restriction en $[0,1]^d$ d'une fonction de distribution dont ses marges sont distribuées uniformément le long du segment $[0,1]$. Depuis les travaux de Sklar, il est de connaissance commune que chaque fonction de distribution F peut être décomposée comme $F(\mathbf{x}) = C(F_1(x_1), \ldots, F_d(x_d))$ pour $\mathbf{x} \in \mathbb{R}^d$ et la copule C est unique si les marges sont continues.

L'étude de la notion de copule dans le cadre des extrêmes mène à la copule de valeurs extrêmes définie de la façon suivante :

$$C(\mathbf{u}) = \exp\left(-\ell(-\ln(u_1), \ldots, -\ln(u_d))\right), \quad \mathbf{u} \in (0,1]^d, \tag{1}$$

où $\ell : [0,\infty)^d \to [0,\infty)$ est la fonction de dépendance caudale qui est une fonction convexe, homogène d'ordre un et qui satisfait $\max(z_1, \ldots, z_d) \leq \ell(z_1, \ldots, z_d) \leq z_1 + \cdots + z_d$. Notons $\Delta^{d-1} = \{\mathbf{w} \in [0,1]^d, w_1 + \cdots + w_d = 1\}$ le simplexe unité de $\mathbb{R}^d$. Par son caractère homogène, la fonction ℓ est caractérisée par la fonction de dépendance de Pickands qui correspond à la restriction de ℓ au simplexe unité Δ^{d-1}, à proprement dit

$$\ell(z_1, \ldots, z_d) = (z_1 + \cdots + z_d) A(w_1, \ldots, w_d), \quad w_j = \frac{z_j}{z_1 + \cdots + z_d}.$$

Inspiré par la notion de variogramme en géostatistique, le λ-madogramme fut défini par [3] afin de capturer les dépendances extrémales bivariées de processus spatiaux. Cette définition fut ensuite généralisée en plus grandes dimensions via le **w**-madogramme définie dans [5] :

$$\nu(\mathbf{w}) = \mathbb{E}\left[\bigvee_{j=1}^{d} \{F_j(X_j)\}^{1/w_j} - \frac{1}{d}\sum_{j=1}^{d} \{F_j(X_j)\}^{1/w_j}\right], \quad \mathbf{w} \in \Delta^{d-1}, \tag{2}$$

si $w_j = 0$ et $0 < u < 1$, alors $u^{1/w_j} = 0$ par convention. Cette quantité caractérise la fonction de dépendance de Pickands via la relation formulée par la Proposition 2.2 de [5],

$$A(\mathbf{w}) = \frac{\nu(\mathbf{w}) + c(\mathbf{w})}{1 - \nu(\mathbf{w}) - c(\mathbf{w})}, \quad c(\mathbf{w}) = d^{-1} \sum_{j=1}^{d} w_j/(1 + w_j).$$

Dans cette communication, les principaux résultats obtenus que nous proposons de présenter sont la définition d'un estimateur du **w**-madogramme en (2) dans le cadre de données manquantes et l'étude de son comportement asymptotique. Un cadre multivarié ($d \geq 2$) avec données manquantes et une structure de dépendance décrite par une copule de

valeurs extrêmes seront considérés dans cette présentation. Nous évoquerons un théorème de la limite centrale fonctionnelle qui fournit la convergence faible du processus empirique centré du madogramme multivarié vers un processus Gaussien tendu dont la fonction de covariance va dépendre des probabilités d'observations, l'énonciation de ce théorème peut être retrouvée en [1]. Lorsque la trajectoire de ce processus empirique est fixée, nous montrons dans la Proposition 1 que cette variable aléatoire converge vers une loi normale centrée dont il est possible de fournir une expression explicite de la variance asymptotique. Enfin, des résultats numériques permettront de vérifier nos contributions sur un échantillon de taille finie.

2 Cadre de données manquantes

On suppose ici que la copule C est de valeurs extrêmes définie en (1). On considère un échantillon $\mathbf{X}_1, \ldots, \mathbf{X}_n$ de $\mathbf{X}$. En présence de données manquantes, on n'observe pas complètement le vecteur $\mathbf{X}_i$ pour $i \in \{1, \ldots, n\}$. Pour formaliser, on considère $\mathbf{I}_i \in \{0,1\}^d$ indiquant, $\forall j \in \{1, \ldots, d\}$, $I_{i,j} = 0$ si $X_{i,j}$ n'est pas observé et $I_{i,j} = 1$ si $X_{i,j}$ est observé. On introduit alors le vecteur aléatoire $\tilde{\mathbf{X}}_i$ à valeurs dans l'espace produit $\bigotimes_{j=1}^{d}(\mathbb{R} \cup \{\texttt{NA}\})$ tel que

$$\tilde{X}_{i,j} = X_{i,j} I_{i,j} + \texttt{NA}(1 - I_{i,j}), \quad i \in \{1, \ldots, n\},\ j \in \{1, \ldots, d\}.$$

On suppose alors que nous observons un $2d$-uplet tel que :

$$(\mathbf{I}_i, \tilde{\mathbf{X}}_i), \quad i \in \{1, \ldots, n\}. \tag{3}$$

Pour chaque $i \in \{1, \ldots, n\}$, $\mathbf{I}_i$ est distribuée identiquement et indépendamment selon $\mathbf{I} = (I_1, \ldots, I_d)$ où I_j est distribuée selon une variable aléatoire de Bernoulli avec probabilité $p_j = \mathbb{P}\{I_j = 1 | \mathbf{X}\}$ pour $j \in \{1, \ldots, d\}$. On note p la probabilité d'observer complètement une réalisation de $\mathbf{X}$, autrement dit $p = \mathbb{P}\{I_1 = 1, \ldots, I_d = 1 | \mathbf{X}\}$. Concernant le mécanisme de données manquantes, nous formulons l'hypothèse suivante :

Hypothèse A. *Les vecteurs* $\boldsymbol{I}$ *et* $\boldsymbol{X}$ *sont indépendants, ainsi*

$$p_j = \mathbb{P}\{I_j = 1\},\ j \in \{1, \ldots, d\}, \quad p = \mathbb{P}\{I_1 = 1, \ldots I_d = 1\}.$$

Sous l'Hypothèse A, nous pouvons définir des estimateurs consistants des distributions marginales et jointes. Par convenance, nous écrivons $\{\tilde{\mathbf{X}}_i \leq \mathbf{x}_i\} = \{\tilde{\mathbf{X}}_{i,1} \leq x_1, \ldots, \tilde{\mathbf{X}}_{i,d} \leq x_d\}$,

$$\hat{F}_{n,j}(x) = \frac{\sum_{i=1}^{n} \mathbb{1}_{\{\tilde{X}_{i,j} \leq x\}} I_{i,j}}{\sum_{i=1}^{n} I_{i,j}},\ \forall x \in \mathbb{R}, \quad \hat{F}_n(\mathbf{x}) = \frac{\sum_{i=1}^{n} \mathbb{1}_{\{\tilde{\mathbf{X}}_i \leq \mathbf{x}\}} \Pi_{j=1}^{d} I_{i,j}}{\sum_{i=1}^{n} \Pi_{j=1}^{d} I_{i,j}},\ \forall \mathbf{x} \in \mathbb{R}^d. \tag{4}$$

Avec ces notations, nous rappelons l'estimateur de la copule hybride formulée par [8], *i.e.*

$$\hat{C}_n^{\mathcal{H}}(\mathbf{u}) = \hat{F}_n(\hat{F}_{n,1}^{\leftarrow}(u_1), \ldots, \hat{F}_{n,d}^{\leftarrow}(u_d)), \quad \mathbf{u} \in [0,1]^d,$$

où $F^{\leftarrow}$ correspond à l'inverse généralisé de la fonction F, *i.e.* $F^{\leftarrow}(u) = \inf\{x \in \mathbb{R} | F(x) \geq u\}$ avec $0 < u < 1$. L'estimateur de la copule hybride est une extension de l'estimateur classique non paramétrique de la copule $\hat{C}_n$ (voir, *e.g.* [4] pour une définition) permettant que les marges de l'estimateur ne soient pas nécessairement égales aux estimateurs des marges ce qui est une propriété très appropriée dans le cadre plus général de données manquantes. Nous formulons une hypothèse permettant de garantir la convergence du présent processus empirique $\sqrt{n}(\hat{C}_n^{\mathcal{H}} - C)$:

Hypothèse B.

1. *La fonction de distribution F a des marges continues $F_1, \ldots, F_d$.*
2. *Pour chaque $j \in \{1, \ldots, d\}$, les dérivées partielles $\dot{\ell}_j$ de ℓ par rapport à x_j existent et sont continues sur l'ensemble $\{x \in [0, \infty)^d : x_j > 0\}$.*

L'hypothèse B1 est classique dans la littérature des copules sans laquelle une quelconque convergence faible du processus $\sqrt{n}(\hat{C}_n - C)$ ne peut être espérée. L'hypothèse B2 a été formulé par [7] permettant de généraliser la condition formulée par [4] demandant initialement que les dérivées partielles de la copule existent et soient continues sur $[0, 1]^d$. Des conditions plus faibles sont toujours le fruit de nouvelles recherches, voir par exemple [2] où l'hypothèse que l'ensemble des points où les dérivées partielles de la copule existent et sont continues soit de mesure de Lebesgue 1 est demandée. Nous proposons à présent un estimateur du $\mathbf{w}$-madogramme en (2) dans le cadre de données manquantes complètement au hasard.

Definition 1. *Soit $(\boldsymbol{I}_i, \tilde{\mathbf{X}}_i)_{i=1}^n$ un échantillon donné par l'Equation (3), l'estimateur hybride du $\mathbf{w}$-madogramme en (2) est défini par*

$$\hat{\nu}_n^{\mathcal{H}}(\mathbf{w}) = \frac{1}{\sum_{i=1}^n \Pi_{j=1}^d I_{i,j}} \sum_{i=1}^n \left[\left(\bigvee_{j=1}^d \left\{ \hat{F}_{n,j}(\tilde{X}_{i,j}) \right\}^{1/w_j} - \frac{1}{d} \sum_{j=1}^d \left\{ \hat{F}_{n,j}(\tilde{X}_{i,j}) \right\}^{1/w_j} \right) \Pi_{j=1}^d I_{i,j} \right], \tag{5}$$

où $\hat{F}_{n,j}(x)$ sont définis en (4).

Comme classiquement rencontré dans la littérature, voir *e.g.* [6], cet estimateur ne vérifie pas la condition $\hat{\nu}_n(\mathbf{e}_j) = (d-1)/2d$ où $\mathbf{e}_j$ est un vecteur de la base canonique pour $j \in \{1, \ldots, d\}$. Néanmoins, une correction préalable permettant d'y remédier est définie ci-dessous.

Definition 2. *Soit $(\boldsymbol{I}_i, \tilde{\mathbf{X}}_i)_{i=1}^n$ un échantillon donné en (3) et $\hat{\nu}_n^{\mathcal{H}}$ l'estimateur hybride en (5). Considérons par $\lambda_1, \ldots, \lambda_d : \Delta^{d-1} \to \mathbb{R}$ des fonctions continues vérifiant $\lambda_j(\boldsymbol{e}_k) = \delta_{jk}$ (delta de Kronecker) pour $j, k \in \{1, \ldots, d\}$, on définit la version corrigée de l'estimateur*

hybride du $\mathbf{w}$*-madogramme par*

$$\hat{\nu}_n^{\mathcal{H}*}(\mathbf{w}) = \hat{\nu}_n^{\mathcal{H}}(\mathbf{w}) - \sum_{j=1}^{d} \frac{\lambda_j(\mathbf{w})(d-1)}{d} \left[\frac{1}{\sum_{i=1}^{n} \Pi_{j=1}^{d} I_{i,j}} \sum_{i=1}^{n} \left(\left\{ \hat{F}_{n,j}(\tilde{X}_{i,j}) \right\}^{1/w_j} \Pi_{j=1}^{d} I_{i,j} \right) - \frac{w_j}{1+w_j} \right]. \quad (6)$$

Sous les hypothèses A et B, nous démontrons que le procesus $\sqrt{n}(\hat{\nu}_n^{\mathcal{H}} - \nu)$ convergent faiblement vers un processus Gaussien dans $\ell^\infty(\Delta^{d-1})$, l'espace des fonctions bornées sur le simplexe, voir Théorème 1 de [1]. Lorsque $\mathbf{w} \in \Delta^{d-1}$ est fixé, en utilisant le précédent résultat, nous pouvons établir la convergence en loi des variables aléatoires $\sqrt{n}(\hat{\nu}_n^{\mathcal{H}}(\mathbf{w}) - \nu(\mathbf{w}))$ vers une Gaussienne centrée dont l'expression fermée est donnée. Les mêmes conclusions peuvent être obtenues pour $\sqrt{n}(\nu^{\mathcal{H}*} - \nu)$.

Proposition 1. *Soit* $\mathbf{p} = (p_1, \ldots, p_d, p)$ *et* $\mathbf{w} \in \Delta^{d-1}$*, sous les hypothèses A et B, nous avons*

$$\sqrt{n}\left(\hat{\nu}_n^{\mathcal{H}}(\mathbf{w}) - \nu(\mathbf{w})\right) \xrightarrow[n\to\infty]{d} \mathcal{N}\left(0, \mathcal{S}^{\mathcal{H}}(\mathbf{p}, \mathbf{w})\right), \sqrt{n}\left(\hat{\nu}_n^{\mathcal{H}*}(\mathbf{w}) - \nu(\mathbf{w})\right) \xrightarrow[n\to\infty]{d} \mathcal{N}\left(0, \mathcal{S}^{\mathcal{H}*}(\mathbf{p}, \mathbf{w})\right).$$

De plus, l'expression des variances asymptotiques a été obtenue (voir [1]).

3 Résultats numériques

Figure 1: $\mathcal{E}_n^{\mathcal{H}}$ (première colonne) et $\mathcal{E}_n^{\mathcal{H}*}$ (deuxième colonne) données par (7), fonction de $\mathbf{w}$, pour le modèle logistique avec $\theta = 2$. Les courbes de niveaux de la variance asymptotique (à gauche) et des contreparties empiriques (à droite) sont représentées.

Dans cette section, nous illustrons nos résultats concernant l'expression de la variance asymptotique à travers des simulations numériques. Pour ce faire, nous comparons la variance asymptotique en Proposition 1 avec des contreparties empiriques obtenues via des simulations de Monte-Carlo. Nous considérons le modèle logistique défini par la fonction de dépendance suivante :

$$A(w_1, \ldots, w_d) = \left(\sum_{j=1}^{d} w_j^{\theta} \right)^{1/\theta}, \quad \theta \in [1, \infty).$$

Pour l'expérience numérique, la version corrigée du $\mathbf{w}$-madogramme est calculée en utilisant $\lambda_j(\mathbf{w}) = \mathbf{e}_j$ et pour $d = 3$. La contrepartie empirique de la variance asymptotique donnée en Proposition 1 est calculée pour chaque élément $\mathbf{w}$ d'un certain grillage du simplexe Δ^{d-1}. Pour cet élément $\mathbf{w}$, $n_{iter} = 100$ échantillons de taille $n = 512$ sont générés du modèle Logistique. Un estimateur du $\mathbf{w}$-madogramme est ainsi associé à chaque échantillon. Nous calculons alors la variance empirique de l'erreur normalisée, à proprement dit :

$$\mathcal{E}_n^{\mathcal{H}}(\mathbf{w}) := \widehat{Var}\left(\sqrt{n}\left(\hat{\boldsymbol{\nu}}_{\boldsymbol{n}}^{\boldsymbol{\mathcal{H}}}(\mathbf{w}) - \nu(\mathbf{w})\right)\right), \quad \mathcal{E}_n^{\mathcal{H}*}(\mathbf{w}) := \widehat{Var}\left(\sqrt{n}\left(\hat{\boldsymbol{\nu}}_{\boldsymbol{n}}^{\boldsymbol{\mathcal{H}*}}(\mathbf{w}) - \nu(\mathbf{w})\right)\right), \quad (7)$$

où $\hat{\boldsymbol{\nu}}_{\boldsymbol{n}}^{\boldsymbol{\mathcal{H}}}$ et $\hat{\boldsymbol{\nu}}_{\boldsymbol{n}}^{\boldsymbol{\mathcal{H}*}}$ sont les vecteurs composés des n_{iter} estimateurs hybrides et corrigées (voir (5) et (6)) du $\mathbf{w}$-madogramme, respectivement. Les résultats de cette expérience sont données par la Figure 1.

Dans les deux cas, hybride comme corrigé, nous observons une adéquation entre la forme explicite de la variance asymptotique donnée à la Proposition 1 avec la contrepartie empirique obtenue via simulation.

References

[1] A. Boulin, E. D. Bernardino, T. Laloë, and G. Toulemonde. Non-parametric estimator of a multivariate madogram for missing-data and extreme value framework, 2021.

[2] A. Bücher, J. Segers, and S. Volgushev. When uniform weak convergence fails: Empirical processes for dependence functions and residuals via epi- and hypographs. *The Annals of Statistics*, 42(4):1598 – 1634, 2014.

[3] D. Cooley, P. Naveau, and P. Poncet. *Variograms for spatial max-stable random fields*, pages 373–390. Springer New York, New York, NY, 2006.

[4] J.-D. Fermanian, D. Radulović, and M. Wegkamp. Weak convergence of empirical copula processes. *Bernoulli*, 10(5):847–860, 2004.

[5] G. Marcon, S. Padoan, P. Naveau, P. Muliere, and J. Segers. Multivariate nonparametric estimation of the pickands dependence function using bernstein polynomials. *Journal of Statistical Planning and Inference*, 183:1–17, 2017.

[6] P. Naveau, A. Guillou, D. Cooley, and J. Diebolt. Modelling pairwise dependence of maxima in space. *Biometrika*, 96(1):1–17, 2009.

[7] J. Segers. Asymptotics of empirical copula processes under non-restrictive smoothness assumptions. *Bernoulli*, 18(3):764–782, 2012.

[8] J. Segers. Hybrid copula estimators. *J. Statist. Plann. Inference*, 160:23–34, 2015.

Méthode de plug-in pour l'estimateur non-paramétrique du coefficient de dépendance de queues

Matthieu Garcin [1] & Maxime Nicolas [2]

[1] *Léonard de Vinci Pôle Universitaire, Research center, 92916 Paris La Défense, France, matthieu.garcin[at]m4x.org*

[2] *Université Paris I Panthéon-Sorbonne, Maison des Sciences Economiques, 106-112, boulevard de l'Hôpital 75013 Paris, France, maxime.nicolas[at]univ-paris1.fr*

Résumé. Une expression théorique est donnée pour l'erreur quadratique moyenne asymptotique d'un estimateur non-paramétrique du coefficient de dépendance de queues, en fonction d'un seuil qui définit quel rang délimite les queues d'une distribution. Nous proposons une nouvelle méthode pour sélectionner de manière optimale ce seuil. Il combine l'erreur quadratique moyenne théorique de l'estimateur avec une estimation paramétrique de la copule reliant les observations dans les queues. À l'aide de simulations, nous comparons cette méthode semi-paramétrique avec d'autres approches proposées dans la littérature, y compris l'algorithme de recherche de plateau.

Mots-clés. Coefficient de dépendance de queues, Estimation non-paramétrique, Copule, Valeurs extrêmes

Abstract. A theoretical expression is derived for the mean squared error of a non-parametric estimator of the tail dependence coefficient, depending on a threshold that defines which rank delimits the tails of a distribution. We propose a new method to optimally select this threshold. It combines the theoretical mean squared error of the estimator with a parametric estimation of the copula linking observations in the tails. Using simulations, we compare this semiparametric method with other approaches proposed in the literature, including the plateau-finding algorithm.

Keywords. Tail dependence coefficient, Nonparametric estimation, Copula, Extreme values

1 Introduction

La description de la dépendance entre des variables aléatoires est un élément fondamental en gestion des risques, notamment en finance pour représenter les possibilités de diversification d'un portefeuille, c'est-à-dire de diminution des risques par l'ajout de titres choisis en fonction de la dépendance de leur rendement futur avec celui des titres déjà

en portefeuille. Si les modèles à dépendance linéaire offrent une solution simple, les copules permettent de décrire de manière plus fidèle cette dépendance, notamment grâce au théorème de Sklar (1959), qui stipule que toute paire (X, Y) de variables aléatoires a une fonction de répartition qui s'écrit

$$F(x, y) = C(F_X(x), F_Y(y)),$$

où C est la fonction copule et F_X et F_Y sont les marginales. Pour tout $(u, v) \in [0, 1]^2$, la copule s'exprime alors

$$C(u, v) = F(F_X^{-1}(u), F_Y^{-1}(v)).$$

La prise en compte des risques extrêmes impose d'analyser le comportement de cette copule pour des u et des v extrêmes. En particulier, nous nous intéressons aux coefficients de dépendance de queues (CDQ) introduits par Sibuya (1960) : le CDQ supérieur

$$\lambda_U = \lim_{u \to 1} \mathbb{P}\left[X > F_X^{-1}(u) | Y > F_Y^{-1}(u)\right] = \lim_{u \to 1} \frac{1 - 2u + C(u, u)}{1 - u}$$

et le CDQ inférieur

$$\lambda_L = \lim_{u \to 0^+} \mathbb{P}\left[X < F_X^{-1}(u) | Y < F_Y^{-1}(u)\right] = \lim_{u \to 0^+} \frac{C(u, u)}{u}.$$

Les CDQ dépendent uniquement de la copule, pas des marginales.

L'estimation des CDQ est délicate dans la mesure où ils quantifient des phénomènes extrêmes pour lesquels, par définition, peu d'observations sont disponibles. Dans ce qui suit, nous présenterons brièvement les différentes méthodes d'estimation existantes et nous en proposerons une nouvelle.

2 Estimateurs du CDQ

L'approche la plus simple pour estimer le CDQ est l'approche paramétrique. Après avoir sélectionné une copule paramétrique, on estime ses paramètres sur les données avec une méthode de maximum de vraisemblance. Un lien simple permet alors de passer de ces paramètres aux CDQ. Par exemple, pour une copule de Gumbel de paramètre θ, le CDQ supérieur est $\lambda_U = 2 - 2^\theta$.

Cependant, l'approche paramétrique ne s'appuie pas spécifiquement sur les événements extrêmes dans la mesure où l'estimation de la copule repose essentiellement sur le corps de la distribution. Dans cette perspective, on peut alors s'intéresser à l'estimateur non-paramétrique qui s'appuie sur des données extrêmes. Ainsi définit-on

$$\widehat{\lambda}_L\left(\frac{i}{n}\right) = \frac{\widehat{C}\left(\frac{i}{n}, \frac{i}{n}\right)}{\frac{i}{n}},$$

où $(u,v) \in [0,1]^2 \mapsto \widehat{C}(u,v)$ est la copule empirique (Deheuvels 1979). De même

$$\widehat{\lambda}_{\mathrm{U}}\left(\frac{i}{n}\right) = \frac{1 - 2\frac{i}{n} + \widehat{C}\left(\frac{i}{n}, \frac{i}{n}\right)}{1 - \frac{i}{n}}.$$

Finalement, tout l'enjeu de cet estimateur non-paramétrique réside dans la sélection du paramètre libre $i \in [\![1, n]\!]$, appelé seuil, le plus pertinent.

L'algorithme de recherche de plateau, proposé par Frahm, Junker et Schmidt (2005), tente de répondre à cette question par une approche heuristique. Cette méthode contient plusieurs régularisations, afin de réduire la variance de l'estimateur. En effet, pour un seuil i proche de 1 ou de n, l'estimateur non-paramétrique de la copule et des CDQ est peu robuste et a donc une variance très élevée, alors que quand i s'éloigne de ces valeurs extrêmes l'estimateur est plus exposé à la possibilité d'un biais conséquent.

3 Raffinement de l'estimateur non-paramétrique

Le sélection du seuil i dans l'estimateur non-paramétrique des CDQ repose, comme on l'a vu, sur un arbitrage classique entre biais et variance. Or, à notre connaissance, personne ne s'est encore intéressé à leur expression. C'est ce que nous faisons maintenant, avant d'en déduire une règle de sélection pour i.

3.1 Biais et variance de l'estimateur non-paramétrique

Avant de fournir une expression de l'erreur quadratique moyenne (et donc du biais et de la variance) de l'estimateur non-paramétrique de CDQ, il est utile d'introduire quelques notations :

- δ est la section diagonale : $\delta(u) = C(u,u)$. Nous avons $\lambda_L = \delta'(0)$ et $\lambda_U = 2 - \delta'(1)$.
- Les "fonctions h" de la copule sont : $h_1(u,v) = \partial C(u,v)/\partial u$ et $h_2(u,v) = \partial C(u,v)/\partial v$. Dans le cas de copules symétriques, nous écrivons $h(u,v) = h_1(u,v) = h_2(u,v)$. Quand $u = v$, nous simplifions la notation en écrivant $h_1(u)$, $h_2(u)$, et $h(u)$.

Grâce à l'expression de la distribution asymptotique de la copule empirique (Fermanian, Radulovic et Wegkamp, 2004), nous pouvons énoncer le théorème suivant.

Theorem 1 *Soit C une copule bivariée à dérivées partielles continues et $i \in [\![1, n]\!]$. Alors, l'erreur quadratique moyenne de l'estimateur non-paramétrique du CDQ inférieur se comporte asymptotiquement de la manière suivante:*

$$\mathbb{E}\left[\left(\widehat{\lambda}_L\left(\frac{i}{n}\right) - \lambda_L\right)^2\right] \overset{n\to\infty}{\sim} \frac{n}{i^2}\sigma^2\left(\frac{i}{n}\right) + \left(\frac{n}{i}\delta\left(\frac{i}{n}\right) - \delta'(0)\right)^2,$$

où

$$\sigma^2(u) = \delta(u)(1-\delta(u))+(1-u)\left[u\left(h_1(u)^2+h_2(u)^2\right)-2\delta(u)(h_1(u)+h_2(u))\right]+2h_1(u)h_2(u)(\delta(u)-u^2).$$

Un résultat similaire est obtenu pour le CDQ supérieur ainsi que pour les estimateurs définis comme une moyenne d'estimateurs non-paramétriques de seuils différents (Garcin et Nicolas, 2021). Par ailleurs, quand la copule C est symétrique, nous avons une expression plus simple pour $\sigma^2(u)$:

$$\sigma^2(u) = \delta(u)(1-\delta(u)) + 2(1-u)h(u)\left[uh(u)-2\delta(u)\right] + 2h(u)^2(\delta(u)-u^2).$$

Dans le cas particulier d'une copule de Clayton, nous avons

$$C(u,v) = \left(u^{-\theta}+v^{-\theta}-1\right)^{-1/\theta},$$

où $\theta > 0$, et l'on peut quantifier le biais et la variance du CDQ inférieur :

$$\left\{\begin{array}{rcl} variance & = & \frac{1}{nu^2}\left(\delta(u)-\delta(u)^2\left[1+2\frac{2(1-u)(1-u^\theta)+1}{u(2-u^\theta)^2}\right]+2\delta(u)^3\left[\frac{1}{u^2(2-u^\theta)^2}\right]\right) \\ biais^2 & = & \left(\left(2-u^\theta\right)^{-1/\theta}-2^{-1/\theta}\right)^2, \end{array}\right.$$

pour $u = i/n$ et avec $\delta(u) = \left(2u^{-\theta}-1\right)^{-1/\theta}$.

3.2 Sélection du paramètre libre

La méthode que nous proposons est une méthode semi-paramétrique : pour une spécification de copule donnée, de paramètre θ, nous avons une expression pour le biais et la variance de l'estimateur non-paramétrique de CDQ pour un seuil i. Nous proposons une méthode avec plug-in, dans laquelle le paramètre θ est estimé sur des données extrêmes, dans un ensemble d'observations dépendant du seuil i. Plus précisément, nous définissons cet ensemble $\Omega^L_{i/n}$ comme les observations au-delà du quantile multivarié de Kendall de probabilité i/n (Garcin, Guégan et Hassani, 2021) :

$$\Omega^L_{i/n} = \left\{(X_j,Y_j)\in\mathbb{R}^2, j\in[\![1,n]\!]\,\middle|\,\widehat{F}\left(X_j,Y_j\right)) \leq \widehat{K}^{-1}\left(\frac{i}{n}\right)\right\},$$

où $\widehat{K}$ est la version empirique de la fonction de Kendall, $K(p) = \mathbb{P}[F(X,Y)\leq p]$, et qui s'appuie sur la fonction de répartition empirique $\widehat{F}$:

$$\widehat{K}(p) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\widehat{F}(X_i,Y_i)\leq p}.$$

On peut alors estimer θ en ne considérant que les données extrêmes, c'est-à-dire appartenant à $\Omega^L_{i/n}$, par un estimateur du maximum de vraisemblance avec censure, le terme

de censure s'écrivant de manière simple grâce au choix du quantile multivarié de Kendall que nous avons fait dans la définition de $\Omega_{i/n}^L$:

$$\widehat{\theta}_{L,i/n} = \underset{\theta}{\operatorname{argmax}} \sum_{j=1}^{n} \left\{ \ln\left(c_\theta(\widehat{F}_X(X_j), \widehat{F}_Y(Y_j))\right) \mathcal{J}_{j,i}^L + \ln\left(1 - K_\theta\left(\widehat{K}^{-1}\left(\frac{i}{n}\right)\right)\right)\left(1 - \mathcal{J}_{j,i}^L\right) \right\},$$

où c_θ est la densité d'une copule paramétrique C_θ de paramètre θ et $\mathcal{J}_{j,i}^L = \mathbb{1}_{(X_j,Y_j)\in\Omega_{i/n}^L}$.

Ainsi peut-on écrire l'estimateur $\widehat{\theta}_{L,i/n}$ comme une fonction du seuil i : $\widehat{\theta}_{L,i/n} = \psi(i/n)$. Nous appliquons maintenant une méthode du plug-in. En supposant une spécification donnée (Clayton par exemple) de paramètre θ pour la copule, du moins pour les observations extrêmes, on peut utiliser le Théorème 1 pour exprimer l'erreur quadratique moyenne asymptotique du CDQ pour un seuil i : $\widetilde{MSE}(i, C_\theta)$. En remplaçant θ par son estimateur de seuil i, $\psi(i/n)$, nous aboutissons donc à une expression de l'erreur que l'on cherche à minimiser :

$$i_{PI}^\star = \underset{i\in[\![1,n]\!]}{\operatorname{argmin}}\ \widetilde{MSE}(i, C_{\psi(i/n)}).$$

4 Étude de simulations et conclusion

La Figure 1 illustre la méthode de plug-in décrite ci-avant.

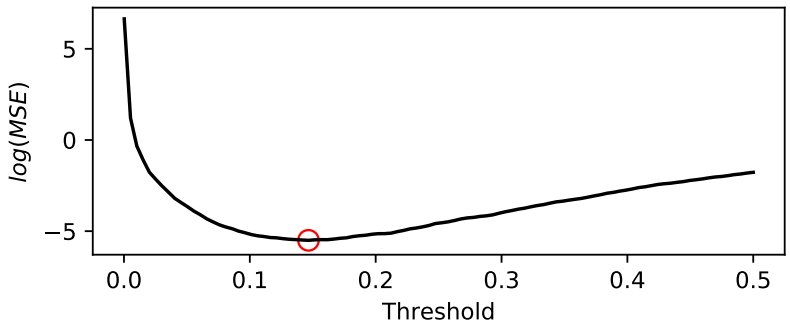


Figure 1: Erreur quadratique moyenne estimée par le Théorème 1 en faisant l'hypothèse d'une spécification Clayton pour 5000 données simulées selon une rotation d'une copule de Gumbel de paramètre 2.

Des résultats de simulations entre les méthodes d'estimation introduites dans ce document, incluant en particulier la méthode de plug-in et la recherche de plateau, montrent un bon comportement de la méthode de plug-in (Garcin et Nicolas, 2021). Cela ouvre la voie à une description plus précise de la dépendance de queue entre des variables aléatoires. Cette précision est requise pour certaines applications des CDQ, comme une allocation de portefeuille s'appuyant sur un partitionnement des titres selon leur CDQ, à des fins de diversification et d'optimisation de portefeuille (De Luca et Zuccolotto, 2011 et 2017).

Bibliographie

De Luca, G. et Zuccolotto, P. (2011), A tail dependence-based dissimilarity measure for financial time series clustering, *Advances in data analysis and classification*, 5(4), 323-340.

De Luca, G. et Zuccolotto, P. (2017), Dynamic tail dependence clustering of financial time series, *Statistical papers*, 58(3), 641-657.

Deheuvels, P. (1979), La fonction de dépendance empirique et ses propriétés. Un test non paramétrique d'indépendance, *Bulletins de l'académie royale de Belgique*, 65(1), 274-292.

Fermanian, J.-D., Radulovic, D. et Wegkamp, M. (2004), Weak convergence of empirical copula processes, *Bernoulli*, 10(5), 847-860.

Frahm, G., Junker, M. et Schmidt, R. (2005), Estimating the tail-dependence coefficient: properties and pitfalls, *Insurance: mathematics and economics*, 37(1), 80-100.

Garcin, M., Guégan, D. et Hassani, B. (2021), A multivariate quantile based on Kendall ordering, *Revstat-statistical journal*, à paraître.

Garcin, M. et Nicolas, M.L. (2021), Nonparametric estimator of the tail dependence coefficient: balancing bias and variance, *arXiv preprint* arXiv:2111.11128.

Sibuya, M. (1960), Bivariate extreme statistics, *Annals of the institute of statistical mathematics*, 11(2), 195-210.

Sklar, M. (1959), Fonctions de répartition à n dimensions et leurs marges, *Publ. inst. statist. univ. Paris*, 8, 229-231.

Estimation fonctionnelle de L^1-expectiles multivariés extrêmes à queue lourde

Cambyse Pakzad [1] & Elena Di Bernardino [2] & Thomas Laloë [3]

[1,2,3] *Laboratoire J.A. Dieudonné UMR CNRS 7351, Université Côte d'Azur, Parc Valrose 06108 Nice Cedex 2, France. cpakzad@unice.fr ; elenadb@unice.fr ; thomas.laloe@unice.fr*

Résumé. Les expectiles et les quantiles peuvent être interprétés comme des solutions de problèmes de minimisation convexes mais seuls les expectiles constituent une mesure de risque, invariante par loi, cohérente au sens d'Artzner et al (1998) et élicitable au sens de Gneiting (2011). Nous étudions ici une généralisation au cas multivarié, à savoir le L^1-expectile. Nous adressons la question de son estimation dans le cadre extrême, avec des distributions sous-jacentes à queue lourde et lorsqu'une covariable fonctionnelle est disponible. Pour cela, nous avons recours à une approche de minimisation qui nécessite l'estimation de différents objets tels que l'indice de queue conditionnel ou les fonctions de dépendance de queue bivariées. Nous abordons la question de la vitesse de convergence de notre méthode d'estimation. Pour ce faire, nous explorons en particulier la théorie de la variation régulière.

Mots-clés. Mesure de risque multivariée, Théorie des extrêmes, Dépendance, Variation régulière, Queue lourde, Statistique fonctionnelle

Abstract. Expectiles and quantiles can be interpreted as solutions of convex minimization problems, but only expectiles constitute a risk measure, law-invariant, consistent in the sense of Artzner et al (1998) and elicitable in the sense of Gneiting (2011). We study here a generalization to the multivariate case, namely the L^1-expectile. We address the question of its estimation in the extreme setting, with heavy-tailed underlying distributions and when a functional covariate is available. For this, we use a minimization approach that requires the estimation of different objects such as the conditional tail index or the bivariate tail dependence functions. We investigate the issue of the speed of convergence of our estimation method. To do so, we explore in particular the theory of regular variation.

Keywords. Multivariate risk measure, Extreme theory, Dependence, Regular variation, Heavy tail, Functional statisticss

1 Génèse

Afin d'établir une stratégie contre l'incertitude, les mesures de risques sont devenues un outil incontournable dans une multitude de domaines nécessitant une gestion du risque,

comme l'actuariat, la finance ou les sciences de l'environnement. En opposition au quantile, Newey et Powell (1987) proposent une mesure de risque alternative, l'expectile, qui utilise l'information sur la fréquence et magnitude des événements extrêmes. Mais le caractère univarié se trouve être très restrictif pour les applications. Récemment, plusieurs généralisations multivariées de l'expectile sont introduites par Maume-Deschamps, Rulliere et Said (2018), notamment le dénommé L^1-expectile. Nous en donnons la définition directement dans sa forme conditionnelle.

Définition 1.1. Soit y un élément d'un espace Polonais (E, dist). Pour un vecteur aléatoire $\boldsymbol{X} = (X_1, \ldots, X_d)$, on définit son L^1-expectile conditionnel (par rapport à Y variable aléatoire dans E) de niveau $\alpha \in (0,1)$ comme le $\mathbb{R}^d$-vecteur aléatoire :

$$\boldsymbol{e}_\alpha(\boldsymbol{X}, y) := \underset{\boldsymbol{x}\in\mathbb{R}^d}{\arg\min}\, \mathbb{E}\left[\alpha \left\|(\boldsymbol{X}-\boldsymbol{x})_+\right\|_1^2 + (1-\alpha)\left\|(\boldsymbol{X}-\boldsymbol{x})_-\right\|_1^2 | Y = y\right]. \tag{1}$$

La stricte convexité de $\varphi_\alpha(x,t) = \alpha \left\|(t-x)_+\right\|^2 + (1-\alpha)\left\|(t-x)_-\right\|^2$ en x assure l'existence et unicité de (1). La version univariée d'origine se retrouve en prenant $d = 1$ et en supprimant la condition $Y = y$.

Traditionnellement, l'étude des événements extrêmes se fait à travers l'estimation du quantile $q_X(\alpha) := F_X^{-1}(\alpha) = \inf\{t \in \mathbb{R} : F_X(t) \geq \alpha\}$, où F est la fonction de répartition de X, dans le régime extrême $\alpha := \alpha_n \to 1$ avec n la taille de l'échantillon et X à queue lourde. L'extension à l'expectile multivarié ou non, dans un régime extrême, avec covariable fonctionnelle n'est que très récente. Dans le cadre univarié, Girard, Stupfler et Usseglio-Carleve (2021) achèvent l'estimation de l'expectile conditionnel extrême à queue lourde, avec covariable fonctionnelle, grâce à une connexion astucieuse entre le quantile et l'expectile univarié. Du côté multivarié, Beck, Di Bernardino et Mailhot (2021) aboutissent à l'estimation du L^1-expectile (non-conditionnel) dans le régime extrême à queue lourde par une approche de minimisation.

Le texte ci-présent est dédié à l'estimation du L^1-expectile conditionnel à covariable fonctionnelle dans le régime extrême à queue lourde, i.e. $\alpha = \alpha_n \to 1$ avec n la taille de l'échantillon et les marginales conditionnelles de $\boldsymbol{X}$ sont à queues lourdes.

2 Variation régulière

Le cadre mathématique pour aborder la queue de distribution est celui de la variation régulière et son poids se quantifie via l'indice de queue noté $\theta = \gamma^{-1}$.

Définition 2.1. Une fonction positive mesurable f est à variation régulière d'indice θ, i.e. $f \in \mathrm{RV}_\theta$, si pour tout $t > 0$, $\frac{f(tx)}{f(t)} \xrightarrow[t\to+\infty]{} x^\theta$. Lorsque $\theta = 0$, on dit que f est à variation lente. Par extension, une variable aléatoire réelle X est à variation régulière, i.e. $X \in \mathrm{RV}_\theta$, si sa fonction de survie $\overline{F}_X = 1 - F_X \in \mathrm{RV}_\theta$.

On a l'équivalence entre $f \in \mathrm{RV}_\theta$ et le fait qu'il existe une fonction ℓ_f à variation lente, dite partie lente de f, telle que $f(x) = x^\theta \ell_f(x)$ pour tout x. Dans sa forme multivariée,

Définition 2.2. La loi d'un vecteur aléatoire $\boldsymbol{X} \in \mathbb{R}^d$ est à variation régulière d'indice $\theta \in (0, +\infty)$ à l'infini, i.e. $\boldsymbol{X} \in \mathrm{MRV}_\theta$, si pour une certaine norme $\|\cdot\|$ sur $\mathbb{R}^d$, il existe un vecteur aléatoire Θ de la sphère unité $\mathbb{S}^{d-1}$ tel que pour tout $x \in (0, +\infty)$,$\mathbb{P}(\|\boldsymbol{X}\| > z)^{-1}\mathbb{P}(\|\boldsymbol{X}\| > xz, \frac{\boldsymbol{X}}{\|\boldsymbol{X}\|} \in \cdot) \xrightarrow[z\to+\infty]{} x^{-\theta}\mathbb{P}(\Theta \in \cdot)$.

En outre, on peut modéliser les fluctuations dans la convergence de Définition 2.1 :

Définition 2.3 (Second ordre)**.** Soit $\gamma > 0$ et $\rho \leq 0$. On dit que $\overline{F} \in 2\mathrm{RV}_{-\frac{1}{\gamma},\frac{\rho}{\gamma}}$ s'il existe une fonction auxiliaire A de signe ultimement constant avec $\lim_{t\to+\infty} A(t) = 0$ telle que pour tout $x > 0$, $\lim_{t\to+\infty} \frac{1}{A\left(1/\overline{F}(t)\right)}\left(\frac{\overline{F}(tx)}{\overline{F}(t)} - x^{-\frac{1}{\gamma}}\right) = x^{-\frac{1}{\gamma}}\left(\frac{x^{\frac{\rho}{\gamma}}-1}{\gamma\rho}\mathbf{1}_{\{\rho<0\}} + \log(x)\mathbf{1}_{\{\rho=0\}}\right)$.

Il est parfois plus commode de prendre le point de vue des fonctions quantiles de queue $U_X(x) := \overline{F}_X^{-1}(x) = q_X(1 - \frac{1}{x})$. Il y a alors équivalence avec $\lim_{t\to+\infty} \frac{1}{A(t)}\left(\frac{U(tx)}{U(t)} - x^\gamma\right) = x^\gamma \int_1^x u^{\rho-1}\mathrm{d}u$, $x > 0$. En particulier, $\overline{F} \in \mathrm{RV}_{-1/\gamma} \iff U \in \mathrm{RV}_\gamma$ et $U \in 2\mathrm{RV}_{\gamma,\rho} \iff \overline{F} \in 2\mathrm{RV}_{-\frac{1}{\gamma},\frac{\rho}{\gamma}}$.

3 Le modèle

Soit un espace Polonais (E, dist) et Y une variable aléatoire dans E ayant le rôle de covariable. Soit un vecteur aléatoire $\boldsymbol{X} = (X_1, \ldots, X_d)$. On fait l'hypothèse de variation régulière multivariée d'indice θ pour le vecteur conditionnelle $\boldsymbol{X}|Y \in \mathbb{R}^d$ avec les queues de ses marginales à variation régulière équivalente.

Hypothèses 3.1. *Pour $1 \leq j \leq d$ et $y \in E$, supposons $\boldsymbol{X}|Y \in \mathrm{MRV}_\theta(+\infty)$ et,*

- $x \mapsto F_{X_j|Y}(x|y) = \mathbb{P}(X_j \leq x|Y = y)$ *est continue.*
- $x \mapsto \overline{F}_{X_1|Y}(\cdot|y) = \mathbb{P}(X_1 > x|Y = y) \in \mathrm{RV}_{-\theta(y)}$ *avec* $\theta(y) = \frac{1}{\gamma(y)} > 0$.
- *Il existe* $c_j(y) \in (0, +\infty)$ *tel que* $\frac{\overline{F}_{X_j|Y}(x|y)}{\overline{F}_{X_1|Y}(x|y)} \xrightarrow[x\to+\infty]{} c_j(y)$.

La 3ème hypothèse est précisément celle de queue lourde, propre au régime extrême, dont l'indice $\gamma(y)$ est uniforme en les marginales. La dernière est fondamentale dans notre approche.

Par continuité, le théorème de Sklar certifie l'existence du copule conditionnel bivarié

$$C^{j,k}_{|Y}(u_j, u_k) = F_{X_j,X_k|Y}(F^{-1}_{X_j|Y}(u_j|y), F^{-1}_{X_k|Y}(u_k|y)\big|y),\ 1 \leq j \neq k \leq d \text{ et } u_j, u_k \in [0,1]. \quad (2)$$

On modélise alors la dépendance deux-à-deux par la fonction de dépendance conditionnelle de queue (supérieure) par

$$\lambda_{|Y}^{j,k}(u_j,u_k|y)=\lim_{t\to 0}t^{-1}\mathbb{P}(\overline{F}_{X_j|Y}(X_j|y)\leq tu_j,\overline{F}_{X_k|Y}(X_k|y)\leq tu_k|Y=y). \tag{3}$$

Son existence est équivalente au caractère MRV de la distribution sous-jacente.

4 Stratégie d'estimation du L^1-expectile

On considère n copies indépendantes $(\boldsymbol{X}_i,Y_i)_{1\leq i\leq n}$ avec $\boldsymbol{X}_i=(X_{ij})_{1\leq j\leq d}$, $i\in\{1,\dots,n\}$, de la paire aléatoire $(\boldsymbol{X},Y)\in\mathbb{R}^d\times E$ où (E,dist) est Polonais, sous les Hypothèses 3.1. Imitant la démarche de Beck, Di Bernardino et Mailhot (2021), l'idée est d'exprimer le problème de minimisation d'origine (1) sous une nouvelle forme.

4.1 Optimalité

Adaptant au cas conditionnel le résultat d'optimalité de Maume-Deschamps, Rullière et Said (2018),

Proposition 4.1. *Sous les Hypothèses 3.1, les composantes du L^1-expectiles $\boldsymbol{e}_\alpha^j(\boldsymbol{X},y)$ pour $j=2,\dots,d$ vérifient $0<\liminf_{\alpha\to 1}\frac{1-\alpha}{\overline{F}_{X_j|Y}(\boldsymbol{e}_\alpha^j(\boldsymbol{X},y)|y)}\leq\limsup_{\alpha\to 1}\frac{1-\alpha}{\overline{F}_{X_j|Y}(\boldsymbol{e}_\alpha^j(\boldsymbol{X},y)|y)}<+\infty$ et $0<\liminf_{\alpha\to 1}\frac{\boldsymbol{e}_\alpha^j(\boldsymbol{X},y)}{\boldsymbol{e}_\alpha^1(\boldsymbol{X},y)}\leq\limsup_{\alpha\to 1}\frac{\boldsymbol{e}_\alpha^j(\boldsymbol{X},y)}{\boldsymbol{e}_\alpha^1(\boldsymbol{X},y)}<+\infty$. De plus, tout vecteur limite $(\eta,\beta_2,\dots,\beta_d)$ de $(\frac{1-\alpha}{\overline{F}_{X_1}(\boldsymbol{e}_\alpha^1(\boldsymbol{X},y)|y)},\frac{\boldsymbol{e}_\alpha^2(\boldsymbol{X},y)}{\boldsymbol{e}_\alpha^1(\boldsymbol{X},y)},\dots,\frac{\boldsymbol{e}_\alpha^d(\boldsymbol{X},y)}{\boldsymbol{e}_\alpha^1(\boldsymbol{X},y)})$, $y\in E$ vérifie le système, pour $1\leq k\leq d$,*

$$\frac{1}{\theta(y)-1}-\eta\frac{\beta_k^{\theta(y)}}{c_k(y)}=-\sum_{j\neq k}^{d}\left\{\int_{\frac{\beta_j}{\beta_k}}^{+\infty}\lambda_{|Y}^{j,k}\left(\frac{c_i(y)}{c_k(y)}t^{-\theta(y)},1|y\right)\mathrm{d}t-\eta\frac{\beta_k^{\theta(y)-1}}{c_k(y)}\beta_i\right\}. \tag{4}$$

Ce résultat nous permet d'accéder à l'expectile via les ratios de ses marginales. Notons bien que ceci n'est possible que sous l'hypothèse de queues équivalentes. L'équation (4) fait intervenir $\theta=\gamma^{-1}$ et $\boldsymbol{c}(y)=(c_2(y),\dots,c_d(y))$ de Hypothèses 3.1 ainsi que $\lambda_{|Y}^{j,k}$ de (3). On comprend donc que leur estimation individuelle est une tâche impérieuse. Précisons maintenant la reformulation du problème de minimisation :

Définition 4.1. Sous les Hypothèses 3.1 et en omettant la notation de dépendance en $y\in E$, soit $\boldsymbol{\beta}^\star=(\beta_j)_{1\leq j\leq d}$ avec $\beta_1=1$ et $\boldsymbol{\Theta}=(\eta,\beta_2,\dots,\beta_d)$. Soit $c_1=1$ et $\boldsymbol{\Lambda}=(\theta,c_2,\dots,c_d,\lambda_{|Y}(\cdot))$. On définit la fonction de perte $L_{\boldsymbol{\Lambda}}(\boldsymbol{\Theta}):=\frac{1}{2}\left\|F_{\boldsymbol{\Lambda}}(\boldsymbol{\Theta})\right\|_2^2$ et un vecteur optimal $\boldsymbol{\Theta}^\star:=\arg\min_{\boldsymbol{\Theta}}L_{\boldsymbol{\Lambda}}(\boldsymbol{\Theta})$ où $F_{\boldsymbol{\Lambda}}(\boldsymbol{\Theta})=\left(F_{\boldsymbol{\Lambda}}^{(k)}(\boldsymbol{\Theta})\right)_{1\leq k\leq d}$ et, pour $k=1,\dots,d$,

$$F_{\boldsymbol{\Lambda}}^{(k)}(\boldsymbol{\Theta}):=\frac{\gamma(y)}{\gamma(y)-1}-\eta\frac{\beta_k^{1/\gamma(y)}}{c_k(y)}\left(1+\sum_{j\neq k}\frac{\beta_j}{\beta_k}\right)+\sum_{j\neq k}^{d}\int_{\frac{\beta_j}{\beta_k}}^{+\infty}\lambda_{|Y}^{j,k}\left(\frac{c_j(y)}{c_k(y)}t^{-1/\gamma(y)},1|y\right)\mathrm{d}t.$$

Les problèmes de minimisation (1) et (4) sont équivalents :

Proposition 4.2. *Sous les Hypothèses 3.1 et l'existence d'un unique vecteur limite* $\boldsymbol{\Theta} = (\eta, \beta_2, \ldots, \beta_d)$ *dans Proposition 4.1 lorsque* $\alpha \to 1$*, le vecteur limite* $\boldsymbol{\Theta}$ *vérifie la minimisation dans Définition 4.1, i.e.* $\boldsymbol{\Theta}$ *est un candidat pour* $\boldsymbol{\Theta}^\star$*.*

De fait, les éléments de $\boldsymbol{\Lambda}$ sont inconnus, donc une application directe d'un algorithme d'optimisation pour $\boldsymbol{\Theta}^\star$ n'est pas envisageable. On considère plutôt une version approximée $\arg\min_{\boldsymbol{\Theta}\in\mathbb{R}^d} L_{\hat{\boldsymbol{\Lambda}}}(\boldsymbol{\Theta})$ pour des vecteurs d'estimation $\hat{\boldsymbol{\Lambda}} = (\hat{\theta}, \hat{c}_2, \ldots, \hat{c}_d, \hat{\lambda}_{|Y})$ à expliciter. Dans cet esprit, posons :

$$F^{(k)}_{\hat{\boldsymbol{\Lambda}}}(\boldsymbol{\Theta}) := \frac{\hat{\gamma}_n(y)}{\hat{\gamma}_n(y)-1} - \eta\frac{\beta_k^{1/\hat{\gamma}_n(y)}}{\hat{c}_{k,n}(y)}\left(1+\sum_{j\neq k}\frac{\beta_j}{\beta_k}\right) + \sum_{j\neq k}\int_{\frac{\beta_j}{\beta_k}}^{+\infty} \widehat{\lambda}^{j,k}_{|Y,n}\left(\frac{\hat{c}_{j,n}(y)}{\hat{c}_{k,n}(y)}t^{-1/\hat{\gamma}_n(y)}, 1|y\right)\mathrm{d}t.$$

5 Résultats

Pour $y \in E$ et $0 < h_n \ll 1$, on définit la probabilité de petite boule $\varphi_y(h_n) := \mathbb{P}(Y \in B_{\mathrm{dist}}(y, h_n))$. On suppose dans la suite que $y \in E$ est tel que $\varphi_y(h_n) > 0$ et $n\varphi_y(h_n) \gg 1$. Notons $\overline{\alpha_n} := 1 - \alpha_n \to 0$. Notre résultat principal s'énonce comme une vitesse de convergence entre les versions théoriques et approximées.

Théorème 5.1. *Sous les Hypothèses 3.1 et certaines hypothèses de régularité, supposons en outre que pour tout* $y \in E$ *et* $1 \leq j \leq d$*, il existe* $\rho_j(y) \leq 0$ *et une fonction auxiliaire* $A_j(\cdot|y) \in \mathrm{RV}_{\rho_j(y)}$ *tels que* $U_{X_j|Y}(\cdot|y) := \overline{F}^{-1}_{X_j|Y}(\cdot|y) \in 2\mathrm{RV}_{-\frac{1}{\gamma(y)},\rho_j(y)}$*. Alors,* $\left|F^{(k)}_{\boldsymbol{\Lambda}}(\boldsymbol{\Theta}) - F^{(k)}_{\hat{\boldsymbol{\Lambda}}}(\boldsymbol{\Theta})\right| = O\left\{\delta_n(y)\right\}$*, où* $\delta_n(y)$ *tend vers* 0 *avec vitesse au plus* $O_{\mathbb{P}}\{A_1(\overline{\alpha}_n^{-1}|y) \vee A_k(\overline{\alpha}_n^{-1}|y) \vee (n\overline{\alpha}_n\varphi_y(h_n))^{-\frac{1}{2}}\} \ll 1$.

La preuve consiste en une analyse terme à terme de la différence $\Delta := F^{(k)}_{\boldsymbol{\Lambda}}(\boldsymbol{\Theta}) - F^{(k)}_{\hat{\boldsymbol{\Lambda}}}(\boldsymbol{\Theta})$. Suivant les pas de Gardes et Girard (2012) et Girard, Stupfler et Usseglio-Carleve (2021), choisissons $1 \leq J < +\infty$ et une subdivision $0 < \tau_J < \tau_{J-1} < \cdots < \tau_2 < \tau_1 = 1$. Soit $1 \leq j \leq d$. On considère alors le quantile empirique conditionnel $\hat{q}_{X_j|Y,n}(\alpha_n|y) = \hat{F}^{-1}_{X_j|Y,n}(\alpha_n|y)$, $\alpha_n \to 1$, où $\widehat{\overline{F}}_{X_j|Y,n}$ est un estimateur conditionnel de $F_{X_j|Y}$ et l'estimateur de Hill conditionnel $\hat{\gamma}^{\mathrm{H}}_n(y) = \sum_{j=1}^J \log\frac{\hat{q}_n(\tau_j\overline{\alpha}_n|y)}{\hat{q}_n(\overline{\alpha}_n|y)}(-\sum_{j=1}^J\log(\tau_j))^{-1}$. Sous des hypothèses de type Lipschitz, les mêmes auteurs prouvent les fluctuations gaussiennes du vecteur $\{(\frac{\hat{q}_{X_k|Y,n}(\tau_j\alpha_n|y)}{q_{X_k|Y}(\tau_j\alpha_n|y)}-1)\}_{j=1,\ldots,J}$ et de $(\hat{\gamma}^{\mathrm{H}}_n(y)-\gamma(y))$ à vitesse $\sqrt{n\overline{\alpha}_n\varphi_y(h_n)}$ que nous mettons à profit par Δ-méthode. Quant aux constantes $c_j(y)$, nous proposons $\hat{c}_{n,j}(y) = (\frac{\hat{q}_{X_j|Y,n}(\alpha_n|y)}{\hat{q}_{X_1|Y,n}(\alpha_n|y)})^{-1/\hat{\gamma}^{\mathrm{H}}_n(y)}$. Une analyse développée autour de la représentation de Karamata et de la variation régulière de second ordre met en évidence l'intervention des fonctions auxiliaries dans la vitesse $\delta_n(y)$. En passant, nous étendons un développement limité de Hua et Joe (2011) au cas $\rho = 0$ via le point de vue des Π-classes :

Lemme 5.2. *Supposons que pour tout* $y \in E$ *et* $1 \leq j \leq d$, $U_{X_j|Y}(\cdot|y) := \overline{F}^{-1}_{X_j|Y}(\cdot|y) \in 2\mathrm{RV}_{-\frac{1}{\gamma(y)},\rho_j(y)}$, *pour* $\gamma(y) > 0$ *et* $\rho_j(y) \leq 0$ *de fonction auxiliaire* $A_j(\cdot|y) \in \mathrm{RV}_{\rho_j(y)}$. *Alors, il existe* $\kappa_{U_{X_j|Y}}(y) \in (0, +\infty)$ *independant de* x *telle que la partie lente de* $U_{X_j|Y}$ *vérifie* $\ell_{U_{X_j|Y}}(x|y) = \kappa_{U_{X_j|Y}}(y) \cdot (1 + O\{A_j(x|y)\})$ *quand* $x \to +\infty$.

Concernant la partie intégrale de la différence Δ où apparaissent les fonctions bivariées de dépendance $\lambda^{j,k}_{|Y}$, on définit les poids de Nadaraya-Watson par $w_{i,n}(y, h_n) := \frac{K(\frac{\mathrm{dist}(Y_i - y)}{h_n})}{\sum_{s=1}^n K(\frac{\mathrm{dist}(Y_s - y)}{h_n})}$ où K est un noyau donné, afin de capturer la conditionnalité. On utilise la version empirique à poids $\hat{F}_{X_j,X_k|Y,n}(x_j, x_k) = \sum_{i=1}^n w_{i,n}(y, h_n)\mathbf{1}_{\{X_{ij} \leq x_j, X_{ik} \leq x_k\}}$ où $x_j, x_k \in \mathbb{R}$ pour estimer $F_{X_j,X_k|Y}$. On note $\hat{F}_{j,n,y}$, $\hat{F}_{k,n,y}$ et ses marginales $\hat{F}_{j,n,y}(x_j) = \hat{F}_{j,k,n,y}(x_j, +\infty), \hat{F}_{k,n,y}(x_k) = \hat{F}_{j,k,n,y}(+\infty, x_k)$. De même, on considère ensuite la version empirique à poids de (2) dont les fluctuations gaussiennes avec vitesse $\sqrt{n\varphi_y(h_n)}$ sont établies par Gijbels, Omelka et Veraverbeke (2012) :

$$\hat{C}^{j,k}_{|Y,n}(u_j, u_k) = \hat{F}_{j,k,n,y}(\hat{F}^{-1}_{X_j|Y,n}(u_j), \hat{F}^{-1}_{X_k|Y,n}(u_k)) = \sum_{i=1}^n w_{i,n}\mathbf{1}_{\{X_{ik} \leq \hat{F}^{-1}_{X_j|Y,n}(u_j), X_{ik} \leq \hat{F}^{-1}_{X_k|Y,n}(u_k)\}}.$$

Bibliographie

Artzner, P., Delbaen, F., Eber, J.-M. and Heath, D. (1999), Coherent Measures of Risk, *Mathematical Finance*, 9: 203-228, DOI:10.1111/1467-9965.00068

Beck, N., Di Bernardino, E., Mailhot, M. (2021), Semi-parametric estimation of multivariate extreme expectiles. *Journal of Multivariate Analysis*, Elsevier, 184, DOI:10.1016/j.jmva.2021.104758

Gardes, L. and Girard, S. (2012), Functional Kernel Estimators of Large Conditional Quantiles, *Electronic Journal of Statistics*, 6, 1715-1744, DOI:10.1214/12-EJS727

Gijbels, I., Omelka, M. and Veraverbeke, N. (2012), Multivariate and functional covariates and conditional copulas, *Electron. J. Statist.*, 6 1273 - 1306, DOI:10.1214/12-EJS712

Girard, S., Stupfler G. and Usseglio-Carleve, A. (2022), Functional estimation of extreme conditional expectiles, *Econometrics and Statistics*, Elsevier, 21, pp.131-158.

Gneiting, T. (2011), Making and Evaluating Point Forecasts, *Journal of the American Statistical Association*, 106:494, 746-762, DOI:10.1198/jasa.2011.r10138

Hua, L. and Joe, H. (2011), Second order regular variation and conditional tail expectation of multiple risks, *Insurance: Mathematics and Economics*, 49, issue 3, p. 537-546, https://EconPapers.repec.org/RePEc:eee:insuma:v:49:y:2011:i:3:p:537-546.

Maume-Deschamps, V., Rullière, D., Said, K. (2018) Extremes for multivariate expectiles. *Statistics & Risk Modeling with Applications in Finance and Insurance*, De Gruyter, 35 (3-4), pp.111-140. DOI:10.1515/strm-2017-0014

Newey, W. K., Powell, J. L. (1987). Asymmetric Least Squares Estimation and Testing. *Econometrica*, 55(4), 819–847. DOI:10.2307/1911031

Statistique mathématique

Full Model Estimation for Non-Parametric Multivariate Finite Mixture Models

Marie du Roy de Chaumaray [1] & Matthieu Marbac [1]

[1] *Univ. Rennes, Ensai, CNRS, CREST - UMR 9194, F-35000 Rennes, France*
{*marie.du-roy-de-chaumaray, matthieu.marbac-lourdelle*}*@ensai.fr*

Résumé. Cet article s'intéresse à l'estimation du modèle complet dans des modèles de mélanges non-paramétriques finis. Il présente une approche pour sélectionner le nombre de composantes ainsi que le sous-ensemble de variables discriminantes, c'est-à-dire le sous-ensemble de variables ayant des distributions différentes pour les composantes du mélange. L'approche proposée repose sur la discrétisation de chaque variable en B intervalles et sur une pénalisation de la log-vraisemblance qui en résulte. Pour un choix judicieux du terme de pénalité, on montre la consistance de l'estimateur du modèle (nombre de composantes et sous-ensemble de variables discriminantes) lorsque le nombre d'intervalles tend vers l'infini avec la taille de l'échantillon.

Mots-clés. Processus empirique, modèle à classes latentes, modèle localement conique, perte d'identifiabilité, sélection de modèle, mélange non-paramétrique.

Abstract. This paper addresses the problem of full model estimation for non-parametric finite mixture models. It presents an approach for selecting the number of components and the subset of discriminative variables, *i.e.,* the subset of variables having different distributions among the mixture components. The proposed approach considers a discretization of each variable into B bins and a penalization of the resulting log-likelihood. Considering that the number of bins tends to infinity with the sample size, we prove that our estimator of the model (number of components and subset of relevant variables for clustering) is consistent under a suitable choice of penalty term.

Keywords. Empirical process, Latent class model, Locally conic model, Loss of identifiability, Model selection, Non-parametric mixture.

1 Introduction

Finite mixture models permits to achieve clustering by estimating the distribution of the observed variables. This paper focuses on a full model selection (*i.e.,* estimation of the number of components and detection of the subset of the relevant variables for clustering) for non-parametric mixture models where no more assumption are made on the component distribution except to be defined as a product of univariate densities. Thus, we consider a sample of n independent observations $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ where $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{iJ})^\top \in \mathcal{X}$ is composed of the J variables collected on subject i defined on the space $\mathcal{X} = \mathcal{X}_1 \times \ldots \times \mathcal{X}_J$ where each $\mathcal{X}_j$ is compact. Each $\boldsymbol{X}_i$ is identically distributed according to the non-parametric mixture of K components defined by the density

$$g(\boldsymbol{x}_i) = \sum_{k=1}^{K} \pi_k \prod_{j=1}^{J} \eta_{kj}(x_{ij}), \tag{1}$$

where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)^\top$ is a finite dimensional parameter belonging to the simplex of size K, $\mathcal{S}_K = \{\boldsymbol{u} \in [0,1]^K : \sum_{k=1}^K u_k = 1\}$ and where the univariate densities η_{kj} constitute infinite dimensional parameters and are supposed to be striclty positive except on a set of lebesgue measure zero.

Before the work of [4], no tools were available for selecting the number of components K. This was the main drawback of non-parametric versus parametric mixture models which allow information criteria for model selection to be used. The two recent papers of [4] and [6] introduced two approaches for determining the smallest value of K such that (1) holds true. One elegant property of both methods is that they determine an estimator of K without performing the density estimation for different numbers of clusters and without choosing ahead a maximum number of clusters. Both start with a step of model selection followed by the estimation of the parameters for the selected model, which is quite unusual. However, their approaches are mainly based on the distribution of a pair of variables. Thus, if the number of variables J is large, computational issues can arise while considering all possible pairs of variables. It restricts the use of their methods to data sets composed of few variables. Additionnally, the method of [4] is only consistent to a lower bound of K. Moreover, the nature of the approach makes impossible a variable selection.

Selecting variables in clustering is challenging because the role of a variable is defined from the unobserved partition. Thus, variable selection and clustering need to be performed simultaneously. Note that selecting the variables in mixture models has two strong benefits: improving estimators accuracy by reducing their number and facilitating the interpretation of each component by considering only the subset of discriminative variables. We consider two types of variables: the relevant variables and the irrelevant variables for clustering. Variable j is irrelevant for clustering if $\eta_{1j} = \ldots = \eta_{Kj}$ and a model $M = \{K, \Omega\}$ is defined by the number of components K and the indexes of the relevant variables $\Omega \subset \{1, \ldots, J\}$. Therefore, considering the task of full model selection in (1) implies that each $\boldsymbol{X}_i$ is identically distributed according to a non-parametric mixture of K components defined by the density

$$g_{M,\boldsymbol{\psi}}(\boldsymbol{x}_i) = \left(\prod_{j \in \bar{\Omega}} \eta_{1j}(x_{ij}) \right) \left(\sum_{k=1}^{K} \pi_k \prod_{j \in \Omega} \eta_{kj}(x_{ij}) \right), \tag{2}$$

where $\bar{\Omega} = \{1, \ldots, J\} \setminus \Omega$ contains the indexes of the irrelevant variables for clustering and $\boldsymbol{\psi} \in \Psi_M$ groups the finite dimensional parameters $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)^\top \in \mathcal{S}_K$ and the infinite dimensional parameters composed by the univariate densities η_{kj}. In a non-model based framework, regularization methods can be used to achieve variable selection in clustering, among which the sparse K-means [8] is the most popular as it requires small computational overhead and is able to manage very high-dimensional datasets. The approach uses a lasso-type penalty to select the relevant variables . The selection of the number of components is difficult since probabilistic tools are not available and it is sensitive to the structure of the penalty term. Model selection is thus performed with an extension of the gap statistics. In the model-based framework, variable selection falls in the scope of model selection. Thus, here, full model selection is meant as the double objective of estimating K and Ω. Note that all existing methods of variables selection for model-based clustering are restricted to parametric distributions. Thus, if the parametric assumptions are violated, bias can occur for the estimator of the number of components or on

the subset of discriminative variables. To the best of our knowledge, this paper presents the first method that permits a full-model selection (*i.e.,* estimation of K and Ω) for non-parametric multivariate mixture models. Moreover, it allows to manage many variables, which makes it a complementary work of [4] and [6], even in the case where all the variables are considered to be relevant and only K needs to be estimated.

To achieve full model selection, we use a discretization of each continuous variable into B bins, which number tends to infinite with the sample size to ensure the consistency of the approach. Indeed, if B were fixed, the estimated model could be a sub-model of the true model (*i.e.,* K or Ω could be underestimated). The distribution of the resulting discretized variables follows a latent class model where each component is a product of multinomial distributions.This discretization permits to achieve model selection for the latent class model by using penalized likelihood (*e.g.,* BIC) whose consistency has been proved for mixture models [5]. Moreover, in this framework, a specific EM algorithm optimizing the penalized likelihood can be used for simultaneously detecting the subset of the relevant variables and estimating the model parameters [7], for a given number of components. Thus, by considering an upper-bound of the number of components, the full-model selection can be achieved. Unlike in [4], it provides a consistent estimation of the univariate densities η_{kj}. Therefore, we prove the consistency of the procedure, for a wide range of number of bins B, at an appropriate rate that is detailled in [3].

Section 2 details the discretization step. and states the consistency of the procedure. Section 3 presents the algorithm used for the full model selection while Section 4 is devoted to numerical experiments to compare the proposed approach to standard ones and illustrate its relevance.

2 Model selection by bin estimation and penalized log-likelihood

Each variable is discretized into B non-overlapping bins $I_{Bj1}, \ldots, I_{BjB}$ such that $\cup_{b=1}^{B} I_{Bjb} = \mathcal{X}_j$ and for any (b, b') with $b \neq b'$, $I_{Bjb} \cap I_{Bjb'} = \emptyset$. We denote by l_{Bjb} the size of the bin I_{Bjb} and by σ_{Bjb} its indicator function. Therefore, the pdf of the discretized subject i is

$$f_{M,B,\theta}(\boldsymbol{x}_i) = \prod_{j\in\bar{\Omega}} \prod_{b=1}^{B} \left(\frac{\alpha_{B1jb}}{l_{Bjb}}\right)^{\sigma_{Bjb}(x_{ij})} \left(\sum_{k=1}^{K} \pi_k \prod_{j\in\Omega} \prod_{b=1}^{B} \left(\frac{\alpha_{Bkjb}}{l_{Bjb}}\right)^{\sigma_{Bjb}(x_{ij})}\right), \tag{3}$$

where θ groups the component proportions π_k and the probabilities α_{Bkjb} that one subject drawn from component k takes level b for the variable j when this variable is discretized into B bins. The parameter space is given by the product of simplexes $S_K \times S_B^{K|\Omega|+(J-|\Omega|)}$.The probabilities α_{Bkjb} are unknown and must be estimated from the observed sample. The maximum likelihood statistics for a model M with B bins per variables is

$$T_{n,M,B} = \sup_{\theta\in\Theta_{M,B,\varepsilon}} \ell_n(f_{M,B,\theta}) \text{ with } \ell_n(f_{M,B,\theta}) = \sum_{i=1}^{n} \ln f_{M,B,\theta}(\boldsymbol{x}_i),$$

where, in order to avoid numerical issues, we introduced a threshold ε being the minimal value of all the elements defined in the simplexes. Fhrther details about this condition are given in

[3]. The penalized likelihood is defined by substracting a penality term $a_{n,M,B}$ which takes into account the sample size and the complexity of model (3). Depending on the choice of $a_{n,M,B}$, different well-known criterion can be considered, among them the AIC if $a_{n,M,B} = \nu$ or the BIC if $a_{n,M,B} = \nu \log(n)/2$, where $\nu = (K-1) + KJ(B-1)$ is the model complexity. The set of competing models $\mathcal{M}$ is defined by all the mixture models with at most $K_{\max}$ components and at least three relevant variables (for identifiability reasons)

$$\mathcal{M} = \{M = \{K, \Omega\} : K \leq K_{\max}, \Omega \subset \{1, \ldots, J\} \text{ and } |\Omega| \geq 3\}.$$

The estimator $\widehat{M}_{n,B}$ of the model maximizes the penalized likelihood as follows

$$\widehat{M}_{n,B} = \arg\max_{M \in \mathcal{M}} W_{n,M,B} \text{ with } W_{n,M,B} = T_{n,M,B} - a_{n,M,B}.$$

The study of the asymptotic properties of the estimator $\widehat{M}_{n,B}$ is covered by the approach of [5] if the number of intervals B does not change with the sample size n. However, due to the discretization, the approach provides a non-consistant estimator which converges to a model included into the true model. By increasing the number of intervals with n, we avoid the issues due to the loss of identifiability. However, we need to study the convergence of $T_{n,M,B}/n$ to the minimum Kullback divergence, which requires to control empirical processes over varying functional spaces. It relies on the locally-conic parametrization proposed by [1, 2].

Theorem 1 *Assume that independent data arise from* (2) *with true model* $M_0 = \{K_0, \Omega_0\}$ *and that the set of competing models* $\mathcal{M}$ *is defined with a known upper bound for the number of clusters* $K_{\max}$. *Then, under mild assumptions (see [3]),* $\widehat{M}_{n,B}$ *converges in probability to* M_0.

3 Estimation of the best model

The estimation of $\widehat{M}_{n,B}$ requires an optimization on a space of cardinal of order $2^J K_{\max}$, which excludes an exhaustive approach computing $W_{n,M,B}$ for each M in $\mathcal{M}$. Thus we follow the approach of [7] which, for a given K, performs simultaneously variable selection and parameter estimation via a specific EM algorithm to estimate

$$\widehat{M}_{n,B,K} = \arg\max_{\{M=(K,\Omega) \text{ with } \Omega \subseteq \{1,\ldots,J\} \text{ and } |\Omega| \geq 3\}} W_{n,M,B}.$$

Then, the following algorithm is performed for each K between 1 and $K_{\max}$. Parameter estimation is achieved by maximum likelihood and model selection is done with penalty $a_{n,K,\Omega,M} = \nu_{K,\Omega,M} c_n$ where $\nu_{K,\Omega,M}$ is the number of model parameters. By considering $c_n = (\ln n)/2$, this algorithm carries out the model selection according to the BIC. The algorithm considers a fixed K and starts at an initial point $(\Omega^{[0]}, \theta^{[0]})$. Its iteration $[r]$ is composed of two steps:
E-step Computation of the fuzzy partition

$$t_{ik}^{[r]} := \frac{\pi_k^{[r-1]} \prod_{j \in \Omega^{[r-1]}} \prod_{b=1}^{B} \left(\alpha_{Bkjb}^{[r-1]}\right)^{\sigma_{Bjb}(x_{ij})}}{\sum_{\ell=1}^{K} \pi_\ell^{[r-1]} \prod_{j \in \Omega^{[r-1]}} \prod_{b=1}^{B} \left(\alpha_{B\ell jb}^{[r-1]}\right)^{\sigma_{Bjb}(x_{ij})}},$$

M-step Maximization of the expectation of the penalized complete-data log-likelihood over Ω and θ such

$$\Omega^{[r]} = \{j : \Delta_j^{[r]} > 0\},\ \pi_k^{[r]} = \frac{n_k^{[r]}}{n} \text{ and } \alpha_{Bkjb}^{[r]} = \begin{cases} \tilde{\alpha}_{Bkjb}^{[r]} & \text{if } j \in \Omega^{[r]} \\ \bar{\alpha}_{Bkjb} & \text{otherwise} \end{cases},$$

where

$$\Delta_j^{[r]} = \sum_{i=1}^{n}\sum_{b=1}^{B} \sigma_{Bjb}(x_{ij}) \sum_{k=1}^{K} t_{ik}^{[r]} \ln\left(\frac{\tilde{\alpha}_{Bkjb}^{[r]}}{\bar{\alpha}_{Bkjb}}\right) - (K-1)(B-1)c_n$$

is the difference between the maximum of the expected value of the penalized complete-data log-likelihood obtained when variable j is relevant and when it is irrelevant, with

$$\tilde{\alpha}_{Bkjb}^{[r]} = \frac{1}{n_k^{[r]}} \sum_{i=1}^{n} t_{ik}^{[r]} \sigma_{jb}(x_{ij}), \quad \bar{\alpha}_{Bkjb} = \frac{1}{n} \sum_{i=1}^{n} \sigma_{jb}(x_{ij}) \quad \text{and } n_k^{[r]} = \sum_{i=1}^{n} t_{ik}^{[r]}.$$

4 Numerical comparison of methods

This section compares approaches for a full model selection on simulated data. We compare the proposed approach with a BIC applied on a Gaussian mixture model, with the sparse K-means approach and with the semi-parametric approach considering all the variables as relevant. The results of the proposed clustering method are obtained by performing full model selection with B levels defined by the empirical quantiles $1/B$,...,B/B where $B = [n^{1/6}]$ and a BIC like penalty and then by estimating the mixture components for the selected model by maximizing the smoothed log-likelihood with a bandwidth, for variable j, equal to $\hat{\sigma}_j n^{-1/5}$ where $\hat{\sigma}_j$ is the empirical standard deviation of variable j. To compare the different methods of clustering, we generate data from a mixture with three components and equal proportions ($\pi_k = 1/3$). The density of X_i given Z_i is a product of univariate densities such that $X_{ij} = \sum_{k=1}^{K} z_{ik}\delta_{kj} + \xi_{ij}$ where all the ξ_{ij} are independent and where $\delta_{11} = \delta_{12} = \delta_{23} = \delta_{24} = \delta_{35} = \delta_{36} = \tau$, while all remaining $\delta_{kj} = 0$, which implies that only the first six variables are relevant for clustering. Three distributions are considered for the ξ_{ij} (standard Gaussian, Student with three degrees of freedom and Laplace) and the value of τ is defined to obtain a theoretical misclassification rate of 5% (1.94, 2.60 and 2.52 for the Gaussian, Student and Laplace distributions respectively).

We investigate the accuracy of the estimated partition by computing the Adjusted Rand Index (ARI) between the true partition and the estimators of the partition given by the semi-parametric and the parametric methods when K is known and then when it is estimated. Moreover, to illustrate the benefit of variable selection, we also estimate the partition by considering all variables as relevant and the true number of components. Results are presented in Figure 1. Thus, when the parametric assumptions are satisfied, the parametric approach outperforms the proposed approach only on small samples (few observations with respect to the number of variables), whenever K is known or not. However, when the parametric assumptions are violated, the proposed approach strongly outperforms the parametric approach. Note that, when the number of irrelevant variables increases, the approach considering all the variables for clustering performs poorly (see row 100), illustrating the benefit of variable selection for clustering.

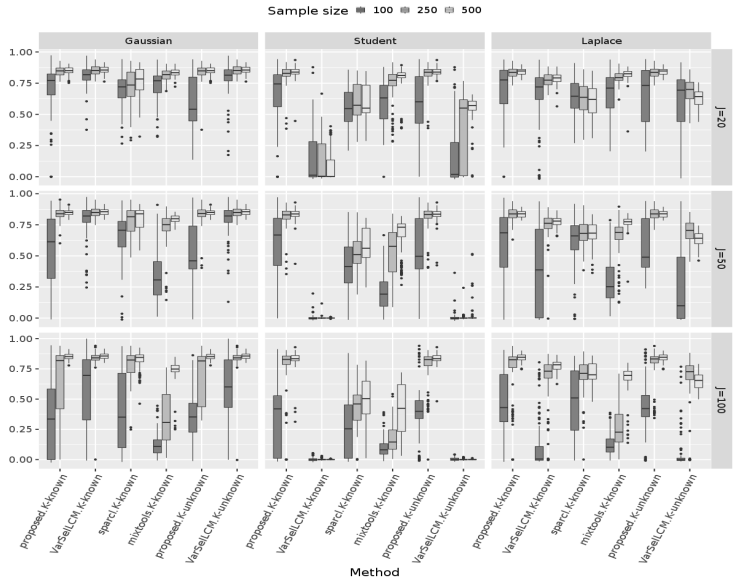


Figure 1: Boxplot of ARI obtained on the resulting partition when variable selection is performed with the true K by the proposed method (proposed.K-known), the parametric method (VarSelLCM.K-known), the sparse K-means (Sparcl.K-known) and the model considering all variables as relevant (mixtools.K-known) and when the full model selection is achieved by the proposed approach (proposed.K-unknown) and the parametric approach (VarSelLCM.K-unknown). Data are generated with theoretical misclassification rate of 5%.

References

[1] D. Dacunha-Castelle and E. Gassiat, *Testing in locally conic models, and application to mixture models*, ESAIM: Probability and Statistics, 1 (1997), pp. 285–317.

[2] ———, *Testing the order of a model using locally conic parametrization: population mixtures and stationary arma processes*, The Annals of Statistics, 27 (1999), pp. 1178–1209.

[3] M. Du Roy de Chaumaray and M. Marbac, *Full model estimation for non-parametric multivariate finite mixture models*, 2021.

[4] H. Kasahara and K. Shimotsu, *Non-parametric identification and estimation of the number of components in multivariate mixtures*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 76 (2014), pp. 97–111.

[5] C. Keribin, *Consistent estimation of the order of mixture models*, Sankhyā: The Indian Journal of Statistics, Series A, (2000), pp. 49–66.

[6] C. Kwon and E. Mbakop, *Estimation of the number of components of nonparametric multivariate finite mixture models*, The Annals of Statistics, 49 (2021), pp. 2178 – 2205.

[7] M. Marbac, M. Sedki, and E. Patin, *Variable selection for mixed data clustering: application in human population genomics*, Journal of Classification, (2019), pp. 1–19.

[8] D. Witten and R. Tibshirani, *A Framework for Feature Selection in Clustering*, Journal of the American Statistical Association, 105 (2010), pp. 713–726.

Modèles linéaires généralisés multivariés à composantes supervisées et facteurs latents, avec partitionnement thématique des variables explicatives

Julien GIBAUD [1] & Xavier BRY [1] & Catherine TROTTIER [1,2]

[1] *IMAG, CNRS, Univ. Montpellier, France / julien.gibaud@umontpellier.fr & xavier.bry@umontpellier.fr*

[2] *UPV Montpellier 3, Montpellier, France / catherine.trottier@univ-montp3.fr*

Résumé. À l'origine, la Régression Linéaire Généralisée sur Composantes Supervisées (SCGLR) a été conçue pour trouver, au sein de très nombreuses covariables redondantes, des composantes explicatives conjointement supervisées par plusieurs réponses, ce qui est nécessaire dans un contexte de grande dimension. Plus tard, SCGLR fut améliorée pour chercher des composantes au sein de variables explicatives partitionnées en thèmes. Dans ce travail, nous proposons d'étendre cette méthode en modélisant la matrice de variance-covariance conditionnelle des réponses de telle sorte que la covariance conditionnelle des réponses soit principalement expliquée par quelques facteurs. Nous chercherons donc non seulement à extraire des thèmes des composantes explicatives, mais aussi à identifier des blocs dans la matrice de variance-covariance des réponses conditionnellement à ces composantes. Après la linéarisation du modèle, un algorithme combinant EM et celui de SCGLR thématique est proposé afin d'estimer les paramètres du modèle.

Mots-clés. SCGLR, modèle à facteurs, algorithme EM, variables latentes

Abstract. Originally, the Supervised Component-based Generalized Linear Regression (SCGLR) was designed to extract, from a large set of redundant covariates, explanatory components jointly supervised by a set of responses, something much needed in a high-dimensional framework. Later on, SCGLR has been improved to search for components in explanatory variables divided up into thematic subsets. In this work, we propose to extend this methodology by modeling the responses' conditional variance-covariance matrix so that the conditional covariance of responses is mainly explained by few factors. Not only do we aim to extract explanatory components from the thematic subsets, but also to identify blocks in the conditional variance-covariance matrix of the responses. After linearising the model, we propose an algorithm combining EM with that of thematic SCGLR to estimate the model parameters.

Keywords. SCGLR, factor model, EM algorithm, latent variables.

1 Contexte

Dans leur article, Bry et al. (2013) proposent une méthode - la régression linéaire généralisée sur composantes supervisées (Supervised Component-based Generalized Linear Regression, SCGLR) - combinant le modèle linéaire généralisé multivarié avec les méthodes à composantes permettant la réduction de dimension. SCGLR optimise un critère compromis entre la qualité d'ajustement (Goodness-of-Fit, GoF) et la pertinence structurelle (Structural Relevance, SR, Bry and Verron, 2015) mesurant la proximité des composantes supervisées à des dimensions d'intérêt. Cette technique ne trouve pas seulement des directions fortes et interprétables, elle produit aussi des prédicteurs régularisés, ce qui permet le traitement de données de grande dimension. Bry et al. (2020) proposent de chercher ces composantes dans un partitionnement thématique des variables explicatives. Dans un contexte écologique par exemple, les variables de température et

de précipitation seraient considérées comme appartenant à deux thèmes distincts. Cependant, SCGLR suppose que les réponses sont indépendantes conditionnellement aux covariables. Pour nous affranchir de cette hypothèse, nous proposons d'étendre cette méthode aux modèles à facteurs. L'objectif est de modéliser la matrice de variance-covariance conditionnelle des réponses afin d'identifier des groupes de réponses liées.

2 Modélisation

Dans cette section, nous présentons la méthode SCGLR, sa généralisation au partitionnement des variables explicatives, puis son extension aux modèles à facteurs latents.

2.1 SCGLR

N individus sont décrits par K réponses y_k, $k = 1, \ldots, K$, ainsi que des covariables explicatives séparées en deux groupes : un groupe $X \in \mathbb{R}^{N \times P}$ de covariables *a priori* nombreuses et possiblement redondantes, et un autre $A \in \mathbb{R}^{N \times Q}$ de covariables additionnelles peu nombreuses et faiblement, voire non-redondantes. Chaque réponse y_k fait l'objet d'un modèle linéaire généralisé (Generalized Linear Model, GLM, McCullagh and Nelder, 1989). Pour la partie explicative du modèle, seule la matrice X requiert réduction de dimension et régularisation. À cette fin, SCGLR cherche dans X des composantes communes à l'ensemble des réponses. Une composante $f \in \mathbb{R}^N$ est donnée par $f = Xu$ où $u \in \mathbb{R}^P$ est un vecteur de coefficients. Dans un modèle à une composante, le prédicteur linéaire associé à la réponse y_k est ainsi donné par :

$$\eta_k = (Xu)\,\gamma_k + A\delta_k,$$

où γ_k et δ_k sont les paramètres de régression. La composante f est commune à l'ensemble des réponses y_k et pour assurer son identifiabilité, nous imposons $u^T u = 1$.

À cause du produit $u\gamma_k$, le modèle "linéarisé" à chaque étape de l'algorithme des scores de Fisher (Fisher Scoring Algorithm, FSA) pour l'estimation du GLM, n'est pas linéaire et doit être estimé de façon alternée sur u et sur $\{\gamma_k, \delta_k\}$. Soient w_k la pseudo-réponse (ou variable de travail) associée à chaque étape du FSA et W_k^{-1} sa matrice de variance-covariance, l'estimateur des moindres carrés de u est solution du programme d'optimisation suivant :

$$\max_{u, u^T u = 1} \psi_A(u) := \sum_{k=1}^{K} \|w_k\|^2_{W_k} \cos^2_{W_k}\left(w_k, \Pi^{W_k}_{\text{span}[f,A]} w_k\right).$$

La quantité ψ_A est une mesure de GoF. Pour trouver des composantes fortes et interprétables, le GoF ne suffit pas. Il faut le combiner avec une mesure de pertinence structurelle (SR).

Dans ce travail nous utilisons une mesure particulière de SR : l'inertie duale généralisée (Variable Powered Inertia, VPI). On appelle W la matrice des poids *a priori* des observations (typiquement, $W = \frac{1}{N} I_N$) et on suppose les colonnes de X centrées et réduites. Nous voulons trouver une direction span$[Xu]$ proche d'un faisceau de covariables (i.e. un ensemble de variables explicatives suffisamment corrélées pour être vues comme alignées autour de la même dimension latente). Pour cela, on considère un paramètre $l \geq 1$ et on formule la SR comme :

$$\phi(u) = \left(\frac{1}{P} \sum_{p=1}^{P} \langle Xu, x_p \rangle^{2l}_W\right)^{1/l}. \tag{1}$$

Le paramètre l permet de trouver une composante proche d'un faisceau plus (l fort) ou moins (l faible) étroit de variables corrélées.

Pour construire un compromis entre le GoF et la SR, SCGLR les combine en utilisant un paramètre $s \in [0,1]$ permettant de régler leurs poids relatifs et considère la maximisation suivante :

$$\max_{u,u^T u=1} s\ln(\phi(u)) + (1-s)\ln(\psi_A(u)).$$

2.2 Généralisation à un partitionnement thématique des variables explicatives

Bry et al. (2020) définissent le "modèle thématique" comme étant le modèle conceptuel indiquant que la matrice réponse Y dépend de R thèmes $X_1, \ldots, X_R$ plus un ensemble de covariables A, et que les dimensions structurellement pertinentes doivent être explicitement identifiées dans chaque X_r. Le prédicteur linéaire associé à la réponse y_k devient :

$$\eta_k = (X_1 u_1)\,\gamma_{k1} + \cdots + (X_R u_R)\,\gamma_{kR} + A\delta_k.$$

Pour ouvrir la voie à la régularisation des thèmes, nous devons adapter le critère compromis de SCGLR. En appelant $f_r = X_r u_r$ la composante du thème X_r, nous avons $\Pi^{W_k}_{\text{span}[f_1,\ldots,f_R,A]} = \Pi^{W_k}_{\text{span}[f_r,A_r]}$ où $A_r = [f_1, \ldots, f_{r-1}, f_{r+1}, \ldots, f_R, A]$. La mesure de GoF devient donc :

$$\psi_{A_r}(u_r) := \sum_{k=1}^{K} \|w_k\|^2_{W_k} \cos^2_{W_k}\left(w_k, \Pi^{W_k}_{\text{span}[f_r,A_r]} w_k\right),$$

où les covariables supplémentaires devront être spécifiques à chaque occasion. La mesure de SR reste inchangée en prenant $\phi(u_r)$ comme en (1). Ainsi on obtient un compromis entre le GoF et la SR spécifique à chaque thème. Enfin, le programme d'optimisation peut être résolu en maximisant itérativement le critère compromis sur chaque u_r :

$$\forall r, \quad \max_{u_r,u_r^T u_r=1} s\ln(\phi(u_r)) + (1-s)\ln(\psi_{A_r}(u_r)). \tag{2}$$

Afin de trouver les composantes d'ordre $h > 1$, nous notons $f_r^h = X_r u_r^h$ la h-ième composante du thème X_r et $F_r^h = [f_r^1, \ldots, f_r^h]$, où $h \leq H_r$, la matrice des h premières composantes du thème X_r. La nouvelle composante f_r^{h+1} doit venir compléter au mieux les composantes précédentes en plus de la matrice A, c'est à dire $A_r^h := [F_1^{H_1}, \ldots, F_{r-1}^{H_{r-1}}, F_r^h, F_{r+1}^{H_{r+1}}, \ldots, F_R^{H_R}, A]$. Ainsi, f_r^{h+1} est calculée en utilisant A_r^h comme nouvelle matrice de covariables additionnelles. De plus, nous imposons que f_r^{h+1} soit orthogonale à F_r^h par la contrainte ${F_r^h}^T W f_r^{h+1} = 0$. Cette maximisation sous contrainte est permise par l'algorithme du gradient normé projeté itéré (Projected Iterated Normed Gradient, PING).

2.3 SCGLR thématique pour modèles à facteurs

Dans cette section, les composantes sont considérées comme connues. Ainsi, pour des raisons de simplicité, nous prenons la matrice $F = [F_1^{H_1}, \ldots, F_R^{H_R}]$ comme la nouvelle matrice des variables explicatives et $\gamma_k = (\gamma_{k1}, \ldots, \gamma_{kR})^T$ le vecteur des paramètres de régression associé à la réponse y_k. Pour une unité statistique n, chaque réponse est modélisée linéairement à l'aide de J variables latentes aléatoires $g_n = (g_{n1}, \ldots, g_{nJ})^T$ appelées facteurs :

$$\eta_{nk} = f_n^T \gamma_k + a_n^T \delta_k + g_n^T b_k,$$

où f_n et a_n sont les vecteurs composés par les n-ième lignes des matrices F et A respectivement, et b_k le vecteur des paramètres de régression de g_n. Les facteurs suivent une loi normale multivariée $g_n \sim \mathcal{N}_J(0, I_J)$ et on suppose qu'ils sont indépendants entre les unités statistiques. Le modèle est construit de telle manière que les facteurs capturent la covariance entre les réponses conditionnellement aux composantes et covariables supplémentaires. En notant $G \in \mathbb{R}^{N\times J}$ la matrice contenant toutes les réalisations des facteurs, le prédicteur linéaire associé à la réponse y_k exprimé en colonne devient donc :

$$\eta_k = F\gamma_k + A\delta_k + Gb_k.$$

Soit $B = [b_1, \ldots, b_K] \in \mathbb{R}^{J\times K}$ la matrice des coefficients des facteurs. Comme montré par Geweke and Zhou (1996), cette matrice n'est pas unique. Pour garantir l'identification du modèle, nous devons imposer que la sous matrice de B de taille $J \times J$ soit un triangle supérieur avec des éléments diagonaux tous positifs. Un avantage principal des modèles à facteurs est d'exprimer la matrice $\Sigma = B^T B \in \mathbb{R}^{K\times K}$ des covariances conditionnelles des réponses de manière parcimonieuse. En effet, outre la contrainte triangulaire, le nombre de facteurs choisi est petit.

Dans un objectif de clarté, la figure 1 présente de manière graphique le modèle thématique avec facteurs.

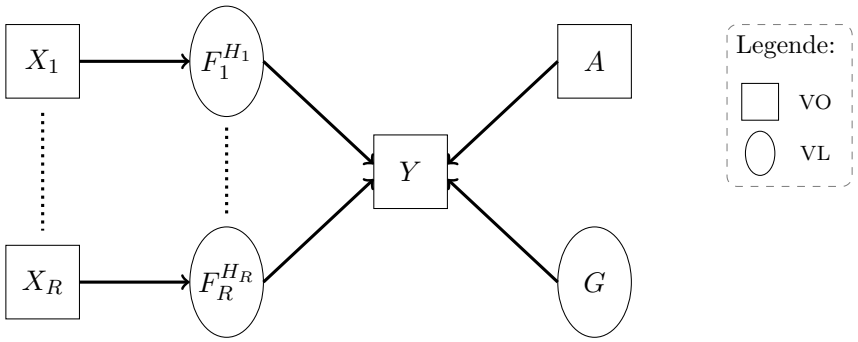


Figure 1: Représentation graphique de SCGLR thématique pour modèles à facteurs. L'intervention d'un groupe de variables sur un autre est représentée par une flèche. Les variables observées (VO) sont placées dans des carrés tandis que les ovales contiennent les variables latentes (VL).

3 Estimation du modèle

Soit $\Theta = \{\gamma_k, \delta_k, b_k; k = 1, \ldots, K\}$ l'ensemble des paramètres. La log-vraisemblance du modèle est obtenue en intégrant sur les facteurs g_n :

$$\begin{aligned} l(\Theta; Y) &= \sum_{n=1}^{N} \ln\left(L(y_n; \Theta)\right) \\ &= \sum_{n=1}^{N} \ln\left(\int \prod_{k=1}^{K} L(y_{nk}|g_n; \Theta) L(g_n)\, dg_n\right). \end{aligned}$$

Cependant, dans un contexte de réponses non Gaussiennes, la maximisation de cette log-vraisemblance n'est pas possible. Dans l'esprit de Saidane et al. (2013), nous avons choisi d'effectuer l'estimation en deux étapes.

3.1 Linéarisation du modèle

Nous considérons d'abord que les facteurs sont connus, c'est à dire que, conditionnellement à G, nous nous plaçons dans le cadre des GLM multivariés classiques. Soient h_k la fonction de lien canonique associée à la réponse y_k, h'_k sa première dérivée et μ_{nk} le paramètre d'espérance pour l'unité statistique n. La pseudo réponse w_{nk} est alors calculée comme le développement limité au premier ordre de h_k au point μ_{nk} :

$$\begin{aligned} w_{nk} &= h_k(\mu_{nk}) + (y_{nk} - \mu_{nk})\, h'_k(\mu_{nk}) \\ &= \eta_{nk} + \zeta_{nk}, \end{aligned}$$

où $\zeta_{nk} = (y_{nk} - \mu_{nk})\, h'_k(\mu_{nk})$. Ce développement conduit alors au modèle linéarisé conditionnel suivant :

$$w_k = F\gamma_k + A\delta_k + Gb_k + \zeta_k,$$

où $\mathbb{E}[w_k|G] = F\gamma_k + A\delta_k + Gb_k$ et $\mathbb{V}[w_k|G] = \mathbb{V}[\zeta_k] = W_k^{-1}$.

3.2 Estimation à l'aide de l'algorithme EM

La deuxième étape consiste à supposer Gaussiennes les pseudo réponses désormais connues. Les facteurs aléatoires étant latents, la log-vraisemblance du modèle linéarisé $l(\Theta; \mathcal{W})$, où $\mathcal{W}$ désigne la matrice des pseudo réponses, possède une expression complexe qui la rend difficile à maximiser. Ainsi, nous utilisons l'algorithme EM (Dempster et al., 1977) pour estimer les paramètres. Nous calculons puis maximisons l'espérance de la log-vraisemblance complétée $l(\Theta; \mathcal{W}, G)$ conditionnellement aux pseudo réponses.

4 Algorithme

L'algorithme 1 alterne le calcul des composantes et celui des autres paramètres via EM. Des essais numériques, sur données simulées et réelles, testant les performances de l'algorithme, seront exposés lors de la présentation orale de ce travail.

Remerciements

Cette recherche a été soutenue par le projet GAMBAS financé par l'Agence Nationale de la Recherche (ANR-18-CE02-0025).

References

Bry, X., Trottier, C., Mortier, F., and Cornu, G. (2020). Component-based regularization of a multivariate GLM with a thematic partitioning of the explanatory variables. *Statistical Modelling*, 20(1):96–119.

Bry, X., Trottier, C., Verron, T., and Mortier, F. (2013). Supervised component generalized linear regression using a PLS-extension of the Fisher scoring algorithm. *Journal of Multivariate Analysis*, 119:47–60.

Bry, X. and Verron, T. (2015). THEME: THEmatic Model Exploration through multiple costructure maximization. *Journal of Chemometrics*, 29(12):637–647.

Algorithm 1: SCGLR thématique pour les modèles à facteurs

while *not convergence* **do**

Calculer les composantes grâce à la maximisation du critère (2) à l'aide de l'algorithme PING

$\forall r = 1, \ldots, R, \forall h = 1, \ldots, H_r, \quad f_r^{h\,(t+1)} = X_r u_r^{h\,(t+1)}$

Calculer les pseudo réponses à l'aide du FSA

$\eta_k^{(t+1)} = F^{(t+1)}\gamma_k^{(t)} + A\delta_k^{(t)} + Gb_k^{(t)}$

$\mu_{nk}^{(t+1)} = h_k^{-1}\left(\eta_{nk}^{(t+1)}\right), \forall n = 1, \ldots, N$

$w_{nk}^{(t+1)} = \eta_{nk}^{(t+1)} + h_k'\left(\mu_{nk}^{(t+1)}\right)\left(y_{nk} - \mu_{nk}^{(t+1)}\right), \forall n = 1, \ldots, N$

$W_k^{(t+1)} = \text{diag}\left(\left[a_{nk}(\phi_k)v_k\left(\mu_{nk}^{(t+1)}\right)h_k'\left(\mu_{nk}^{(t+1)}\right)^2\right]^{-1}\right)_{n=1,\ldots,N}$

Calculer les paramètres à l'aide l'algorithme EM

$\Theta^{(t+1)} = \underset{\Theta}{\text{argmax}}\; l(\Theta^{(t)}; \mathcal{W})$

Incrémenter

$t \leftarrow t + 1$

end

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

Geweke, J. and Zhou, G. (1996). Measuring the pricing error of the arbitrage pricing theory. *The review of financial studies*, 9(2):557–587.

McCullagh, P. and Nelder, J. (1989). Generalized Linear Models, Chapman and Hall, New York, NY.

Saidane, M., Bry, X., and Lavergne, C. (2013). Generalized linear factor models: A new local EM estimation algorithm. *Communications in Statistics-Theory and Methods*, 42(16):2944–2958.

Calibrations de pénalités pour la régression linéaire gaussienne en grande dimension

Perrine Lacroix [1,2]

[1] *Université Paris-Saclay, CNRS, Laboratoire de Mathématiques d'Orsay, Rue Michel Magat Bâtiment 307, 91405 Orsay, France.*
[2] *Université Paris-Saclay, CNRS, INRAE, Institut des Sciences des Plantes de Paris-Saclay (IPS2), Rue Noetzlin Bâtiment 630, 91190 Gif-sur-Yvette, France.*

perrine.lacroix@universite-paris-saclay.fr

Résumé. Dans un contexte de grande dimension, une approche classique pour estimer le paramètre inconnu en régression linéaire gaussienne est de minimiser les moindres carrés pénalisés. Pour obtenir une inégalité oracle sur le risque prédictif, la théorie développée par [Birgé and Massart, 2001] fournit des formes de pénalité connues à constantes multiplicatives près. Cependant, le groupe de variables sélectionnées ne correspond pas aux variables dites actives, c'est-à-dire celles intervenant réellement dans la régression linéaire pour expliquer Y. Ainsi, contrôler la qualité de prédiction n'est pas suffisant pour limiter la sélection des variables dites inactives (par opposition aux variables actives). C'est pourquoi, dans un premier temps et sous un modèle simplifié, nous proposons une étude théorique et une heuristique dépendante des données permettant de calibrer l'une des constantes de la pénalité pour éviter la sélection de variables inactives tout en conservant une bonne qualité de prédiction. Dans un second temps, nous proposons un algorithme qui étend le principe de l'heuristique de pente pour calibrer les deux constantes restantes tout en assurant un contrôle du risque prédictif.

Mots-clés. Sélection de modèle, Grande dimension, Pénalisation, Calibration de constantes.

Abstract. In a high-dimensional context, a classical approach to estimate the unknown parameter in a Gaussian linear regression consists in minimizing the penalized least-squares criterion. To get an oracle inequality on the predictive risk, the theory developed by [Birgé and Massart, 2001] gives some penalty shapes known up to multiplicative constants. However, the group of the selected variables does not correspond to the so-called active variables, which are those actually involved in the linear regression to explain Y and therefore. Thus, to control the prediction quality is not sufficient to limit the selection of the so-called inactive variables (as opposed to the active variables). That is why, in a first step, under a simplified model, we propose a theoretical study and a data-dependent heuristics to calibrate one of the constants to avoid selecting inactive variables while maintaining a high prediction quality. Secondly, we propose an algorithm that extends the slope heuristics principle to calibrate the last two constants while maintaining a reasonnable control of the predictive risk.

Keywords. Model selection, High-dimension, Penalization, Constants calibration.

1 Le modèle Statistique

Nous considérons le modèle de régression linéaire gaussien suivant:

$$Y = X\beta^* + \varepsilon \tag{1}$$

La variable réponse Y, constituée de n observations indépendantes, est régressée par p variables, chacune de taille n, notées $X_1, \cdots, X_p$ et résumée dans la matrice de design fixe X de taille $n \times p$. Le bruit ε suit une distribution gaussienne centrée et de covariance $\sigma^2 I_n$. Le paramètre inconnu β^* est de taille p. Si $\beta_j^* \neq 0$, alors la variable associée X_j est impliquée dans la relation linéaire et est appelée variable active. En opposition, si $\beta_j^* = 0$, alors X_j est dite inactive. L'objectif est d'identifier l'ensemble des variables actives, ceci à travers l'estimation de β^*. Nous nous intéressons au cadre de la grande dimension: p est grand ($p \sim n$ voire $p \geq n$) mais pas excessivement pour éviter le cadre de la ultra haute dimension [Verzelen, 2012]. Dans ce contexte de grande dimension, les outils statistiques classiques basés sur une estimation empirique ne parviennent pas à retrouver les variables actives. Des hypothèses de régularité doivent être ajoutées au modèle. Ici, nous nous plaçons dans une hypothèse de parcimonie, c'est-à-dire que le nombre de variables actives est faible par rapport à p. La procédure de sélection de modèle prends en compte ce fléau de la grande dimension et permet d'estimer β^* d'un point de vue prédictif: les variables sélectionnées permettent de prédire Y.

2 La sélection de modèles

Les méthodes de sélection de modèle sont basées sur trois étapes. La première consiste à générer une collection $\mathcal{M}$ contenant quelques sous-ensembles de variables de tailles variées. Cette étape évite l'exploration de tous les sous-ensembles de variables possibles parmi les p variables et dont le nombre grandit exponentiellement avec p. Les variables finales sélectionnées seront celles d'un des sous-ensembles de la collection. La seconde étape consiste simplement en la minimisation des moindres carrés à l'intérieur de chaque sous-ensemble afin d'obtenir une collection d'estimateurs sans biais:

$$\hat{\beta}_m := \underset{\beta,\ \text{s.t.}\ X\beta \in m}{\arg\min} \big\{ ||Y - X\beta||_2^2 \big\},$$

où $||.||_2$ est la norme euclidienne usuelle. Enfin, la dernière étape consiste à sélectionner $\hat{\beta}_{\hat{m}}$ par la résolution du problème de minimisation:

$$\hat{m} = \underset{m \in \mathcal{M}}{\arg\min} \Big\{ ||Y - X\hat{\beta}_m||_2^2 + \text{pen}(D_m) \Big\}, \tag{2}$$

où D_m désigne la dimension du modèle m et pen est une fonction de pénalité dépendant de D_m. [Birgé and Massart, 2007] proposent des formes de pénalité en fonction de la richesse de la collection de telle sorte qu'un contrôle optimal non-asymptotique sur le risque prédictif (RP) est obtenu, avec comme définition du risque prédictif pour $m \in \mathcal{M}$:

$$\mathrm{RP}(m) = \mathbb{E}_{\tilde{Y}}\Big[||\tilde{Y} - \tilde{X}\hat{\beta}_m||_2^2\Big],$$

où $\mathbb{E}_{\tilde{Y}}$ désigne l'espérance sous tous les $\tilde{Y}$ satisfaisant (1)

3 Compromis entre risque prédictif (RP) et taux de fausses découvertes (FDR)

3.1 Cadre

Simplifions dans un premier temps le problème en supposant que la variance σ^2 est connue et que la collection de modèle est donnée par:

$$\mathcal{M} = \Big\{m_0 = \emptyset, m_1 = \mathrm{Vect}(X_1), m_2 = \mathrm{Vect}(X_1, X_2), \cdots, m_q = \mathrm{Vect}(X_1, X_2, \cdots, X_q)\Big\}, \tag{3}$$

où $q = \min(n, p)$. Cette collection fixe de modèles emboîtés peut être utilisée en pratique si un ordre d'importance au sens de la régression (1) sur les X_j est connu. Supposons enfin que le vrai modèle m^* appartient à $\mathcal{M}$. Puisque la taille de la collection (3) est linéaire en p, un choix approprié de pénalité pour obtenir une inégalité oracle sur le RP est $\mathrm{pen}(D_m) = KD_m$ avec $K > 1$ [Birgé and Massart, 2007]. Le choix du paramètre libre K a été longuement discuté à la suite de ces travaux pionniers et $K = 2$ est utilisée en pratique car cette constante apparaît lors de considérations asymptotiques. Le False Discovery Rate (FDR), défini pour $m \in \mathcal{M}$ par:

$$\mathrm{FDR}(m) = \mathbb{E}\Big[\frac{\mathrm{FP}(m)}{D_m \vee 1}\Big],$$

où $FP(m)$ désigne le nombre de variables inactives contenues dans m, permet d'évaluer la proportion de variables inactives sélectionnées. Notons $\hat{m}(K)$ le modèle sélectionné dans (2) avec $\mathrm{pen}(D_m) = KD_m$. En nous basant sur des simulations, nous observons que le RP associé à $\hat{m}(2)$ est petit, cependant ce n'est pas le cas pour le FDR. Ceci suggère que certaines variables sélectionnées sont inactives. Rendre la procédure plus conservative est une idée pour éviter leur sélection et une solution est d'augmenter la pénalisation. Dans $\mathrm{pen}(D_m) = KD_m$, le seul paramètre libre est K que nous proposons d'augmenter. Une faible valeur du RP nous garantit qu'assez de variables actives sont sélectionnées tandis qu'une faible valeur du FDR nous garantit que peu de variables inactives sont sélectionnées. C'est pourquoi, pour choisir K, nous proposons d'intégrer un contrôle du FDR dans la procédure de sélection de modèle en plus d'un contrôle du RP déjà garantit par celle-ci.

3.2 Le choix de K

Bien que la procédure de sélection de modèle ait été construite pour un contrôle du RP, nous proposons une borne inférieure et une borne supérieure non-asymptotiques de la fonction $\left[K \longrightarrow \mathrm{FDR}(\hat{m}(K))\right]$. Ces bornes sont satisfaisantes car, contrairement au FDR, elles sont totalement explicites et déterministes. Lors de l'exposé, ces bornes théoriques seront présentées. Celles-ci dépendent du paramètre β^*, inconnu en pratique. Nous proposons de le remplacer par un estimateur judicieusement choisi. Enfin, à partir d'un jeu de données, nous proposons une heuristique dépendante des données pour calibrer K. Celui-ci repose sur l'étude des bornes obtenues et évaluées en un estimateur bien choisi de β^* simultanément avec l'estimation de la fonction $\left[K \longrightarrow \mathrm{RP}(\hat{m}(K))\right]$, obtenue en séparant classiquement le jeu de données en deux: le jeu d'entraînement sur lequel $\hat{m}(K)$ est calculé pour chaque K et un jeu test sur lequel le risque est évalué (Figure 1). L'objectif est de garantir une faible valeur pour le FDR tout en conservant de bonnes qualités de prédiction.

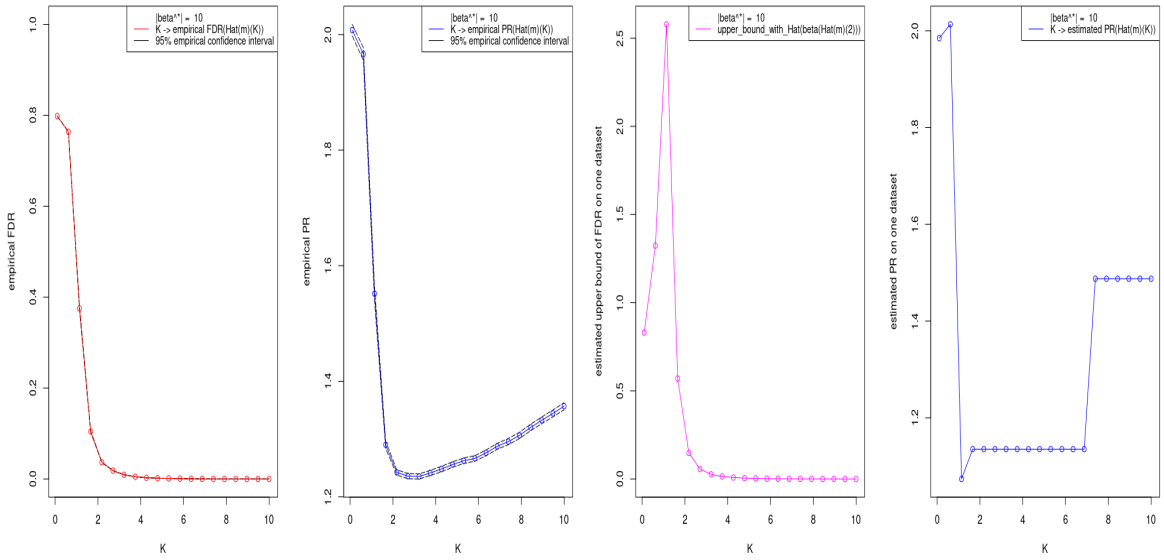


Figure 1: Les deux figures de gauche: courbes des estimations empiriques de $\left[K \longrightarrow \mathrm{FDR}\big(\hat{m}(K)\big)\right]$ et $\left[K \longrightarrow \mathrm{RP}\big(\hat{m}(K)\big)\right]$. Les deux figures de droite: courbes de la borne sup estimée du FDR et de la courbe du RP obtenue sur un jeu de données

4 Heuristique de pente en dimension 2 pour la prédiction

4.1 Cadre

Cette fois-ci, nous supposons que la variance σ^2 est inconnue et qu'il n'y a aucune connaissance a priori sur les modèles et donc sur la collection $\mathcal{M}$. Dans ce cas, celle-ci est générée aléatoirement, par exemple en minimisant le critère Lasso [Tibshirani, 1996] ou Elastic-Net [Zou and Hastie, 2005] et un choix approprié de pénalité pour obtenir une qualité prédictive optimale est [Birgé and Massart, 2007]:

$$\text{pen}(m) = 2\left(C_1(\sigma^2)D_m + C_2(\sigma^2)\log\left(\binom{p}{D_m}\right)\right) \tag{4}$$

où $C_1(\sigma^2)$ et $C_2(\sigma^2)$ sont deux constantes qui ne dépendent que de σ^2. Généralement la pénalité est utilisée en supposant que $C_1(\sigma^2) = \sigma^2 C_1$, $C_2(\sigma^2) = \sigma^2 C_2$ et en fixant le rapport $\frac{C_1}{C_2}$ à 2.5, valeur obtenue par [Lebarbier, 2005] lors d'une large étude de simulation dans le cadre de la détection de rupture avec collection fixe. La pénalité définie en (4) s'écrit alors

$$\text{pen}(m) = 2\kappa(\sigma^2)\left(2.5D_m + \log\left(\binom{p}{D_m}\right)\right)$$

où $\kappa(\sigma^2)$ est calibrée à l'aide des données par la méthode de l'heuristique de pente implémentée dans le package capushe du logiciel R [Baudry et al., 2012]. Cette calibration repose sur la relation linéaire attendue entre les valeurs du risque empirique et les valeurs $\left(2.5\frac{D_m}{n} + \frac{\log\left(\binom{p}{D_m}\right)}{n}\right)$ pour les modèles m de grande dimension (Figure 2 tirée de [Baudry et al., 2012]).

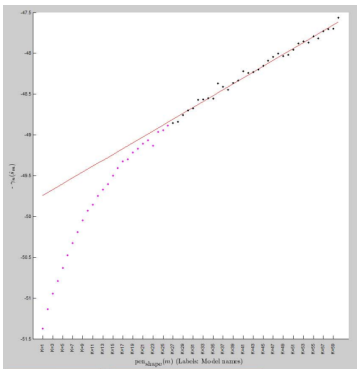


Figure 2: Illustration de la méthode de l'heuristique de pente.

5 Étude de simulation

Nous proposons de tester cette pénalité (4) lorsque la collection n'est plus fixe mais aléatoire et dépendante des données. Dans un premier temps, nous avons étudié le risque prédictif associé à des valeurs de C_1 et C_2 prises dans une grille fixée au préalable. Nous avons observé sur des simulations qu'il est toujours possible d'observer un comportement linéaire entre les vecteurs $\left(\frac{D_m}{n}, \frac{\log\left(\binom{p}{D_m}\right)}{n}\right)$ et les valeurs du risque empirique pour les modèles de grandes dimensions. Cependant, en selectionnant le couple de constantes $\left(\hat{C}_1, \hat{C}_2\right)$ qui offre le plus petit risque prédictif sur la grille considéree, nous constatons que le rapport $\frac{\widehat{C_1}}{\widehat{C_2}}$ n'est plus constant. Dans cet exposé, je proposerai un nouvel algorithme. Celui-ci permet de s'affranchir du choix d'une grille pour C_1 et C_2 car il permet de calibrer directement et à l'aide des données les deux constantes $C_1(\sigma^2)$ et $C_2(\sigma^2)$.

6 Remerciements

Ce travail a été soutenu par une subvention publique dans le cadre du projet Investissement d'avenir, reference ANR-11-LABX-0056-LMH, LabEx LMH.

References

[Baudry et al., 2012] Baudry, J.-P., Maugis, C., and Michel, B. (2012). Slope heuristics: overview and implementation. *Statistics and Computing*, 22(2):455–470.

[Birgé and Massart, 2001] Birgé, L. and Massart, P. (2001). Gaussian model selection. *Journal of the European Mathematical Society*, 3(3):203–268.

[Birgé and Massart, 2007] Birgé, L. and Massart, P. (2007). Minimal penalties for gaussian model selection. *Probability theory and related fields*, 138(1-2):33–73.

[Lebarbier, 2005] Lebarbier, É. (2005). Detecting multiple change-points in the mean of gaussian process by model selection. *Signal processing*, 85(4):717–736.

[Tibshirani, 1996] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

[Verzelen, 2012] Verzelen, N. (2012). Minimax risks for sparse regressions: Ultra-high dimensional phenomenons. *Electronic Journal of Statistics*, 6:38–90.

[Zou and Hastie, 2005] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320.

Beyond L1: Faster and Better Sparse Models with skglm

Mathurin Massias [1] & Quentin Bertrand [2] & Quentin Klopfenstein [3]

[1] *INRIA, 46 allée d'Italie, Lyon, 69007, mathurin.massias@inria.fr*
[2] *Mila, Université de Montréal, Montréal, quentin.bertrand@mila.quebec*
[3] *LCSB, Université du Luxembourg, quentin.klopfenstein@uni.lu*

Résumé. Nous proposons un nouvel algorithme rapide pour les modèles linéaires généralisés parcimonieux avec pénalités séparables non convexes. Notre algorithme est capable de résoudre des problèmes avec des millions de variables et d'observations en quelques secondes, en s'appuyant sur la descente de coordonnées, les *working sets* et l'accélération d'Anderson. Une version longue est disponible à : https://arxiv.org/abs/2204.07826.

Mots-clés. Optimisation, parcimonie, Anderson, Lasso, descente par coordonnée

Abstract. We propose a new fast algorithm to estimate sparse generalized linear models with non-convex separable penalties. It is able to solve problems with millions of samples and features in seconds, by relying on coordinate descent, working sets and Anderson acceleration. A long version is at https://arxiv.org/abs/2204.07826.

Keywords. Optimization, sparsity, acceleration, Anderson, Lasso, coordinate descent

1 Introduction

Les modèles linéaires généralisés parcimonieux jouent un rôle central dans l'apprentissage moderne. L'impact pratique du Lasso (Tibshirani, 1996) et ses dérivés a été rendu possible par deux facteurs clés : des algorithmes, et des implémentation logicielles efficaces.

Les algorithmes de pointe pour les problèmes "séparables lisses + non lisses" reposent principalement sur la descente de coordonnées (CD, Tseng et Yun 2009), qui, lorsqu'elle peut être appliquée, est plus efficace que les méthodes de gradient complet. La descente par coordonnées peut même être améliorée avec une accélération de type Nesterov, pour obtenir de meilleurs taux de convergence (Fercoq et Richtárik, 2015). Cependant, ces meilleurs taux peuvent ne pas se refléter dans les accélérations pratiques. Au contraire, Bertrand et Massias (2021) se sont appuyés sur l'accélération d'Anderson (Anderson, 1965) pour fournir à la fois de meilleurs taux *et* une accélération pratique.

Malgré tout, même ces algorithmes efficaces sont considérés comme trop lents pour des problèmes comportant plus de millions de variables. Pour atteindre cette échelle, on peut utiliser des règles de filtrage, qui éliminent les prédicteurs non importants du problème

(El Ghaoui *et al.*, 2010; Bonnefoy *et al.*, 2015; Fercoq *et al.*, 2015). Une aultre alternative est de résoudre de manière itérative des sous-problèmes de plus en plus grands (Massias *et al.*, 2018), qui incluent progressivement les variables identifiées comme pertinentes.

Du point de vue statistique, le Lasso souffre de limites importantes : les modèles parcimonieux non convexes bénéficient de meilleures propriétés théoriques et empiriques (Soubies *et al.*, 2015). Pourtant, ils n'ont pas été appliqués aussi souvent à grande échelle, principalement pour une raison algorithmique : alors que la descente de coordonnées peut être appliquée à des pénalités non convexes (Breheny et Huang, 2011), les règles actuelles de filtrage et les ensembles de travail dépendent fortement de la convexité (Rakotomamonjy *et al.*, 2022). Nous proposons, à notre connaissance, le premier algorithme d'ensembles de travail pour résoudre les modèles linéaires généralisés parcimonieux *génériques*.

2 Cadre et hypothèses

Dans cet article, nous considérons des problèmes de la forme :

$$\hat{\beta} \in \operatorname*{argmin}_{\beta \in \mathbb{R}^p} \Phi(\beta) \triangleq \underbrace{F(X\beta)}_{\triangleq f(\beta)} + \sum_{j=1}^{p} g_j(\beta_j) \ , \tag{1}$$

où F est une fonction lisse, et les fonctions g_j sont des fonctions semi-continues inférieures propres mais pas nécessairement convexes, dont l'opérateur proximal peut être calculé exactement. On écrit $g = \sum_j g_j$. Les instances de (1) incluent des estimateurs convexes : le Lasso, l'elastic net, la régression logistique parcimonieuse, le dual du SVM. Elles incluent également des pénalités non convexes : les pénalités $\ell_{0.5}$ et $\ell_{2/3}$, la pénalité concave minimax (MCP, Zhang 2010) ou SCAD (Zhang, 2010).

Hypothèse 1 *$f : \mathbb{R}^p \to \mathbb{R}$ est convexe et différentiable et pour tout $j \in [p]$, la restriction de $\nabla_j f$ à la j-ième coordonnée est L_j-Lipschitz : pour tout $(x, h) \in \mathbb{R}^p \times \mathbb{R}$. $|\nabla_j f(x + he_j) - \nabla_j f(x)| \leq L_j |h|$.*

Hypothèse 2 *Pour tout $j \in [p]$, $g_j : \mathbb{R} \to \mathbb{R}$ est propre, fermée, semi-continue inférieure et bornée inférieurement.*

Définition 3 *En utilisant la sous-différentielle de Fréchet, un point critique $x \in \mathbb{R}^p$ est un point qui satisfait: $-\nabla f(x) \in \partial g(x)$.*

Les Hypothèses 1 et 2 sont classiques et, sous réserve que les itérés soient bornés, assurent la convergence des algorithmes *forward-backward* et par coordonnées vers un point critique (Bolte *et al.*, 2014, Thm. 3.1). Notre travail se concentre sur le cas où les g_j présentent des points de non-différentiabilité.

Algorithm 1 `skglm` (proposed)

input: $X, \beta \in \mathbb{R}^p, n_{\text{out}} \in \mathbb{N}, n_{\text{in}} \in \mathbb{N}$, ws_size $\in \mathbb{N}, \epsilon > 0$

for $t = 1, \dots, n_{\text{out}}$ **do**

- score $= \left(\text{dist}\left(-\nabla_j f(\beta), \partial g_j(\beta_j)\right) \right)_{j \in [p]}$
- ws_size $= \max(\text{ws_size}, 2|\,\text{gsupp}(\beta)|)$
- `// ws_size features with largest scores`
- ws $= \text{arg_topK}(\text{score}, \text{K} = \text{ws_size})$
- **if** $\max_{j \in [p]} \text{dist}\left(-\nabla_j f(\beta), \partial g_j(\beta_j)\right) \leq \epsilon$ **then** stop
- **else** `// accelerated CD on working set`
- $\beta \leftarrow$ `inner_solver`$(X, \beta, \text{ws}, n_{\text{in}}, \epsilon)$

return β

Algorithm 2 `inner_solver`

input: $X, \beta^{(0)} \in \mathbb{R}^p, \text{ws} \subset [p], n_{\text{in}} \in \mathbb{N}$, $\epsilon > 0$, $M = 5$

for $k = 1, \dots, n_{\text{in}}$ **do**

- $\beta^{(k)} \leftarrow \texttt{CD}(X, \beta^{(k-1)}, X\beta, \text{ws})$
- **if** $k \bmod M = 0$ **then**
  - $\beta^{\text{extr}}_{\text{ws}} \leftarrow \texttt{Anderson}(\beta^{(k-M)}_{\text{ws}}, \dots, \beta^{(M)}_{\text{ws}})$
  - **if** $\Phi(\beta^{\text{extr}}_{\text{ws}}) < \Phi(\beta^{(k)}_{\text{ws}})$ **then**
    - $\beta^{(k)}_{\text{ws}} \leftarrow \beta^{\text{extr}}_{\text{ws}}$; $X\beta \mathrel{+}= X_{\text{ws}}(\beta^{\text{extr}}_{\text{ws}} - \beta^{(k)}_{\text{ws}})$
- **if** $\max_{j \in \text{ws}} \text{dist}\left(-\nabla_j f(\beta), \partial g_j(\beta_j)\right) \leq \epsilon$ **then** stop

return ws

Définition 4 (support généralisé) *Le* support généralisé *de $\beta \in \mathbb{R}^p$ est l'ensemble des indices tels que g_j est différentiable en β_j :* $\text{gsupp}(\beta) = \{j \in [p] : \partial g_j(\beta_j) \text{ est un singleton}\}$.

Les pénalités telles que ℓ_1, ℓ_q $(0 < q < 1)$, MCP ou SCAD sont seulement non différentiables en 0, et cela correspond à la notion habituelle de parcimonie. Mais cette définition s'étend aux estimateurs tels que le SVM, où $g_j = \iota_{[0,C]}$ et le support généralisé correspond aux vecteurs de support. Notre algorithme de *working set* exploite cette structure afin de converger plus rapidement.

3 Algorithme proposé et propriétés

L'algorithme proposé combine deux idées :

- Une stratégie de working set, capable de traiter une grande classe de pénalités convexes et non convexes.
- Une accélération d'Anderson pour la descente par coordonnées pour le problème interne.

Pour éviter de gaspiller des calculs sur des variables en dehors du support généralisé, les algorithmes de working set sélectionnent itérativement un sous-ensemble de variables/coordonnées jugées importantes (le *working set*), et résolvent (1) en se limitant à celles-ci. La question clé est donc la notion de variables *importantes*. En partant de la Définition 3, nous classons les variables selon leur violation de la condition d'optimalité :

$$\text{score}^{\partial}_j = \text{dist}(-\nabla_j f(\beta), \partial g_j(\beta)) \ . \tag{2}$$

Pour contrôler la croissance du working set, nous classons les caractéristiques par score (2), et avec $n_k = \max(n_{k-1}, 2\,|\,\mathrm{gsupp}(\beta^{(t)})|)$ prendre les n_k les plus grands d'entre eux dans le working set, tout en conservant les variables actuellement dans le working set. Cette croissance permet d'atteindre rapidement la taille inconnue du support généralisé tout en évitant une surestimation.

Proposition 5 *Soit $\mathcal{W}_t$ le t-ième ensemble de travail. Supposons que pour tout $t \geq 0$, $\mathcal{W}_t \subset \mathcal{W}_{t+1}$, alors la séquence $(\beta^{(t)})_{t\geq 0}$ générée par les itérés de Algorithme 1 converge vers un point critique de (1).*

Le deuxième ingrédient clé de notre algorithme consiste à utiliser un algorithme très rapide pour le problème interne: la descente par coordonnées avec accélération d'Anderson. Nous montrons d'abord que la descente de coordonnées identifie le support en temps fini pour une grande classe de problèmes non convexes (Proposition 9), ce qui entraîne un taux de convergence linéaire local et une accélération (Proposition 12). Dans cette section, nous étudions notre solveur interne (Algorithme 2) ; par commodité, nous écrivons toujours β et X pour leurs versions restreintes au working set.

Hypothèse 6 (α-semi-convexité) *Pour tout $j \in [p]$, g_j/L_j est α-semi-convexe, i.e. $g_j/L_j + \alpha\|\cdot\|^2/2$ est convexe, avec $\alpha < 1$.*

En particulier, l'hypothèse 6 est vérifiée pour le MCP et le SCAD.

Hypothèse 7 (Existence) *Le problème (1) admet au moins un point critique.*

Dans le cas de la Proposition 9, on suppose la convergence de Algorithme 2 vers un point critique $\hat{\beta}$, et les hypothèses suivantes sont faites sur ce point critique.

Hypothèse 8 (Non dégénérescence) *Le point critique considéré $\hat{\beta} \in \mathbb{R}^p$ est non dégénéré : pour tout $j \notin \mathrm{gsupp}(\hat{\beta})$, $-\nabla f_j(\hat{\beta}) \in \mathrm{int}(\partial g_j(\hat{\beta}_j))$.*

3.1 Identification du modèle en un nombre fini d'itérations

Avec les hypothèses précédentes, nous montrons que la descente par coordonnées identifie le support généralisé de sa limite pour cette classe de problèmes non convexes.

Proposition 9 (Identification de modèle de CD) *Supposons que*

1. *Les Hypothèses 1, 2, 6 et 7 sont satisfaites.*
2. *La séquence $(\beta^{(k)})_{k\geq 0}$ générée par la descente par coordonnées (Algorithme 2 sans extrapolation) converge vers un point critique $\hat{\beta}$.*
3. *$\hat{\beta}$ satisfait l'Hypothèse 8.*

Alors, l'Algorithme 2 (sans extrapolation) identifie le modèle en un nombre fini d'itérations : il existe $K > 0$ tel que pour tout $k \geq K$, $\beta^{(k)}_{\mathcal{S}^c} = \hat{\beta}_{\mathcal{S}^c}$.

En d'autres termes, pour k suffisamment grand, $\beta^{(k)}$ partage le support généralisé de $\hat{\beta}$.

3.2 Taux de convergence local accéléré

Si en plus f et g sont localement régulières sur le support généralisé au point critique considéré, notre algorithme bénéficie d'une convergence linéaire locale.

Hypothèse 10 (Localement $\mathcal{C}^3$) *Pour tout $j \in \mathcal{S} \triangleq \operatorname{gsupp}(\hat{\beta})$, g_j est localement $\mathcal{C}^3$ autour de $\hat{\beta}_j$, et f est localement $\mathcal{C}^3$ autour de $\hat{\beta}$.*

Hypothèse 11 (Convexité forte locale) *La Hessienne au point critique considéré $\hat{\beta}$, restreinte à son support généralisé $\mathcal{S}$, est définie positive, i.e. $\nabla^2_{\mathcal{S},\mathcal{S}} f(\hat{\beta}) + \nabla^2_{\mathcal{S},\mathcal{S}} g(\hat{\beta}) \succ 0_{\mathcal{S}}$.*

Proposition 12 (Taux de convergence local) *Considérons un point critique $\hat{\beta}$ et supposons que*

1. *Les Hypothèses 1, 2 et 7 tiennent.*
2. *Les fonctions f et g_j, $j \in [p]$ sont quadratiques par morceaux (ce qui est le cas pour la régression MCP).*
3. *La séquence $(\beta^{(k)})_{k\geq 0}$ générée par la descente accélérée des coordonnées d'Anderson avec des mises à jour de 1 à p et de p à 1 (Algorithme 2 avec extrapolation) converge vers un point critique $\hat{\beta}$.*
4. *$\hat{\beta}$ satisfait les Hypothèses 8, 10 et 11.*

Alors il existe $K \in \mathbb{N}$, et une fonction $\mathcal{C}^1$. $\psi : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$. telle que, pour tout $k \in \mathbb{N}, k \geq K$: $\beta_j^{(k)} = \hat{\beta}_j$, for all $j \in \mathcal{S}^c$, Soit $T \triangleq \mathcal{J}\psi(\hat{\beta})$, $H \triangleq \nabla^2_{\mathcal{S},\mathcal{S}} f(\hat{\beta}) + \nabla^2_{\mathcal{S},\mathcal{S}} g(\hat{\beta})$, $\zeta \triangleq (1 - \sqrt{1-\rho(T)})/(1 + \sqrt{1-\rho(T)})$ et $B \triangleq (T - \mathrm{Id})^\top (T - \mathrm{Id})$. Alors $\rho(T) < 1$ et les itérés de l'extrapolation d'Anderson bénéficient d'un taux de convergence local accéléré : $\|\beta_{\mathcal{S}}^{(k-K)} - \hat{\beta}_{\mathcal{S}}\|_B \leq \left(\sqrt{\kappa(H)} \frac{2\zeta^{M-1}}{1+\zeta^{2(M-1)}}\right)^{(k-K)/M} \|\beta_{\mathcal{S}}^{(K)} - \hat{\beta}_{\mathcal{S}}\|_B$.

4 Validation expérimentale

Sur la Figure 1 nous illustrons sur des données simulées[a] la flexibilité de notre algorithme, utilisé pour calculer le chemin de régularisation des pénalités Lasso, MCP $\ell_{0.5}$ et $\ell_{2/3}$. Sur la Figure 2, nous illustrons les performances en temps sur le Lasso, montrant que `skglm` obtient une performance au niveau de l'état de l'art.

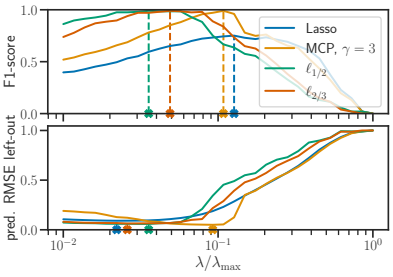


Figure 1

[a] https://contrib.scikit-learn.org/skglm/auto_examples/plot_sparse_recovery.html#sphx-glr-auto-examples-plot-sparse-recovery-py

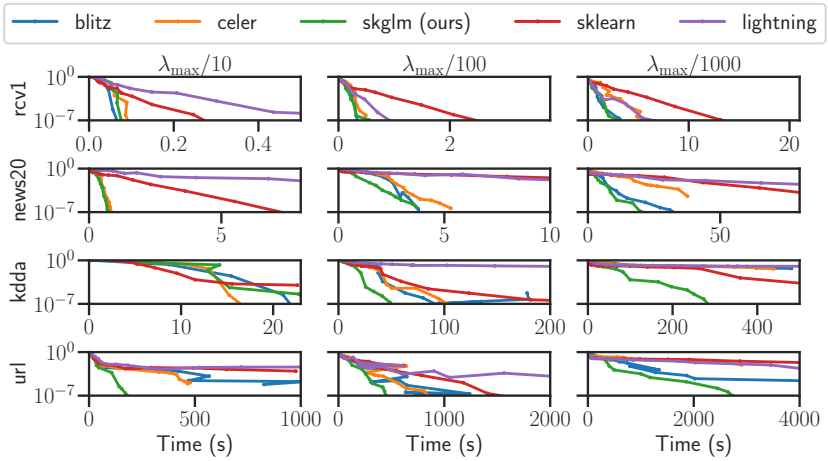


Figure 2

Références

Anderson, D. G. (1965). Iterative procedures for nonlinear integral equations. *Journal of the ACM*, 12(4):547–560.

Bertrand, Q. et Massias, M. (2021). Anderson acceleration of coordinate descent. *In AISTATS*.

Bolte, J., Sabach, S. et Teboulle, M. (2014). Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1):459–494.

Bonnefoy, A., Emiya, V., Ralaivola, L. et Gribonval, R. (2015). Dynamic screening: accelerating first-order algorithms for the Lasso and Group-Lasso. *IEEE Trans. Signal Process.*, 63(19):20.

Breheny, P. et Huang, J. (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *The annals of applied statistics*, 5(1):232.

El Ghaoui, L., Viallon, V. et Rabbani, T. (2010). Safe feature elimination for the lasso and sparse supervised learning problems. *arXiv preprint arXiv:1009.4219*.

Fercoq, O., Gramfort, A. et Salmon, J. (2015). Mind the duality gap: safer rules for the lasso. *In ICML*, pages 333–342. PMLR.

Fercoq, O. et Richtárik, P. (2015). Accelerated, parallel, and proximal coordinate descent. *SIAM Journal on Optimization*, 25(4):1997–2023.

Massias, M., Gramfort, A. et Salmon, J. (2018). Celer: a fast solver for the lasso with dual extrapolation.

Rakotomamonjy, A., Flamary, R., Gasso, G. et Salmon, J. (2022). Provably convergent working set algorithm for non-convex regularized regression. *In AISTATS*.

Soubies, E., Blanc-Féraud, L. et Aubert, G. (2015). A continuous exact ℓ_0 penalty (cel0) for least squares regularized problem. *SIAM Journal on Imaging Sciences*, 8(3):1607–1639.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 58(1):267–288.

Tseng, P. et Yun, S. (2009). A coordinate gradient descent method for nonsmooth separable minimization. *Mathematical Programming*, 117(1):387–423.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, 38(2):894–942.

ISDE : estimation de densité multidimensionnelle sous hypothèse de structure d'indépendance

Louis Pujol [1]

[1] *Université Paris-Saclay, CNRS, Laboratoire de mathématiques d'Orsay, 91405, Orsay.*
louis.pujol@universite-paris-saclay.fr

Résumé. Nous nous intéressons au problème de l'estimation de densité d'un point de vue pratique et dans un contexte multidimensionnel. On note N la taille de l'échantillon. Pour une classe de régularité de type Hölder sur des densités d-dimensionnelles, $\mathcal{H}_d^\beta(L)$, le risque minimax en perte L_2 au carré est asymptotiquement équivalent à $N^{-\frac{2\beta}{2\beta+d}}$. Cette vitesse de convergence se déteriore lorsque d augmente. Pour y remédier, on peut ajouter l'hypothèse de structure d'indépendance , consistant à supposer que la densité inconnue peut se décomposer comme un produit de certaines de ses marginales de dimension au plus $k < d$ et impliquant des variables disjointes. Dans ce cas la vitesse minimax est $N^{-\frac{2\beta}{2\beta+k}}$.

Nous présentons ISDE (Independence Structure Density Estimation), un algorithme permettant d'estimer simultanément une partition des variables et un estimateur de densité comme produit de marginales s'appuyant sur cette partition en maximisant un critère de log-vraisemblance. Nous prouvons l'efficacité de la méthode sur données simulées et son intérêt sur des données réelles issues d'expériences de cytométrie de masse.

Mots-clés. Estimation de Densité, Réduction de Dimension, Structure d'Indépendance

Abstract. We are focusing on practical density estimation in multidimensional setting. Let N denotes the number of observations. For an Hölder regularity class on d-dimensional densities $\mathcal{H}_d^\beta(L)$, the minimax speed under squared L_2 loss is asymptotically equivalent to $N^{-\frac{2\beta}{2\beta+d}}$. This convergence speed deteriorates as d increases. It is possible to limit the impact of the dimension by introducing the independence structure hypothesis, supposing the unknown density can be decomposed as a product of marginal densities over a partition of the features. In this context the minimax speed becomes $N^{-\frac{2\beta}{2\beta+k}}$.

We present ISDE (Independence Structure Density Estimation), an algorithm providing simultaneously a partition of the features and a density estimator written as a product of marginals over this partition maximizing a log-likelihood criterion. We prove the practical efficiency of it on synthetic data and show its interest on real datasets from mass cytometry experiments.

Keywords. Density Estimation, Dimensionality Reduction, Independence Structure

1 Estimation de densité multidimensionnelle

Soit $X_1, \ldots, X_N$ un ensemble de points de $\mathbb{R}^d$, nous modélisons le phénomène menant à la génération de ces points comme N réalisations indépendantes et identiquement distribuées d'une même variable aléatoire de densité f. L'objectif est de produire un estimateur $\hat{f}$ proche de f.

D'un point de vue conceptuel, la difficulté de ce problème peut être quantifiée par le risque minimax. Ce risque est calculé pour un modèle Σ et une perte ℓ. Le modèle représente une classe de fonctions dont nous supposons que f est membre et $\ell(\hat{f}, f)$ une quantité positive représentant l'écart entre la densité f et l'estimateur $\hat{f}$. Un exemple étudié classiquement est un modèle de régularité de type Hölder pour que l'on notera $\mathcal{H}_d^\beta(L)$ où β représente la constante de Hölder et L le rayon de la boule considérée avec la perte L_2 au carré :

$$\ell(\hat{f}, f) = \int \left(f - \hat{f}\right)^2. \tag{1}$$

Dans ce cas le risque minimax dépend asymptotiquement du nombre d'observations de la manière suivante (Hasminskii et Ibragimov (1990)):

$$\mathcal{R}_{\|.\|_2^2}\left[\mathcal{H}_d^\beta(L)\right] \asymp N^{-\frac{2\beta}{2\beta+d}}. \tag{2}$$

On note que ce risque tend vers 0 lorsque le nombre d'observations augmente et que plus le modèle contient des fonctions régulières (donc que β est plus élevé), plus la vitesse de convergence est rapide.

En revanche, lorsque la dimension ambiante d augmente, cette vitesse de convergence se détériore, ce phénomène est une manifestation de la malédiction de la dimension. Pour y remédier, une solution est de construire des modèles imposants des hypothèses structurelles sur la densité inconnue et permettant de limiter l'impact de la dimension sur la vitesse de convergence. Un exemple d'hypothèse de ce type est donné par le modèle de structure d'indépendance.

2 Le modèle de structure d'indépendance

Introduit par Oleg Lepski (Lepski (2013)) et étudié par Gilles Rebelles (Rebelles (2015)), le modèle de structure d'indépendance consiste à supposer que la densité f s'écrit comme un produit de densité marginales (la liste des sous-ensembles de variables sur lesquels sont estimées ces densités marginales forment une partition de $\{1, \ldots, d\}$)

Soit k un entier inférieur à d. On note Set_d^k l'ensemble des sous-ensembles de $\{1, \ldots, d\}$ de taille au plus k. On note Part_d^k l'ensemble des partitions de $\{1, \ldots, d\}$ que l'on peut construire en utilisant des ensembles contenus dans Set_d^k. On définit alors $\mathcal{D}_d^k$, l'ensemble des densité sur $\mathbb{R}^d$ pouvant se décomposer comme un produit de densités marginales selon

une partition dont les blocs sont de taille au plus k :

$$\mathcal{D}_d^k = \left\{ f : \mathbb{R}^d \mapsto \mathbb{R} \mid \exists \mathcal{P} \in \mathrm{Part}_d^k : f = \prod_{S \in \mathcal{P}} f_S \right\} \tag{3}$$

où f_S désigne la densité marginale de f sur l'ensemble de variables S. Dans Rebelles (2015), il a été montré que dans ce cas, le risque minimax était lié à k et non plus à la dimension ambiante d:

$$\mathcal{R}_{\|.\|_2^2} \left[N_{p,d}(\beta, L) \cap \mathcal{D}_d^k \right] \asymp N^{-\frac{2\beta}{2\beta+k}} \tag{4}$$

Malheureusement, l'estimateur introduit pour atteindre cette vitesse minimax nécessite la résolution d'un problème d'optimisation qui ne peut pas facilement être implémenté. Nous proposons une reformulation du problème d'estimation de densité sous l'hypothèse de structure d'indépendance permettant de construire un algorithme permettant de calculer une structure d'indépendance optimale et une estimation de densité associée sur un jeu de données de taille raisonnable.

3 ISDE : Estimation de Densité par Structure d'Indépendance

La clé pour construire un estimateur de densité s'appuyant sur une estimation de partition des variables réside dans la capacité à découpler le problème d'estimation des densités marginales et de la sélection de la partition optimale. En effet il apparaît que le nombre de partitions dans Part_d^k est bien supérieur au nombre d'ensemble de variables dans Set_d^k. Le problème étudié par Lepski et Rebelles était un problème d'optimisation en norme L_2, cette formulation ne fait pas apparaître la possibilité de découpler ainsi le problème d'estimation. Pour dépasser cette difficulté, nous nous sommes placés dans le cadre de la minimisation de la perte de Kullback-Leibler (KL) entre f et $\hat{f}$:

$$\mathrm{KL}\left(f \| \hat{f}\right) = \int \log\left(\frac{f}{\hat{f}}\right) f. \tag{5}$$

Le problème empirique associé est le problème de maximisation de la log vraisemblance empirique. Nous pouvons alors tirer avantage du fait que nous cherchons des densités qui s'écrivent comme des produits de densités marginales et du fait que le log transforme un produit en somme pour réécrire le problème d'optimisation de la manière suivante :

$$\max_{\mathcal{P} \in \mathrm{Part}_d^k} \sum_{S \in \mathcal{P}} \left\{ \max_{\theta_S \in \Theta} P\left[\log(\hat{f}_S^{\theta(S)}) \right] \right\} \tag{6}$$

où Θ représente un espace de paramètres correspondant aux estimations des densités marginales. Nous faisons ainsi apparaître le découplage entre estimation des densitées marginales et sélection de la partition optimale.

En pratique, nous séparons le jeu de données $X_1, \ldots, X_N$ en deux sous-échantillons disjoints : $W = (W_1, \ldots, W_m)$ et $Z = (Z_1, \ldots, Z_n)$. W est utilisé pour construire une collection d'estimateurs marginaux $\left(\hat{f}_S\right)_{(S \in Sdk)}$ et Z pour estimer la log-vraisemblance pour chacun de ses estimateurs $(\ell_n(S))_{(S \in Sdk)}$. Il reste ensuite à résoudre le problème :

$$\max_{\mathcal{P} \in \mathrm{Part}_d^k} \sum_{S \in \mathcal{P}} \ell_n(S). \tag{7}$$

Ce problème peut se réécrire comme un problème d'optimisation binaire sous contrainte et peut être résolu en pratique via l'utilisation de librairies dédiées. Nous utilisons la librairie pyhton PulP (Mitchell, OSullivan, et Dunning (2011)). Nous appelons l'algorithme ainsi défini ISDE (independance structure density estimation) (Pujol (2021)). Notons qu'il peut être utilisé avec n'importe quel estimateur pour construire $\left(\hat{f}_S\right)_{(S \in Sdk)}$, ce qui le rend utilisable dans une large variété de contextes.

4 Résultats empiriques

La formulation du problème d'estimation de densité en perte de Kullback-Leibler nous a permis de proposer une version implémentable de l'estimation de densité sous hypothèse de structure d'indépendance. Cherchons à valider empiriquement la pertinence de l'algorithme proposé, sur données simulées et sur données réelles.

4.1 Données simulées

Etant donnée une liste de nombres entiers strictement positifs $S = (s_1, \ldots, s_M)$, et un entier $N > 0$ (nous fixons $N = 5000$ dans nos expériences), on définit une variable aléatoire sur $\mathbb{R}^d$ où $d = s_1 + \cdots + s_M$ pour laquelle les groupes de variables consécutifs de taille $s_1, \ldots, s_M$ sont indépendants.

On observe qu'en utilisant comme estimateur de densité marginales un estimateur à noyau gaussien avec largeur de bande isotrope, dont la largeur de bande est sélectionnée par validation croisée, on est capable de retrouver la vraie structure d'indépendence (dans 100% des expériences présentées ici) et on observe un gain en log-vraisemblance par rapport à l'estimateur à noyau ne prenant pas en compte la structure d'indépendance comme illustré par la Table 1.

4.2 Données réelles

Pour illustrer la pertinence de notre méthode sur des données réelles nous nous sommes intéressés à des données publiques de cytométrie de masse. Les deux jeux de données que nous présentons sont de dimension 13 et 32 et sont issues de mesures réalisées à l'échelle

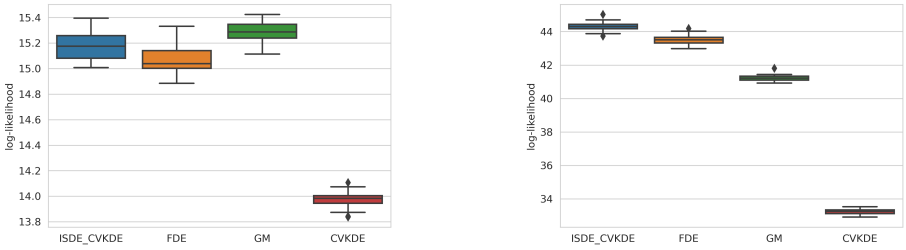
Structure	[2, 2, 1]	[3, 3, 3]	[4, 4, 2, 2]
ISDE avec KDE	$\mathbf{-1.79 \pm 0.09}$	$\mathbf{-3.918 \pm 0.139}$	$\mathbf{-6.44 \pm 0.16}$
KDE	-0.56 ± 0.02	-3.39 ± 0.11	-4.06 ± 0.12

Table 1: Log-vraisemblance négative empirique (avec écart-type correspondant à 10 tirages de l'échantillon de validation de taille 2000)

unicellulaire à partir d'extractions de moelle osseuse sur des patients sains (Levine et al. (2015)). Les variables représentent l'intensité de présence de certaines protéines à la surface de cellule ainsi que des mesures des propriétés optiques des cellules liées à leur taille et à leur configuration.

Nous extrayons de ces jeux de données un échantillon d'entraînement de taille $N = 5000$ sur lequel nous entraînons différents algorithmes d'estimation de densité : ISDE en utilisant comme estimateur de densité marginales à noyau gaussien dont la largeur de bande est sélectionnée par validation croisée (ISDE_CVKDE), un estimateur à noyau gaussien usuel avec largeur de bande isotrope dont le paramètre de largeur de bande est sélectionné par validation croisée (CVKDE), un estimateur basé sur un modèle graphique de type forêt et sur des estimateurs marginaux à noyau gaussien (Liu et al. (2011)) (FDE) et un estimateur paramétrique basé sur un mélange de gaussiennes dont le nombre de gaussiennes est sélectionné via une validation croisée (GM).

Nous évaluons ensuite chacune des quatre méthode en calculant la log-vraisemblance empirique sur 20 tirage de 2000 événements parmi ceux qui n'ont pas été utilisés pour l'entraînement. les résultats sont données par la Figure 1. Pour conserver des temps de calculs raisonnables, on choisit $k = 5$ pour le jeu de données Levine13 et $k = 3$ pour le jeu de données Levine32.



(a) Levine13 (b) Levine32

Figure 1: Comparaison de la log-vraisemblance sur données test pour différentes méthodes d'estimation de densité multidimensionnelle

Ces résultats valident l'intérêt de l'utilisation de ISDE dans le but de fournir une estimation de densité. En effet ses performances sont toujours comparables aux autres méthodes et c'est ISDE qui donne le meilleur score en dimension 32.

L'autre avantage de ISDE est de fournir une information qualitative sur les données via la structure d'indépendance.

Bibliographie

Hasminskii, R., Ibragimov, I. (1990). On density estimation in the view of Kolmogorov's ideas in approximation theory. The Annals of Statistics, 999-1010.

Lepski, O. (2013). Multivariate density estimation under sup-norm loss: oracle approach, adaptation and independence structure. The Annals of Statistics, 41(2), 1005-1034.

Levine, J. H., Simonds, E. F., Bendall, S. C., Davis, K. L., El-ad, D. A., Tadmor, M. D., ... Nolan, G. P. (2015). Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. Cell, 162(1), 184-197.

Liu, H., Xu, M., Gu, H., Gupta, A., Lafferty, J., Wasserman, L. (2011). Forest density estimation. The Journal of Machine Learning Research, 12, 907-951.

Mitchell, S., OSullivan, M., Dunning, I. (2011). PuLP: a linear programming toolkit for python. The University of Auckland, Auckland, New Zealand, 65.

Pujol, L. (2021). ISDE: Independence Structure Density Estimation.

Rebelles, G. (2015). $\mathbb{L}_p$ adaptive estimation of an anisotropic density under independence hypothesis. Electronic journal of statistics, 9(1), 106-134.

Statistique mathématique 2

Comparaison d'arbres CART par sous-échantillonnage sans remplacement

Felix Cheysson [1]

[1] *LPSM, Sorbonne Université, felix.cheysson@upmc.fr*

Résumé. Les arbres CART sont des méthodes non paramétriques intéressantes pour les problèmes de classification et de régression, car ils peuvent souvent modéliser des relations complexes entre covariables et résultats sans connaissances préalables, et le processus de décision s'apparente à la façon dont les humains prennent des décisions. Cependant, ces méthodes sont très sensibles à l'ensemble d'apprentissage, ce qui peut poser des problèmes lors de la comparaison des arbres construits à partir de deux échantillons différents. Nous proposons un test pour la comparaison d'arbres CART, reposant sur la théorie des U-statistiques, où le seuil critique du test d'hypothèse est estimé par sous-échantillonnage sans remplacement. Nous prouvons notamment un théorème central limite pour cet estimateur. Une courte étude de simulation et une application aux données Covid-19 illustrent les performances de la méthode proposée.

Mots-clés. Arbres CART, U-statistiques, Estimation de quantile, Test d'hypothèse, Méthodes d'ensemble

Abstract. CART trees are appealing non parametric methods for classification and regression problems, as they can often model complex covariate-outcome relationships without prior knowledge, and the decision process is akin to how humans make decisions. However, they are highly sensitive to the learning set, and established theoretical properties for CART are few and far between, which may cause some issues when trying to compare the trees built from two different samples. We propose a test for the comparison of CART trees, based on a U-statistic framework, where the critical threshold of the hypothesis test is estimated via subsampling without replacement. Notably, we prove a central limit theorem for this estimator. A short simulation study and an application to Covid-19 data illustrate the performances of the proposed method.

Keywords. CART trees, U-statistics, Quantile estimation, Hypothesis test, Ensemble methods

1 Long summary

We consider binary trees constructed from the CART (*Classification And Regression Tree*) algorithm, for which some background can be found in the seminal work of Breiman et al. (1984). Briefly, CART generates a binary tree by recursively splitting the input space

along one of its dimension, where at each step the node to split, the dimension along which to split, and the position of the split are chosen optimally such that the decrease in a criterion function — an impurity criterion for classification, the mean square error for regression — from the parent node to its two children nodes is maximised. The algorithm stops when a stopping rule is achieved, usually when the number of observations in each furthest node is below a given bound or when no decrease in the criterion function greater than a threshold exists. Then the terminal nodes — the leaves of the tree — form a partition of the input space, and are labelled through averaging (or majority vote, *i.e.* according to the most frequent class within, in the case of classification).

However, established theoretical properties for CART trees are few and far between, with Breiman et al. (1984) proving an initial consistency result for tree structured regression and classification. CART trees are also plagued with high sensitivity to the learning set, which may cause issues for prediction. Thus random forests and more generally ensemble methods, constructed by averaging results from a large number of classifiers or regression learners, are often preferred as they thrive on this instability. Consequently, the last decade has seen many articles establishing asymptotic and non-asymptotic results for ensemble methods (Scornet et al., 2015; Biau and Scornet, 2016; Lopes et al., 2020).

Recently, Mentch and Hooker (2016) then Peng et al. (2019) have developed a framework whereby ensemble methods based on subsampling can be viewed as U-statistics, which allowed them to prove central limit theorems and Berry-Esseen results for random forests. Building on this framework, we propose a hypothesis test for the comparison of CART trees. To derive the critical region of the test, we extend the U-quantile-statistic proposed by Mayer (2009), which provides an estimator for the $1-\alpha$-th quantile of the test statistic under the null, to the case of infinite-order U-statistics (when the rank of the kernel is allowed to grow with the sample size).

While we originally proposed this procedure for CART trees, it can be painlessly extended to most ensemble methods, namely random forests, as in Peng et al. (2019). It also provides a computational way to estimate quantiles for the predictions of CART trees and ensemble methods, backed up by a central limit theorem. Furthermore, note that while this procedure was motivated by the comparison of CART trees, it is actually a two-sample test for the conditional expectation, when the latter is estimated via CART trees (or ensemble methods).

A short simulation study, focusing on the growth rate of the different hyperparameters associated with the test (rank of the kernel, subsampling size) will accompany the presentation, as well as an application to Covid-19 data, evaluating the evolution of the impact of certain comorbidities on the death rates through the different waves of the pandemic.

1.1 Notations

Let $\{X_i\}_{1\leq i\leq m}$ and $\{Y_j\}_{1\leq j\leq n}$ denote two independent samples of i.i.d. variables, from possibly different distribution functions (d.f.) P_X and P_Y, where the X_i and Y_j can be

written as doublets $(u, v) \in \mathcal{U} \times \mathcal{V}$, with u representing the explanatory variables to be used as input to the regression (or classification) problem and v its outcome. Let $T_X = T(X_1, \ldots, X_m)$ denote the tree built from the first sample, and similarly $T_Y = T(Y_1, \ldots, Y_n)$. Finally, for an input $u \in \mathcal{U}$ and a tree T, write $T(u) \in \mathcal{V}$ its prediction at u.

1.2 Comparison of CART trees

We can evaluate the (dis)similarity between T_X and T_Y at a collection of test points $(u_1, \ldots, u_t)$ via the kernel defined by

$$h(X; Y) = \sum_{i=1}^{t} d\big(T_X(u_i), T_Y(u_i)\big),$$

where d denotes a distance function. As CART trees are $\mathbb{L}^2$-consistent, a priori good choices of d are the absolute difference, and the squared difference.

Then we can establish whether T_X is similar to T_Y by conducting a test of the following hypotheses

$$\mathcal{H}_0 : \forall 1 \leq i \leq t,\ \mathbb{E}_{P_X}[V \mid U = u_i] = \mathbb{E}_{P_Y}[V \mid U = u_i]$$
$$\mathcal{H}_1 : \exists 1 \leq i \leq t,\ \mathbb{E}_{P_X}[V \mid U = u_i] \neq \mathbb{E}_{P_Y}[V \mid U = u_i]$$

where $(U, V) \in \mathcal{U} \times \mathcal{V}$ denotes a typical observation.

1.3 Connection with infinite-order U-statistics

Let $0 < r \leq m$ and $0 < s \leq n$ two integers, and note that $h(X_1, \ldots, X_r; Y_1, \ldots, Y_s)$ is a kernel that is permutation symmetric in its first r arguments and its last s arguments separately. Then it stands that the two-sample U-statistic with kernel h has the form

$$U_{m,n,r,s} = \binom{m}{r}^{-1} \binom{n}{s}^{-1} \sum_{(m,r)} \sum_{(n,s)} h(X_{i_1}, \ldots, X_{i_r}; Y_{j_1}, \ldots, Y_{j_s}),$$

where (m, r) and (n, s) are shorthands for the binomial coefficients $\binom{m}{r}$ and $\binom{n}{s}$, and the sums range over the collections of all distinct subsets of r elements from $\{1, \ldots, m\}$ and s elements from $\{1, \ldots, n\}$ respectively. $U_{m,n,r,s}$ is an unbiased estimator of the parameter

$$\theta_{rs} = \mathbb{E}h(X_{i_1}, \ldots, X_{i_r}; Y_{j_1}, \ldots, Y_{j_s}).$$

The sequence $U_{m,n,r,s}$ will be shown to be asymptotically normal when $m, n, r, s \to \infty$ in such a way that $m/(m+n) \to \lambda \in (0, 1)$, $n/(m+n) \to 1 - \lambda$ and $r/m \sim s/n$, *i.e.* that the numbers of X_i and Y_j are of the same order, and that the numbers of X_i and Y_j

arguments for the kernel h grow with the same order. To further simplify the notations, denote $n_0 = m + n$ and s_0 the integer such that $r/m \sim s_0/n_0$. We can summarise these relations in the following assumption

Assumption 1. *As m and n grow to infinity, we have*

$$\begin{aligned} &m + n = n_0 && r/m \sim s/n \sim s_0/n_0 \\ &m = \lambda n_0 + o(n_0) && r = \lambda s_0 + o(s_0) \\ &n = (1-\lambda) n_0 + o(n_0) && s = (1-\lambda) s_0 + o(s_0) \end{aligned}$$

Since all indices grow to infinity with respect to n_0, we will simplify most notations by dropping all references to m, n, r and s and replacing them with a single reference to n_0 instead: for example, the U-statistic $U_{m,n,r,s}$ will be denoted U_{n_0}.

1.4 Incomplete U-statistics

In practice, there is a high computational cost associated with computing the value $h(X_{i_1}, \dots, X_{i_r}; Y_{j_1}, \dots, Y_{j_s})$ for each possible subsamples of r elements of $\{X_1, \dots, X_m\}$ and s elements of $\{Y_1, \dots, Y_s\}$. It is therefore natural to only consider a smaller ensemble of these subsamples, of size N, and the associated *incomplete* U-statistic given by

$$U_{n_0,N} = N^{-1} \sum h(X_{i_1}, \dots, X_{i_r}; Y_{j_1}, \dots, Y_{j_s}),$$

where the sum is computed over N subsamples $(X_{i_1}, \dots, X_{i_r}; Y_{j_1}, \dots, Y_{j_s})$ chosen uniformly with replacement. This statistic can also be rewritten more rigorously using an intermediary term ρ_{ij} which denotes how many times each subsample is selected, such that

$$U_{n_0,N} = N^{-1} \sum_{(m,r)} \sum_{(n,s)} \rho_{ij}\, h(X_{i_1}, \dots, X_{i_r}; Y_{j_1}, \dots, Y_{j_s}),$$

where the tuple (ρ_{ij}) is multinomial with N trials with equal probability $\xi = (m,r)^{-1}(n,s)^{-1}$.

To prove distributional results for the incomplete two-sample U-statistic U_{n_0}, we employ the same method as in Peng et al. (2019), which is first applying the Hoeffding decomposition on the complete U-statistic, then relating the incomplete U-statistic to the complete one via their characteristic functions. This yields the following theorem

Theorem 1. *Consider that $m, n, r, s \to \infty$ as in Assumption 1. Further assume that $\mathbb{E}h^2 < \infty$, $\mathbb{E}h^4 < \infty$ and $\mathbb{E}h^6 < \infty$ uniformly with respect to s_0, and that the following condition is verified*

$$\lim_{n_0 \to \infty} \frac{s_0}{n_0} \frac{\zeta_{rs}}{s_0(\lambda\zeta_{10} + (1-\lambda)\zeta_{01})} = 0,$$

where

$$\begin{aligned}
\zeta_{rs} &= \mathrm{Var}\big(h(X_1,\ldots,X_r;Y_1,\ldots,Y_s)\big)\\
\zeta_{10} &= \mathbb{E}\big[h(X_1,X_2,\ldots,X_r;Y_1,\ldots,Y_s)h(X_1,X_2',\ldots,X_r';Y_1',\ldots,Y_s')\big]-\theta_{rs}\\
\zeta_{01} &= \mathbb{E}\big[h(X_1,\ldots,X_r;Y_1,Y_2,\ldots,Y_s)h(X_1',\ldots,X_r';Y_1,Y_2',\ldots,Y_s')\big]-\theta_{rs}.
\end{aligned}$$

Then, as $n_0 \to \infty$*,*

$$\frac{U_{n_0,N}-\theta_{n_0}}{\sqrt{s_0^2(\lambda\zeta_{10}+(1-\lambda)\zeta_{01})/n_0+\zeta_{rs}/N}} \longrightarrow \mathcal{N}(0,1).$$

1.5 Critical region and U-quantile-statistics

Rather than the expected value of the kernel $h(X;Y)$, we are interested in its p-th quantile under $\mathcal{H}_0$. It turns out that, in the context of complete and finite-order one-sample U-statistics, the sample p-th quantile of the set

$$\{h(X_{i_1},\ldots,X_{i_r}) : 1 \leqslant i_1 < \ldots < i_r \leqslant m\}$$

inherits some of the asymptotic properties of the associated U-statistic (Mayer, 2009)

$$U = \binom{m}{r}^{-1} \sum_{(m,r)} h(X_{i_1},\ldots,X_{i_r}).$$

We investigate the asymptotic behaviour of the U-quantile-statistic in the incomplete, infinite order and two-sample cases. Denote F_{n_0} the cumulative d.f. (c.d.f.) of the random variable $h_{n_0}(X_1,\ldots,X_r;Y_1,\ldots,Y_s)$ and H_{pn_0} its p-th order quantile, defined by

$$H_{pn_0} = \inf\{x : F_{n_0}(x) \geqslant p\}.$$

Similarly define the empirical c.d.f. $\hat{F}_{n_0}$ given by

$$\hat{F}_{n_0}(x) = N^{-1} \sum_{i=1}^{(m,r)} \sum_{j=1}^{(n,s)} \rho_{ij} \mathbb{1}\{h_{n_0}(X_{i_1},\ldots,X_{i_r};Y_{j_1},\ldots,Y_{j_s}) \leqslant x\},$$

and its sample p-th quantile

$$\hat{H}_{pn_0} = \inf\{x : \hat{F}_{n_0}(x) \geqslant p\},$$

and note that, for all x, $\hat{F}_{n_0}(x)$ is an incomplete infinite-order U-statistic with kernel $(X;Y) \mapsto \mathbb{1}\{h(X;Y) \leqslant x\}$. This yields the following central limit theorem for the U-quantile-statistic

Theorem 2. *Assume F_{n_0} converges uniformly to a c.d.f. F which is of class $\mathcal{C}^1$, that F_{n_0} is differentiable everywhere, and that F'_{n_0} converges uniformly to F'. Further assume that the empirical c.d.f. $\hat{F}_{n_0}(x)$ satisfies the conditions of Theorem 1 for all x in a neighbourhood of H_{pn_0}, and let $\sigma^2_{n_0}(H_{pn_0})$ denote the asymptotic variance of $\hat{F}_{n_0}(H_{pn_0})$. Then, as $n_0 \to \infty$,*

$$F'_{n_0}(H_{pn_0})\frac{\hat{H}_{pn_0} - H_{pn_0}}{\sigma_{n_0}(H_{pn_0})} \longrightarrow \mathcal{N}(0,1).$$

1.6 A simulation study and an application to Covid-19 data

A short simulation study (figures not included in this summary) shows that the proposed hypothesis test, whereby the $1-\alpha$-th quantile of the test statistic under $\mathcal{H}_0$ is estimated via the U-quantile-statistic, yields empirically satisfactory results for a large range of s_0 to n_0 and N to n_0 growth rates.

An application to death rates during the different waves of the pandemic will also be presented, where the flexibility of the test points $\{u_1, \ldots, u_t\}$ allows the comparison of different subpopulations, and to evaluate how certain comorbidities affected the death rates through the pandemic.

References

Biau, G. and Scornet, E. (2016). A random forest guided tour. *TEST*, 25(2):197–227, 1511.05741.

Breiman, L., Friedman, J., Olshen, R. A., and Stone, C. J. (1984). *Classification and regression trees*. The Wadsworth statistics / probability series. CRC.

Lopes, M. E., Wu, S., and Lee, T. C. M. (2020). Measuring the Algorithmic Convergence of Randomized Ensembles: The Regression Setting. *SIAM J. Math. Data Sci.*, 2(4):921–943, 1908.01251.

Mayer, M. (2009). U-Quantile-Statistics. arXiv:0906.1266.

Mentch, L. and Hooker, G. (2016). Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *J. Mach. Learn. Res.*, 17:1–41, 1404.6473.

Peng, W., Coleman, T., and Mentch, L. (2019). Asymptotic Distributions and Rates of Convergence for Random Forests via Generalized U-statistics. arXiv:1905.10651.

Scornet, E., Biau, G., and Vert, J. P. (2015). Consistency of random forests. *Ann. Stat.*, 43(4):1716–1741, 1405.2881.

Online multiple-testing with super-uniformity reward

Iqraa Meah [1], Etienne Roquain[2], and Sebastian Döhler[3]

[1]Laboratoire de Probabilités, Statistiques et Modélisations – Sorbonne Universite : UMR$_8$001, $CentreNationaldelaRechercheScientifique : UMR_8001, UniversitédeParis : UMR_8001 - -France$
[2]Laboratoire de Probabilités, Statistique et Modélisation – Sorbonne Universite : UMR$_8$001 $- -France$
[3]Hochschule Darmstadt, University of Applied Sciences – Allemagne

Résumé

Le test multiple en ligne fait référence au contexte où un nombre potentiellement infini d'hypothèses sont testées, et les p-valeurs sont disponibles une par une, de manière séquentielle. Cela diffère du contexte classique où le nombre d'hypothèses à tester $m < \infty$ $estconnuàl'avance, etles$p$ - valeurssonttoutesdisponiblesensemble.Lesméthodesenligneexistantespeuventsouffrird'unepertedepuissancesignifico valeurssontobtenuesàpartirdetestsdiscrets.Pourrésoudreceproblème, nousintroduisonsuneméthodede"récompense"bas valeurssousl'hypothèsenulle.Nousprouvonsquelesprocéduresrécompenséesgardentlecontrôledel'erreurdetypeItoutenp

Annulé

Test de linéarité dans un modèle de régression non paramétrique

Zaher Mohdeb [1] & Abdelkader Mokkadem [2]

[1] *Faculté de Génie des Procédés*
Université Salah Boubnider de Constantine 3
Constantine, Algérie
et
Laboratoire de Mathématiques et Sciences de la Décision,
Université frères Mentouri, Constantine, Algérie
E-mail: zaher.mohdeb@umc.edu.dz, z.mohdeb@gmail.com
[2] *Département de Mathématiques, Université de Versailles-St-Quentin,*
45, Avenue des Etats-Unis, 78035 Versailles Cedex, France
E-mail: abdelkader.mokkadem@uvsq.fr

Résumé. Une procédure de test d'hypothèse linéaire sur la fonction de régression f dans un modèle de régression non paramétrique est proposée. Plus précisément, on teste l'hypothèse que f est un élément de E, où E est un espace vectoriel de dimension finie. En supposant que les fonctions considérées sont höldériennes d'ordre plus grand que $1/2$ et on obtient le comportement asymptotique du test proposé, on a donc ainsi le niveau et la puissance asymptotique du test. Une étude par simulation a été menée, pour des petites tailles d'échantillon, afin de montrer la performance du test proposé.

Mots-clés. Hypothèse linéaire, régression non linéaire, régression non paramétrique, test non paramétrique.

Abstract. A procedure for testing linear hypothesis on the regression function f in a nonparametric regression model. More precisely, we test that f is an element of E, where E is a finite dimensional vector space. We assume that the functions satisfy the Hölder condition with order strictly greater than $1/2$, and we obtain the asymptotic weak behaviour of the proposed test, then we have the level and the asymptotic power of the test. A simulation study is conducted, for small sample size, to demonstrate the performance the proposed test.

Keywords. Linear hypothesis, nonlinear regression, nonparametric regression, nonparametric test.

1 Introduction

On considère le modèle de régression suivant

$$Y_{i,n} = f(t_{i,n}) + \varepsilon_{i,n}\,, \quad i = 1, \ldots, n\,, \tag{1}$$

où f est une fonction réelle inconnue, définie sur l'intervalle $[0,1]$ et $t_{1,n} = 0 < t_{2,n} < \cdots < t_{n,n} = 1$, est un échantillonnage fixé de l'intervalle $[0,1]$. Les erreurs $\varepsilon_{i,n}$ forment un tableau triangulaire de variables aléatoires d'espérance nulle et de variance finie σ^2, tel que pour tout n les variables aléatoires $\varepsilon_{1,n}, \ldots, \varepsilon_{n,n}$ soient indépendantes. Notre objectif est la construction de tests d'hypothèses linéaires sur la fonction de régression f. Plus précisément, soient $g_1, \ldots, g_p$ des fonctions définies sur $[0,1]$ et linéairement indépendantes et soit E_p l'espace vectoriel engendré par $g_1, \ldots, g_p$. On veut tester l'hypothèse:

$$\mathrm{H}_0 : \quad f \in E_p \qquad \text{contre} \qquad \mathrm{H}_1 : \quad f \notin E_p\,. \tag{2}$$

Le problème de test d'hypothèses dans le modèle (1) a donné lieu à de nombreux travaux, l'usage des coefficients de Fourier empiriques de f est abordé par Eubank et Spiegelman (1990) et Mohdeb et Mokkadem (2001). Dette et Munk (1998) introduit une statistique de test basée sur l'estimation du carré de la distance de f à E_p. Mohdeb et Mokkadem (2002) proposent une statistique de test basée sur une approche similaire à celle de Dette et Munk (1998), mais sans utilisation de poids et montrent qu'elle a le même comportement asymptotique. Mohdeb et Mokkadem (2004 a, 2004 b) proposent également une autre statistique de test basée sur une autre estimation du carré de la distance de f à E_p.
Dans ce travail, on procède comme dans Mohdeb et Mokkadem (2002), mais en utilisant une autre estimation de la variance des erreurs $(\varepsilon_{i,n})$ intervenant dans la statistique de test. Nous obtenons une loi asymptotique de celle-ci avec une variance différente de celle donnée dans Mohdeb et Mokkadem (2002). Des simulations ont été menées pour étudier la performance du test proposé.

2 Résultat principal

Soit h une fonction densité sur l'intervalle $[0,1]$, positive et de Hölder d'ordre $\gamma > 1/2$ et que l'échantillonnage $\{t_{1,n}, \ldots, t_{n,n}\}$ est associé à h. On note $L^2(d\mu)$ où $d\mu = h(t)dt$, l'espace des fonctions de carré intégrables, muni du produit scalaire usuel $\langle ., . \rangle$. La distance entre f et E_p est notée par $\mathcal{D}(f)$. On suppose que les hypothèses suivantes sont satisfaites.

(C1) $\displaystyle\max_{i=2,\ldots,n} \left| \int_{t_{i-1,n}}^{t_{i,n}} h(t)\,dt - \frac{1}{n} \right| = o\left(\frac{1}{n}\right)$;

(C2) h , f , g_k, , $k = 1, \ldots, p$, sont des fonctions de Hölder d'ordre $\gamma > 1/2$;

(C3) $\forall n$, $\varepsilon_{1,n}, \ldots, \varepsilon_{n,n}$ sont indépendantes et $\exists C \in \mathbb{R}^+$ tel que $E(\varepsilon_{i,n}^4) < C$, $\forall i, n$.

Soit $(\widehat{\beta}_1, \ldots, \widehat{\beta}_p)' = \underset{(b_1,\ldots,b_p)\in\mathbb{R}^p}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left| Y_{i,n} - \sum_{k=1}^{p} b_k g_k(t_{i,n}) \right|^2 \right\}$.

L'approche que nous proposons repose sur le faite que l'hypothèse:

$$\text{"}f \in E_p\text{"} \quad \text{est équivalente à} \quad \text{"}\phi = f - \sum_{k=1}^{p} \widehat{\beta}_k g_k \in E_p\text{"}.$$

Cela nous amène à utiliser la distance $\mathcal{D}(\phi)$ entre ϕ et E_p au lieu de $\mathcal{D}(f)$.

On suppose que $E_p \subset L^2(d\mu)$ et on veut estimer:

$$\mathcal{D}^2(\phi) = \min_{u\in E_p} \int_0^1 |\phi(t) - u(t)|^2 h(t) dt\,.$$

Cette distance peut aussi s'écrire sous la forme:

$$\mathcal{D}^2(\phi) = \frac{\Gamma(\phi, g_1, \ldots, g_p)}{\Gamma(g_1, \ldots, g_p)}\,, \tag{3}$$

où $\Gamma(u_1, \ldots, u_p)$ est le déterminant de Gram $|(\langle u_i, u_j)|_{i,j=1,\ldots,k}$ pour $u_1, \ldots, u_k$ dans $L^2(d\mu)$.

Pour estimer $\mathcal{D}^2(\phi)$, nous introduisons les observations

$$Z = (Z_{1,n}, \ldots, Z_{n,n}) \quad \text{où} \quad Z_{i,n} = Y_{i,n} - \sum_{k=1}^{p} \widehat{\beta}_k g_k(t_{i,n})\,, \quad i = 1, \ldots, n\,.$$

On procède comme dans Dette et Munk (1998), mais sans utilisation de poids, pour construire un estimateur empirique de $\mathcal{D}^2(\phi)$. Soit $\Delta_{i,n} = t_{i,n} - t_{i-1,n}$, $i = 2, \ldots, n$, $\Delta_{1,n} = \Delta_{2,n}$, $W = \operatorname{diag}\Big(\Delta_{i,n} h(t_{i,n})\Big)_{i=1,\ldots,n}$ et $g_{k,n} = \Big(g_k(t_1), \ldots, g_k(t_n)\Big)'$, $k = 1, \ldots, n$.

Soit $E_{p,n}$ le sous espace de $\mathbb{R}^n$ engendré par $(g_{1,n}, \ldots, g_{p,n})$ et $\Pi_n^{\perp}$ la matrice de projection sur l'hortogonal à $E_{p,n}$.

On définit $\Gamma_n(Z, g_1, \ldots, g_p)$ comme étant le déterminant obtenu en remplaçant dans $\Gamma(\phi, g_1, \ldots, g_p)$, les produits scalaires $\langle \phi, \phi \rangle$ et $\langle \phi, g_k \rangle$, $k = 1, \ldots, p$, respectivement par les quantités:

$$Z'WZ = \sum_{i=1}^{n} \Delta_{i,n} h(t_{i,n}) Z_{i,n}^2 \quad \text{et} \quad Z'W g_{k,n} = \sum_{i=1}^{n} \Delta_{i,n} h(t_{i,n}) g_k(t_{i,n}) Z_{i,n}\,, \quad k = 1, \ldots, p\,.$$

On obtient un estimateur de $\mathcal{D}^2(\phi)$ donné par:

$$\widehat{\mathcal{D}}_n^2 = \frac{\Gamma_n(Z, g_1, \ldots, g_p)}{\Gamma(g_1, \ldots, g_p)} - \widehat{\sigma}_R^2 \operatorname{tr}(W\Pi_n^{\perp})\,,$$

où $\widehat{\sigma}_R^2$ est l'estimateur de Rice (1984), défini par

$$\widehat{\sigma}_R^2 = \frac{1}{2(n-1)} \sum_{i=2}^{n} (Y_{i,n} - Y_{i-1,n})^2 .$$

On rejette H_0 si $\widehat{\mathcal{D}}_n^2 > c$. Le seuil ainsi que la puissance asymptotique du test sont obtenus en étudiant la loi asymptotique de $\widehat{\mathcal{D}}_n^2$. On a le comportement asymptotique de la statistique de test proposée.

Théorème 1 *Si les hypothèses* $(C1)$, $(C2)$ *et* $(C3)$ *sont satisfaites alors*

$$\sqrt{n} \left(\widehat{\mathcal{D}}_n^2 - \mathcal{D}(f)^2 \right) \xrightarrow[n \to +\infty]{\mathcal{L}} \mathcal{N} \left(0, \sigma^4 + 4\sigma^2 \, \mathcal{D}^2(f) \right), \quad \textit{quand} \quad n \to \infty.$$

3 Mise en œuvre du test

Le théorème précédent nous donne le niveau et la puissance du test. La procédure de test est la suivante: soit $\widehat{\sigma}^2$ un estimateur consistant de σ^2, puisque $\mathcal{D}^2(f) = 0$, lorsque $f \in E_p$, on rejette l'hypothèse nulle H_0, au niveau α, si

$$\sqrt{n} \, \widehat{\mathcal{D}}_n^2 \, / \widehat{\sigma}^2 > c_{1-\alpha},$$

où $c_{1-\alpha}$ est le $(1-\alpha)$-quantile d'une loi normale standard. $\widehat{\sigma}^2$ peut être remplacé par $\widehat{\sigma}_R^2$, cependant, il apparait dans nos simulations que l'utilisation $\widehat{\sigma}_R^2$ ne donne pas toujours une très bonne approximation par une loi normale; pour cela, nous proposons l'utilisation de l'estimateur consistant sous H_0 suivant:

$$\widehat{\sigma}_e^2 = \frac{1}{n} \sum_{i=1}^{n} \left\{ Y_{i,n} - \sum_{k=1}^{p} \widehat{\beta}_k g_k(t_{i,n}) \right\}^2 ,$$

qui donne une meilleur approximation par une loi normale pour une petite taille d'échantillon.

4 Simulations

Dans nos simulations, on étudie le test de l'hypothèse H_0: $f \in E_2$, où E_2 est le sous-espace vectoriel de $L^2(d\mu)$, engendré par $g_1(t) = t$ et $g_2(t) = 1$. On a mené une étude Monte Carlo en simulant le modèle (1), avec $t_{i,n} = (i-1)/(n-1)$, $\varepsilon_{i,n} \sim$ i.i.d.$\mathcal{N}(0, \sigma^2)$. L'étude consiste à faire une comparaison avec la statistique $\widehat{M}_n^2$ proposée par Dette et Munk (1998) pour une petite taille d'échantillon $n = 50$. Pour étudier la puissance du

test au niveau $\alpha = 0.05$, comme alternative à l'hypothèse H_0, on considère la fonction suivante: $f(t) = a_1 t + a_2 t + \gamma t e^{-2t}$, avec plusieurs valeurs de γ dans l'intervalle $[0, 2]$. Les résultats obtenus par les simulations montrent que les tests basés respectivement sur les deux statistiques $\widehat{M}_n^2$ et $\widehat{\mathcal{D}}_n^2$ sont comparables.

Bibliographie

[1] Dette, H., and Munk, A. (1998). Validation of linear regression models. *Ann. Stat.*, **26**, 778-800.
[2] Eubank, R. L. and Spiegelman, C. H. (1990). Testing the goodness-of-fit of a linear model via nonparametric regression techniques. *J. Amer. Stat. Assoc.* **85**, 410, 387-392.
[3] Mohdeb, Z. and Mokkadem, A. (2001). Testing Hypotheses On Fourier Coefficients in Nonparametric Regression Model. *Journal of Nonparametric Statistics*, **13**, 605-629.
[4] Mohdeb, Z. and Mokkadem, A. (2002). Testing the goodness-of-fit of a linear model in nonparametric regression. *Goodness-of-Fit Tests and Model Validity,* Stat. Ind. Technol., *Birkhauser Boston, Boston, MA.*, 185-193.
[5] Mohdeb, Z. and Mokkadem, A. (2004 a). Average squared residuals approach for testing linear hypotheses in nonparametric regression. The International Conference on Recent Trends and Directions in Nonparametric Statistics. *J. Nonparametr. Stat.* **16**, no. 1-2, 3–12.
[6] Mohdeb, Z. and Mokkadem, A. (2004 b). On the Use of Nonparametric Regression for Testing Linear Hypotheses. *Ann. I.S.U.P.* **48**, no. 3, 63–77.
[7] Rice, J. (1984). Bandwidth choice for nonparametric regression. *Ann. Stat.* **12**, 1215-1230.

Multi-sample comparison of multivariate copulas and clustering

Yves I. Ngounou Bakam[1] & Denys Pommeret[1,2]

[1] *CNRS, Ecole Centrale, I2M, Aix-Marseille Univ, Marseille, France*
yves-ismael.ngounou-bakam@univ-amu.fr
[2] *ISFA, Univ Lyon, UCBL, LSAF EA2429, F-69007, Lyon, France*
denys.pommeret@univ-amu.fr

Abstract. The comparison of copulas is a major challenge in copula modelling. Copulas are still an important topic with many applications in finance, actuarial science, environmental data, ecology, and so on. They can be used to model the dependence structures of multivariate observations.
In the two-sample case, Rémillard and Scaillet (2009) proposed a test to compare two nonparametric copulas, that is to test $H_0 : C1 = C2$, where C_1 and C_2 are two copulas observed on two iid samples, which may be paired.
To our knowledge, there is no extension to the K sample case. However, the increasing amount of data requires sometimes more comprehensive analyzes. It is in this sense that we propose an equality test of K copulas simultaneously when K populations are observed. We propose to test the following hypothesis:

$$H_0 : \quad C_1 = \cdots = C_K,$$

from K iid samples, possibly paired. It is therefore a generalization of Rémillard and Scaillet (2009). However, we obtain the exact asymptotic distribution of the test statistic and the convergence of the test. The idea of the test is to transform the observations to uniform laws, then to use the decomposition of the density of the copula in the Legendre polynomials orthogonal basis. Returning to the copula function we obtain what are called copula coefficients which characterize each copula. The test then allows to simultaneously comparing these coefficients. The number of involved copula coefficients is automatically selected by a data driven selection. In addition, we suggest a clustering algorithm to classify populations with similar forms of dependence structure. A simulation study, analyse the level and the power of the test to show the good behavior of our test procedure and its performances to Rémillard and Scaillet (2009) approach within two-sample case. We provide some illustrations of the test and clustering through a real datasets in insurance and finance to demonstrate the method.

Keywords. Copula, Copula coefficients, Multi-sample, Clustering, Data driven.

Consistance de l'intervalle de confiance plug-in en régression

Ahmed ZAOUI & Christophe DENIS & Mohamed HEBIRI

Laboratoire LAMA, Université Gustave Eiffel,
Ahmed.Zaoui/Christophe.Denis/Mohamed.Hebiri@univ-eiffel.fr

Résumé. Nous nous intéressons au problème de régression hétéroscédastique. L'objectif est de construire un ensemble de confiance auquel la variable réponse Y appartient avec une grande probabilité. L'ensemble optimal repose sur le seuillage de la densité conditionnelle. La construction de cet ensemble repose sur une procédure d'estimation semi-supervisée en deux étapes. Dans une première étape, à partir d'un premier échantillon de données étiquetées, nous estimons la densité conditionnelle. Dans une deuxième étape, nous calibrons le seuil à l'aide d'un deuxième échantillon de données non étiquetées. L'ensemble construit offre des garanties de consistance et de bonnes performances numérique.
Mots-clés. Régression, Plug-in, Ensemble de confiance.

Abstract. We are interested in the heteroscedastic regression problem. The objective is to construct a confidence set to which the response variable Y belongs with high probability. The optimal set is based on the thresholding of the conditional density. The construction of this set is based on a two-step semi-supervised estimation procedure. In a first step, we estimate the conditional density from a first sample of labeled data. In a second step, we calibrate the threshold using a second sample of unlabeled data. The constructed ensemble offers guarantees of consistency and good numerical performance.
Keywords. Regression, Plug-in, Confidence set.

1 Introduction

Les ensembles de confiance jouent un rôle central en statistique. Les techniques d'estimation sont basées sur différentes méthodes : les splines (Wahba, 1983), l'estimation à noyau (Eubank et al, 1993), le boostrap (Efron ,1987) etc. Néanmoins, les procédures d'estimation commettent des erreurs de prédiction. En pratique, il est important de disposer de mesures précises de l'incertitude des prévisions d'un modèle pour éviter de lourdes conséquences (par exemple dans le domaine médical).

La prédiction conforme fait l'objet de nombreux efforts de recherche récents en statistique sur les plans théoriques, méthodologiques et informatiques. Elle a été initialement formulée pour la tâche de classification (Vovk et al 2005, Shafer et al 2008 $\cdots$). Puis, elle été utilisée en régression (Lei et al 2013, Lei et al 2014, Lei et al 2018). Sa théorie suggère de minimiser la taille de l'ensemble de confiance avec un niveau de confiance fixe sur le niveau d'erreur. Étant donné l'échantillon $\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n$, l'objectif est de prédire Y_{n+1}

associée à la nouvelle entrée X_{n+1}. Plus précisément, étant donné un niveau de confiance $\alpha \in [0,1]$, la prédiction conforme permet de fournir un prédicteur de confiance mesurable $\mathbf{C}_n(X_{n+1}) = \mathbf{C}_n(X_1, Y_1, X_n, Y_n, X_{n+1}) \subseteq \mathbb{R}$ qui satisfait

$$\mathbf{P}\left(Y_{n+1} \in \mathbf{C}_n(X_{n+1})\right) \geq 1 - \alpha,$$

où $\mathbf{P} = \mathbb{P}^{n+1}$ est la distribution jointe de $(X_1, Y_1), \cdots, (X_{n+1}, Y_{n+1})$. Nous pouvons aussi souhaiter des garanties plus fortes avec la validité marginale. Nous parlons ici de la validité conditionnelle :

$$\mathbf{P}\left(Y_{n+1} \in \mathbf{C}_n(X_{n+1}|X_{n+1} = x)\right) \geq 1 - \alpha \ .$$

Elle garantit une couverture adéquate pour chaque nouvelle variable d'entrée. Lei et al 2014 ont construit des bandes de prédiction avec une validation locale. Lei et al 2018 ont étudié plusieurs stratégies de construire un prédicteur de confiance pour la variable de réponse en utilisant n'importe quel estimateur de la fonction de régression.

Notre travail consister à construire un intervalle de confiance en contrôlant sa longueur. Nous déterminons une formule explicite de cet intervalle. Puis, nous proposons une méthode d'estimation de ce dernier et étudions les propriétés statistiques.

2 Cadre de travail

Nous nous plaçons dans le cadre de la régression. Une donnée observée est de la forme $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$ admettant une loi de probabilité $\mathbb{P}$ sur $\mathbb{R}^d \times \mathbb{R}$ telle que

$$Y = f^*(X) + \sigma(X)\xi \ ,$$

où $f^*(x) = \mathbb{E}\left[Y|X = x\right]$ est la fonction de régression f^*, et $\sigma^2(x) = \mathbb{E}\left[(Y - f^*(X)^2|X = x\right]$ est la fonction de variance pour tout $x \in \mathbb{R}^d$. La variable aléatoire ξ, dite de bruit, est centrée et indépendante de X. Nous nous concentrons sur deux cadres : Y est borné ou Y satisfait le modèle gaussien où $\xi \sim \mathcal{N}(0,1)$.
Soit $\mathcal{I}_{\mathbb{R}}$ un ensemble d'intervalles réels. Un intervalle de confiance $\texttt{CI}$ est un ensemble de $\mathbb{R}^d$ dans $\mathcal{I}_{\mathbb{R}}$ tel que $\texttt{CI}(X) \in \mathcal{I}_{\mathbb{R}}$. Il a deux caractéristiques :

- l'erreur : $\text{Err}(\texttt{CI}) = \mathbb{P}\left(Y \notin \texttt{CI}(X)\right)$;
- le taux de la longueur : $\texttt{EL}(\texttt{CI}) := \mathbb{E}_X\left[L(\texttt{CI}(X))\right]$ où $L(\cdot)$ représente la longueur de l'intervalle.

Pour mesurer la performance d'un intervalle $\texttt{CI}$, nous considérons la mesure de risque suivante

$$\mathcal{R}_\lambda(\texttt{CI}) = \text{Err}(\texttt{CI}) + \lambda\texttt{EL}(\texttt{CI}) \ ,$$

où le paramètre $\lambda > 0$ contrôle l'équilibre entre les deux termes. Nous nous intéressons à minimiser l'erreur de l'intervalle de confiance en ajoutant une contrainte sur le taux de la longueur du $\texttt{CI}(X)$. Fixons $\ell > 0$. Le ℓ-intervalle de confiance optimal est défini par :

$$\texttt{CI}^*_\ell \in \arg\min\{\text{Err}\left(\texttt{CI}\right) \ : \ \texttt{CI} \text{ tel que } \texttt{EL}(\texttt{CI}) \leq \ell\} \ .$$

Concernant ce problème d'optimisation : Est-ce que $\texttt{CI}_\ell^*$ admet une expression explicite ? La réponse est oui, sous des hypothèses techniques. D'abord, nous supposons que la variable $Y|X$ admet une densité $p(Y|X)$. Puis, nous introduisons la quantité suivante : $\forall t > 0$

$$G(t) := \int_{\mathbb{R}} \mathbb{P}_X(p(y|X) \geq t)dy.$$

La fonction G est décroissante sur $\mathbb{R}_+^*$. Elle est continue si $t \mapsto \mathbb{P}_X(\hat{p}(y|X) \geq t)$ est continue sur $\mathbb{R}_+^*$ pour tout $y \in \mathbb{R}$. Sous la continuité de G et les propriétés de la pseudo inverse G^{-1} [1] de G (voir Embrechts et al 2013), il est clair que $G(p(y|X)) \leq \ell \Leftrightarrow p(Y|X) \geq G^{-1}(p(Y|X)$. Cette équivalence permet de déduire le résultat suivant :

Théorème 1. *Soit $\ell > 0$. Les assertions suivantes sont vérifiées :*
(i) le ℓ-intervalle de confiance optimal *est défini par :*

$$\texttt{CI}_\ell^*(X) = \{y : p(y|X) \geq \lambda_\ell\}, \tag{1}$$

où $\lambda_\varepsilon = G^{-1}(\ell)$. De plus, $\texttt{EL}(\texttt{CI}) = \ell$.
(ii) l'excès de risque $\mathcal{E}(\texttt{CI}) := \mathcal{R}_{\lambda_\ell}(\texttt{CI}) - \mathcal{R}_{\lambda_\ell}(\texttt{CI}_\ell^*)$ de tout intervalle de confiance $\texttt{CI}$ peut s'écrire explicitement sous la forme :

$$\mathcal{E}(\texttt{CI}) = \mathbb{E}_X \left[\int_{\texttt{CI}(X) \triangle \texttt{CI}_\ell^*(X)} |p(y|X) - \lambda_\ell| dy \right] .$$

Ce théorème établit un point crucial, à savoir la valeur de λ pour laquelle l'intervalle de confiance a exactement un taux de longueur égal à ℓ. De plus, le point (ii) du Théorème 1 donne une formulation explicite de l'excès de risque d'un intervalle de confiance $\texttt{CI}$. En particulier, il ne dépend que du comportement de la densité conditionnelle $p(Y|X)$ autour du seuil λ_ℓ.

2.1 Stratégie d'estimation

L'objectif de cette section est de fournir un estimateur de $\texttt{CI}_\ell^*$ ayant des propriétés statistiques proches de $\texttt{CI}_\ell^*$. La nature de l'oracle donné par (1) suggère une approche semi-supervisée de type plug-in qui repose sur l'estimation de la densité conditionnelle $p(Y|X)$ et de seuil λ_ℓ. Nous utilisons deux échantillons indépendants : $\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n$, et $\mathcal{D}_N = \{X_i\}_{i=n+1}^{n+N}$. Nous construisons un estimateur $\hat{p}(Y|X)$ de $p(Y|X)$ sur $\mathcal{D}_n$. Puis, nous calibrons λ_ℓ sur $\mathcal{D}_N$. Enfin, nous obtenons l'estimateur construit : voir Algorithme 1. Afin de garantir des propriétés de consistance du $\hat{\texttt{CI}}_\ell$ en termes de risque et de taux de longueur, nous ajoutons une perturbation ζ dans l'étape 3. Par l'inégalité triangulaire, la consistance de $\hat{p}(Y|X)$ implique la consistance de $\hat{\hat{p}}(Y|X)$ quand u tend vers 0. En

1. $\mathrm{G}^{-1}(\ell) := \inf\{t > 0 : G(t) \leq \ell\}$.

outre, différents types de méthodes non paramétriques ont été examinées pour estimer des densités conditionnelles : les estimateurs à noyau (Hyndman et al 1996), les estimateurs par polynômes locaux (Hyndman et al 2002), les estimateurs par projection (Efromovich 2007).

Algorithme 1 *Intervalle de confiance plug-in de longueur* ℓ

1. Nous introduisons deux échantillons indépendants : $\mathcal{D}_n$ et $\mathcal{D}_N = \{X_i\}_{i=n+1}^{n+\mathcal{N}}$ qui contient N observations indépendantes, et *non étiquetées*, de même loi que X.
2. À partir de $\mathcal{D}_n$, nous construisons un estimateur $\hat{p}(Y|X)$ de $p(Y|X)$.
3. Sur $\mathcal{D}_N$, nous estimons G :
   (a) Randomisation :
   $$\hat{\hat{p}}(Y|X_i) = \hat{p}(Y|X_i) + \zeta_i, \qquad \zeta_i \sim \mathcal{U}([0,u]),$$
   (b) Fonction empirique de G :
   $$\hat{G}(t) = \int_{\mathbb{R}} \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{\{\hat{\hat{p}}(y|X_i) \geq t\}} dy.$$
4. Sortie d'*intervalle de confiance plug-in de longueur* ℓ
   $$\hat{\mathtt{CI}}_\ell(x) = \{y \in \mathbb{R} : \hat{G}(\hat{p}(y|x)) \leq \ell\} \ . \tag{2}$$

2.2 Consistance

Cette section est dédié au résultat principal de ce papier. Nous étudions les propriétés de consistance d'*intervalle de confiance plug-in de longueur* ℓ défini par l'Eq. (2). Pour cela, nous imposons l'hypothèse suivante :

$$\mathbb{E}\left[\int_{\mathbb{R}} |\hat{p}(y|X) - p(y|X)| dy\right] \underset{n\to+\infty}{\longrightarrow} 0. \tag{3}$$

Cela signifie que l'estimateur $\hat{p}(y|X)$ est consistant par rapport l'erreur L^1 intégrée. Sous cette hypothèse nous obtenons le résultat principal :

Théorème 2. *Soit* $\ell > 0$ *et* $\lambda_\ell = G^{-1}(\ell)$. *Supposons que* $p(y|X)$ *est borné. Sous des hypothèses de régularité, nous avons*

$$\mathbb{E}\left[\mathcal{E}_{\lambda_\ell}\left(\hat{\mathcal{CI}}_\ell\right)\right] \underset{n,N\to+\infty}{\longrightarrow} 0, \quad \textit{et} \quad \mathbb{E}\left[\mathit{EL}(\hat{\mathcal{CI}}_\ell\right] \underset{N\to+\infty}{\longrightarrow} \ell \ .$$

Ce résultat illustre le fait que $\hat{\mathtt{CI}}_\ell$ est asymptotiquement aussi bon que le prédicteur optimal et de longueur ℓ.

2.3 Consistance : cas du modèle gaussien

Dans ce qui suit, nous nous plaçons dans le modèle gaussien où ξ suit une loi gaussienne centrée réduite. En particulier, la variable $Y|X$ suit une loi gaussienne de moyenne $f^*(X)$ et de variance $\sigma^2(X)$. Sa densité est donnée comme suit :

$$p(Y|X) = \frac{1}{\sqrt{2\pi\sigma^2(X)}} \exp\left(-\frac{(Y-f^*(X))^2}{2\sigma^2(X)}\right) \ .$$

Nous proposons un estimateur de type plug-in de $p(Y|X)$

$$\hat{p}(Y|X) = \begin{cases} \frac{1}{\sqrt{2\pi\hat{\sigma}^2(X)}} \exp\left(-\frac{(Y-\hat{f}(X))^2}{2\hat{\sigma}^2(X)}\right) & \text{if } \hat{\sigma}^2(X) > 0 \\ 0 & \text{sinon .} \end{cases} \tag{4}$$

où $\hat{f}$ et $\hat{\sigma}^2$ sont deux estimateurs de f^* et σ^2 respectivement. Le problème de l'estimation de la fonction de régression f^* est classique et largement étudié par exemple par Wand et Jones (1994), Györfi et al (2002), ou encore Biau et Devroye (2015). De nombreuses méthodes sont proposées pour estimer la fonction de variance conditionnelle. Dans ce papier, nous nous intéressons à la méthode basée sur les erreurs résiduelles. Elle consiste en deux étapes. Dans une première étape, nous construisons un estimateur $\hat{f}$ de la fonction de régression f^*. Dans une deuxième étape, nous obtenons un estimateur de σ^2 en résolvant le problème de régression où la variable d'entrée est X et la variable à prédire est $(Y-\hat{f}(X))^2$. Pour plus de détails, nous renvoyons à Ruppert et al (1997), Fan et al (1998), Kulik et al (2011), Denis et al (2020) et Zaoui (2021).

Le but de cette partie est de prouver que l'estimateur (4) satisfait l'hypothèse (3). Pour cela, nous faisons l'hypothèse suivante

Hypothèse 1. *Nous supposons qu'il existe des constantes positives σ^2_{min} et σ^2_{max} telle que*

$$0 < \sigma^2_{min} \leq \sigma^2(X) \leq \sigma^2_{max} \ .$$

Cette hypothèse est faible. Elle garantit que la fonction de variance est strictement positive. Nous établissons le résultat suivant

Théorème 3. *Supposons que f^* est borné, $\hat{f}$ et $\hat{\sigma}^2$ sont deux estimateurs consistants de f^* et de σ^2 respectivement en norme L^2. C'est-à-dire que*

$$\mathbb{E}\left[|\hat{f}(X) - f^*(X)|^2\right] \underset{n\to+\infty}{\longrightarrow} 0, \ \text{ et } \ \mathbb{E}\left[|\hat{\sigma}^2(X) - \sigma^2(X)|^2\right] \underset{n\to+\infty}{\longrightarrow} 0.$$

Alors, sous l'hypothèse 1, nous avons

$$\mathbb{E}\left[\int_{\mathbb{R}} |\hat{p}(y|X) - p(y|X)|dy\right] \underset{n\to+\infty}{\longrightarrow} 0.$$

En particulier,

$$\mathbb{E}\left[\mathcal{E}_{\lambda_\ell}\left(\hat{\mathcal{C}}\mathcal{I}_\ell\right)\right] \underset{n,N\to+\infty}{\longrightarrow} 0 \ .$$

Bibliographie

Biau, G. et Devroye, L. (2015). Lectures on the Nearest Neighbor Method, *Springer Series in the Data Sciences, Springer New York*.

Denis, C., Hebiri, M. et Zaoui, A. (2020). Regression with reject option and application to knn. *NeurIPS 2020*.

Györfi, L., Kohler, M., Krzyżak, A. et Walk, H. (2002). A distribution-free theory of nonparametric regression, *Springer Ser. Statist.*

Efromovich, S. (2007). Conditional density estimation in a regression setting. *Ann. Statist.*, 35(6) :2504–2535.

Efron, B 1987. Better Bootstrap Confidence Intervals. *Journal of the American Statistical Association*, 82(397) :171–185.

Embrechts, P. et Hofert, M. (2013). A note on generalized inverses. *Mathematical Methods of Operations Research*,77 ;423–432.

Eubank, R.L. and Speckman P.L. 1993. Confidence Bands in Nonparametric Regression. *Journal of the American Statistical Association*, 88 :1287–1301.

Hyndman, R. J. et Yao, Q. (2002). Nonparametric estimation and symmetry tests for conditional density functions. *J. Nonparametr. Stat.*, 14(3) :259–278.

Hyndman, R. J., Bashtannyk, D. M., et Grunwald, G. K. (1996). Estimating and visualizing conditional densities. *J. Comput. Graph. Statist.*, 5(4) :315–336.

Kulik, R., et Wichelhaus, C. (2011). Nonparametric conditional variance and error density estimation in regression models with dependent errors and predictors. *Electron. J. Statist.*, 5 :856–898.

Lei, J. et G'Sell, M. et Rinaldo, A. et Tibshirani, R. J. et WassermanL. 2018. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523) :1094–1111.

Lei, J. et Robins, J. et Wasserman, L. 2013. Distribution-free prediction sets. *Journal of the American Statistical Association*, 108(501) :278–287.

Lei, J et Wassermann L. 2014. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society*, Series B, 76(1) :71–96.

Ruppert, D., Wand, M.P., Holst, U. et HöSJER, O.(1997). Local polynomial variance function estimation. *Technometrics*, 39(3) :262–273.

Shafer, G. et Vovk, V. 2008. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(12) :371–421.

Vovk, V. et Gammerman, A. et Shafer, G. 2005. Algorithmic learning in a random world. *Springer*, New York.

Wahba, G (1983). Bayesian "Confidene Intervals" for the Cross-Validated Smoothing Spline. *Journal of the Royal Statistical Society*. Series B, 45(1) :133–150.

Wand, M. P. et Jones M.C. (1994). Kernel smoothing, *CRC Press.*

Zaoui, A. (2021). Variance function estimation in regression model via aggregation procedures. *preprint*.

SFdS - Biométrie

Modélisation conjointe des relations temporelles entre marqueurs longitudinaux multivariés de progression de la Maladie d'Alzheimer et événements cliniques

Anaïs Rouanet [1] & Cécile Proust-Lima [2]

[1] *Inserm U1219 - Bordeaux Population Health research Center, France, anais.rouanet@u-bordeaux.fr*
[2] *Inserm U1219 - Bordeaux Population Health research Center, France, cecile.proust-lima@u-bordeaux.fr*

Résumé. La richesse des données de biomarqueurs désormais disponibles dans les cohortes sur le vieillissement offre l'opportunité de mieux appréhender les mécanismes complexes de la maladie d'Alzheimer (MA), ce qui est essentiel à l'amélioration de stratégies de prévention et de prise en charge des patients. Cependant, les outils statistiques actuels ne permettent pas de modéliser conjointement les multiples biomarqueurs de la MA tout en capturant leurs relations temporelles. Nous proposons un nouvel outil de modélisation conjointe pour décrire la dynamique complexe des différentes dimensions impliquées dans la progression de la MA et appréhender leurs relations temporelles causales. Ce modèle causal dynamique combine un modèle mixte multivarié avec équations de différence pour expliquer l'évolution dans le temps de chaque dimension en fonction des caractéristiques des autres. Les associations possibles avec le diagnostic de la MA et le décès sont prises en compte via une approche de modélisation conjointe à effets aléatoires partagés. La méthodologie est appliquée à la cohorte prospective française PAQUID pour décrire les relations temporelles entre la cognition, la dépression et l'autonomie fonctionnelle, en lien avec les deux principaux événements cliniques de la progression de la MA : le diagnostic de démence et le décès. Ce travail a pour but d'aider à comprendre l'interaction complexe entre les biomarqueurs de la MA et à identifier les cibles pertinentes capables de ralentir la progression à chaque stade de la maladie.

Mots-clés. Maladie d'Alzheimer, causalité, modélisation conjointe, données manquantes, marqueurs multivariés, relations temporelles.

Abstract. The wealth of biomarker data now available offers an unprecedented opportunity to better understand the complex pathways of Alzheimer's Disease (AD), essential to improve prevention strategies and individual healthcare. However, current statistical tools are inefficient at modelling jointly the multiple AD biomarkers and capture their inter-relationships rigorously. We propose an original joint modelling framework to describe the complex dynamics of the different dimensions involved in AD progression and apprehend their causal temporal relationships. This dynamic causal model combines a multivariate mixed model with difference equations to explain the change over time of

the dimensions according to different characteristics of the others. The possible associations with AD diagnosis and death are taken into account via a shared random effect joint modelling approach.The methodology is applied on a French epidemiological Paquid study to disentangle the temporal relationships between depression, cognition and functional autonomy, in link with the two major clinical events in AD progression: diagnosis and death. These methodology and application aim to help in understanding the complex interplay between the biomarkers of AD, and in identifying relevant targets which may slow down the progression at each stage of the disease.

Keywords. Alzheimer's disease, causality, joint modelling, missing data, multivariate markers, temporal relationships.

1 Contexte

La maladie d'Alzheimer (MA) est l'une des principales causes de handicap chez les personnes âgées (Dartigues et al., 2012) et aucun traitement ne s'est avéré efficace pour ralentir sa progression, probablement parce qu'ils sont administrés trop tard dans le processus de dégradation (Waite, 2015). La richesse des données biologiques, cliniques et épidémiologiques désormais disponibles dans les cohortes sur le vieillissement permettent de mieux appréhender les voies neuropathologiques de la MA. Jack et al. (2013) ont proposé un modèle théorique fondamental de la cascade pathologique de la MA, établissant un ordre temporel entre les biomarqueurs de la MA et les symptômes cognitifs et cliniques. Cependant, la littérature manque de méthodes statistiques permettant de tester formellement ces hypothèses et, plus généralement, d'étudier leurs relations temporelles causales, c'est-à-dire la manière dont une dimension peut avoir un impact sur le changement ultérieur d'une autre dimension. Les outils statistiques actuels sont inefficaces pour modéliser rigoureusement de telles relations et ne tiennent souvent pas compte de la complexité des données provenant de cohortes épidémiologiques de personnes âgées. L'absence de modèles statistiques appropriés pour rendre compte de la dynamique complexe des marqueurs biologiques et cliniques de la MA constitue un obstacle majeur à la compréhension de son histoire naturelle et donc au développement de traitements efficaces.

2 Modélisation

Nous proposons un cadre novateur de modélisation de la MA pour décrire la dynamique complexe des dimensions pathologiques impliquées et appréhender leurs relations temporelles causales. Les différentes dimensions sont représentées par des processus latents, décrits par des modèles mixtes linéaires, et leur corrélation est capturée par des équations de différence en temps discret. Chaque processus est mesuré par un ou plusieurs marqueurs

répétés, en utilisant des équations d'observation non linéaires qui traitent des variables continues et/ou ordinales. Par exemple, une transformation spline permet d'accommoder des tests cognitifs non gaussiens dont on sait qu'ils ont des propriétés psychométriques telles que des effets plancher ou plafond. Enfin, nous explorons également la relation entre les dimensions pathologiques et les événements cliniques, le diagnostic de démence et le décès, en modélisant ces derniers conjointement, en tant que fonctions de la valeur actuelle des processus latents. Cela nous permet également de tenir compte des données manquantes informatives dues à la démence et au décès, et de réduire les biais de sélection potentiels. Ce modèle conjoint pour les résultats multivariés et les événements concurrents est estimé par maximisation de la vraisemblance, en utilisant l'approximation Quasi Monte Carlo (Philipson et al., 2020) et un algorithme Levenberg-Marquardt avec des critères de convergence stricts (Philipps et al., 2020). Ce travail est basé sur Tadde et al. (2019) qui, initialement, ne prenait pas en compte les résultats ordinaux ni les événements concurrents.

3 Application

Ce modèle dynamique est appliqué à la cohorte prospective française Paquid pour décrire les relations temporelles entre la dépression, la cognition et l'autonomie fonctionnelle. L'étude Paquid comprend des participants du Sud-Ouest de la France, âgés de 65 ans et plus, vivant de manière autonome à leur domicile en 1988-1989, suivis pendant plus de 25 ans. Ce modèle permet de tester différentes hypothèses sur les relations entre les dimensions impliquées dans la MA et d'identifier les domaines cibles pertinents qui peuvent ralentir sa progression, à chaque stade de la maladie. Plus généralement, ce type de résultats peut fournir des indications pour le développement de traitements et le déploiement de stratégies de prévention en santé publique.

Bibliographie

Dartigues J-F, Helmer C, Letenneur L, Pérès K, Amieva H, Auriacombe S, Orgogozo JM, Commenges D, Jacqmin-Gadda H, Richard-Harston S, Delva F, Foubert-Samier A, Barberger-Gateau P. (2012), Paquid 2012: illustration et bilan. *Geriatrie et Psychologie Neuropsychiatrie du Vieillissement*, 10(3):325-331.

Waite LM. (2015), Treatment for Alzheimer's disease: has anything changed? *Australian Prescriber*, 38(2): 60-63.

Jack CR Jr, Knopman DS, Jagust WJ, Petersen RC, Weiner MW, Aisen PS, Shaw LM, Vemuri P, Wiste HJ, Weigand SD, Lesnick TG, Pankratz VS, Donohue MC, Trojanowski

JQ. (2013), Tracking pathophysiological processes in Alzheimer's disease: an updated hypothetical model of dynamic biomarkers. *Lancet Neurology*, 12(2):207-216.

Philipson, P., Hickey, G. L., Crowther, M. J., & Kolamunnage-Dona, R. (2020), Faster Monte Carlo estimation of joint models for time-to-event and multivariate longitudinal data. *Computational Statistics & Data Analysis*, 151, 107010.

Philipps, V., Hejblum, B. P., Prague, M., Commenges, D., Proust-Lima, C. (2021), Robust and Efficient Optimization Using a Marquardt-Levenberg Algorithm with R Package marqLevAlg, *R journal*, in press, arXiv:2009.03840.

Taddé, B. O., Jacqmin-Gadda, H., Dartigues, J.-F., Commenges, D., & Proust-Lima, C. (2020), Dynamic modelling of Multivariate Latent Processes and Their Temporal Relationships: Application to Alzheimer's Disease, *Biometrics*, 76(3):886-899.

ÉVALUATION PAR LA MODÉLISATION DES STRATÉGIES DE RAPPEL VACCINAL CONTRE LA COVID-19 DANS UNE POPULATION PARTIELLEMENT VACCINÉE

Clément R Massonnaud[1*], Jonathan Roux[2*], Vittoria Colizza[3], Pascal Crépey[4]

[1] *EHESP, Université de Rennes, CNRS, Inserm, Arènes - UMR 6051, RSMS – U 1309, 35043 Rennes, France. clement.massonaud@ehesp.fr*
[2] *EHESP, Université de Rennes, CNRS, Inserm, Arènes - UMR 6051, RSMS – U 1309, 35043 Rennes, France et Groupe Hospitalier Universitaire AP-HP Nord-Université de Paris, Hôpital Bichat Claude-Bernard, Département d'Épidémiologie, Biostatistiques et Recherche Clinique, Paris, France. jonathan.roux@ehesp.fr*
[3] *INSERM, Sorbonne Université, Pierre Louis Institute of Epidemiology and Public Health, Paris, France. vittoria.colizza@inserm.fr*
[4] *EHESP, Université de Rennes, CNRS, Inserm, Arènes - UMR 6051, RSMS – U 1309, 35043 Rennes, France. pascal.crepey@ehesp.fr*
[*] *Authors contributed equally to the work.*

Résumé. Alors que les preuves montrant que l'immunité vaccinale contre le COVID-19 diminue avec le temps et l'apparition de nouveaux variants, plusieurs pays mettent en œuvre des campagnes de rappel. L'objectif de cette étude est d'analyser la morbidité et la mortalité associée à différentes stratégies de primo-vaccination et de rappel contre la COVID-19, en utilisant la France comme cas d'étude. Nous utilisons un modèle compartimental déterministe, structuré par âge, calibré sur les données d'admission à l'hôpital et validé par rapport aux données de séroprévalence en France. Ce modèle permet d'analyser l'impact des stratégies de rappel vaccinal sur la morbidité et la mortalité en supposant une diminution de l'immunité et une transmissibilité accrue du virus pendant l'hiver. L'efficacité des stratégies de rappel ciblant différentes tranches d'âge varie selon les niveaux de transmissibilité du virus, et selon la perte d'immunité supposée pour chaque tranche d'âge. Si la baisse de l'immunité touche toutes les tranches d'âge, les personnes âgées de 30 à 49 ans devraient être boostées en priorité, même pour des niveaux de transmissibilité faibles. Si la réduction de l'immunité est limitée aux personnes âgées de plus de 65 ans, le rappel vaccinal des personnes plus jeunes ne devient efficace qu'au-delà de certains niveaux de transmissibilité. L'augmentation de la couverture vaccinale primaire doit rester une priorité pour réduire la morbidité et la mortalité dues à la COVID-19. Si un plateau de primo-vaccination a été atteint, le renforcement de l'immunité dans les groupes d'âge les plus jeunes pourrait prévenir plus d'hospitalisations et de décès que le renforcement de l'immunité des personnes âgées, en particulier dans des conditions d'augmentation la transmissibilité du SARS-CoV-2, ou face à de nouveaux variants.

Mots-clés. COVID-19, vaccination, stratégies vaccinales, booster, modèle épidémiologique

Abstract. As evidence shows that vaccine immunity to COVID-19 wanes with time and decreases due to variants, several countries are implementing booster vaccination campaigns. The objective of this study was to analyze the morbidity and mortality burdens of different booster vaccination strategies against COVID-19, using France as a case study. We used a deterministic, age-structured, compartmental model fitted to hospital admission data and validated against seroprevalence data in France to analyze the impact of booster vaccination strategies on morbidity and mortality assuming waning of immunity and increased virus transmissibility during winter. The effectiveness of booster strategies targeting different age groups varied with the levels of virus

transmissibility, and according to the assumed loss of immunity for each age group. If the immunity reduction affects all age groups, people aged 30 to 49 years should be boosted in priority, even for low transmissibility levels. If the immunity reduction is restricted to people older than 65 years, boosting younger people becomes effective only above certain levels of transmissibility. Increasing the primary vaccination coverage should remain a priority to reduce morbidity and mortality due to COVID-19. If a plateau of primary vaccination has been reached, boosting immunity in younger age-groups could prevent more hospitalizations and deaths than boosting the immunity of older people, especially under conditions increasing SARS-CoV-2 transmissibility, or when facing new variants.

Keywords. COVID-19, vaccination, vaccination strategy, booster, epidemiological model

1 Introduction

Several highly effective COVID-19 vaccines have been made available worldwide since the end of 2020. In France, four vaccines have been authorized (1) and used in vaccination campaigns to help mitigate the epidemic. The vaccination campaign started with the elderly population and at-risk population in January 2021 and was progressively extended to younger age-groups down to 12 years of age. The vaccination coverage for people aged 65 years or more reached 60% in June 2021 and plateaued at the end of the summer at 85%. However, coverage remained low in specific senior age classes; for example, only 78% of 80+ received a full vaccination. The vaccination of individuals aged 18 to 64 started later during the spring and took off in May for the 50-64 age groups, and June for the 18-49 age groups. The vaccination coverage increased rapidly but then slowed down significantly by the end of the summer, at levels ranging from 70% (30-39 age group) to 85% (60-64 age group). The vaccination of the 12-17 age group started in July, and reached 50% at the end of August. (2)

Among the SARS-CoV-2 variants that spread worldwide, the variant of concern Alpha (VOC B.1.1.7) became dominant in metropolitan France in February 2021,(3) and was then replaced by the more transmissible Delta VOC (B.1.617.2), which has been dominant since July 2021.(4) Several studies now suggest that the efficacy of vaccines against the Delta VOC is lower than against other variants.(5) Moreover, recent studies show that vaccine immunity, especially regarding infection and symptomatic form, is waning with time.(5,6) Therefore, many countries have recommended vaccine booster shots for all or part of their already vaccinated population. In France, only people aged 65 years or more (and specific people at risk) were initially eligible to receive a booster dose, administered at least 6 months after completion of the primary vaccination. Recommendations were extended to the 18+ vaccinated population in August 2021, similarly to other countries.(7) With the surging wave in Europe, and in the context of global vaccine supply constraints, it is important to assess the effectiveness of booster vaccination campaigns. The objective of this study was to analyze the impact, on hospital admissions and deaths, of COVID-19 booster vaccination strategies targeting different age groups, under various hypotheses of vaccine efficacy and virus transmission levels, using France as a case study.

2 Methods

2.1 COVID-19 transmission model

We used a deterministic, age-structured, compartmental epidemic model, based on the demographic and age profile of the population of metropolitan France. This model, previously described in (8), accounts for differences in susceptibility, severity, and contacts across age groups. To consider the impact of vaccines, we take a multi-branch approach, duplicating the main branch for each vaccine considered. In this framework, non-vaccinated individuals from susceptible and removed compartments could be vaccinated and integrate a vaccinated branch. This multi-branch model allowed us to use vaccine-specific efficacies in each branch, and therefore to model booster doses.

The main outcomes returned by the model were the daily number of newly exposed and hospitalized individuals, among others.

2.2 Parametrization

The model is fitted dynamically on the number of observed hospitalizations from the start of the pandemic in March 2020 to September 2021. We use a 7-day sliding window to estimate transmission parameters and shift the window by steps of 3 days. This method allows us to precisely reproduce epidemiological dynamics due to variations of transmissibility related to external factors like mitigation measures, season, or new variants. To account for the impact of VOCs on transmission parameters, we considered a 40%-increase in the risk of hospitalization between Alpha VOC and wild type (9) and an 80%-increase between Delta VOC and Alpha VOC.(10) Moreover, to take into account an increase in transmissibility of the virus due to seasonality during the simulated period, we progressively increased the transmission over two weeks, starting on September 20, 2021, and then held it constant until the end of the simulation (March 1st, 2022), to reproduce a 41% increase in wintertime (11) of the average reproduction number estimated in August (R_{eff}=1.0). In addition, we explored a range from 0% to 200% of increase in transmissibility due to seasonality or new variants. The model successfully reproduced the observed evolution of the epidemic in France, and estimates of the immune status of the population are in accordance with what is reported by observational studies.(12)

2.3 Vaccine efficacies

Our transmission model allowed us to consider vaccine efficacies against three COVID-19 stages: infection, symptomatic case, and hospitalization. Two scenarios were considered when the Delta VOC was predominant: 1) a decrease in vaccine effectiveness only in 65 years and older, which would correspond to a case where this population was vaccinated first and is facing a waning of immunity, and 2) a decrease in all age groups, which would correspond either to a scenario where all age groups started the primary vaccination at the same time, and are facing a waning of immunity or to a scenario where the immunity reduction is due to the Delta VOC itself, even for people recently vaccinated. We assumed the booster dose confers an immunity against the Delta VOC equivalent to the one observed against the Alpha VOC which may be a conservative estimate compared to the latest evaluations. (13)

2.4 Vaccination strategies

We present a comparison of booster strategies targeting different age groups, in a context of plateauing primary vaccination. The effectiveness of each strategy is evaluated by the proportion of hospitalizations and deaths avoided, compared to an artificial reference scenario in which all vaccination is stopped on September 1st, 2021. We simulated the administration of 10 million booster doses across five age groups: 30-49 years, 50-64 years, 30 years or older, 50 years or older, and 65 years or older, starting September 1st, 2021, at a rate of 200,000 doses per day. We then analyzed the impact of different levels of increase of virus transmissibility from no increase to a 200%-increase compared to the mean transmissibility level estimated during summer 2021 (R_{eff}=1.0). The model was run from March 2020 to the end of August 2021, then, the different strategies were simulated from September 1st, 2021, to March 1st, 2022.

3 Results

3.1 Prioritization of booster doses

Figure 1 shows the impact on hospital admissions and deaths due to COVID-19 of 10 million booster shots depending on the targeted age groups compared to a baseline scenario of no booster shots and no primary vaccination. If vaccine efficacy is decreased for all age groups (Figures 1-B, 1-D, 1-F), boosting people aged 30 to 49 years would be the most effective strategy, reducing the total number of hospitalizations by 71%, while boosting people aged 65 years or older would lead

to an 18% reduction only (Figure 1-D – left panel). In the other scenarios, in which the efficacy is decreased only for people aged 65 years or older, the strategy targeting the 30-49 age group was the most effective, with a 25% reduction. The strategy targeting the 30+, and the one targeting the 65+, were next with similar levels of reduction in the total number of hospitalizations (Figure 1-C – left panel). Analysis of the mortality impact (Figure 1 – right panel) raises similar conclusions.

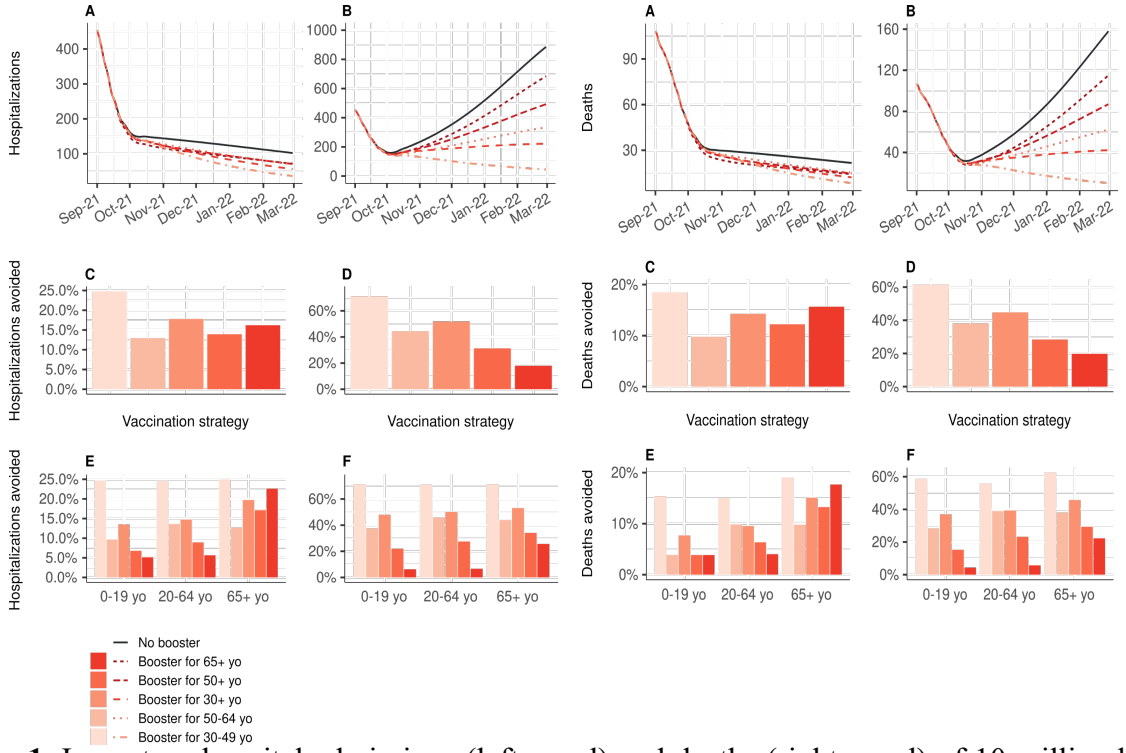


Figure 1. Impact on hospital admissions (left panel) and deaths (right panel) of 10 million booster shots depending on the targeted age groups. **(A, C, E)** Vaccine efficacy decreased only for people aged 65 years and older. **(B, D, F)** Vaccine efficacy decreased for all age groups. **(A, B)** Daily new hospitalizations, left panel, daily deaths, right panel, **(C, D)** proportion of hospitalizations avoided, left panel, proportions of deaths avoided, right panel **(E, F)** proportion of hospitalizations avoided by age group, left panel, proportions of deaths avoided by age-group, right panel.

3.2 Impact of variations of transmissibility

Previous results were obtained considering a 41% increase in viral transmissibility during winter compared to the summer. We also evaluated whether the prioritization of vaccination strategies is sensitive to variations of transmissibility, either due to seasonality, increased contact rate, or new variants. When vaccine efficacy is decreased for all age groups, the ranking of the strategies did not change, and targeting the 30-49 age group remained the most effective boosting strategy to reduce both hospital admissions and deaths. However, when the vaccine efficacy is decreased only for people aged 65 years or older, the ranking varied depending on the assumed viral transmissibility. For more than 20% transmissibility increase, relative to the mean transmissibility in summer (corresponding to R_{eff}=1), targeting the 30-49 age group was the most effective boosting strategy to reduce hospitalizations. Regarding the proportion of deaths avoided, the differences between the strategies were smaller, and a larger degree of transmissibility (>40% increase) was needed to identify the 30-49 age group as the optimal group to target with boosters.

4 Discussion

A growing body of evidence shows that vaccine immunity to COVID-19 wanes with time (5,6) and decreases due to variants. (5,10) Hence, we analyzed the impact of booster strategies targeting different age groups, in a context of decreased efficacy of primary vaccination. When the decrease affects all age groups, the most effective strategy was to boost the immunity of the 30-49 age group. If vaccine efficacy is decreased only for the 65 years and older, the different boosting strategies had

varying effectiveness depending on the level of virus transmissibility in the population. These results suggest that, when deciding which populations to boost, two mechanisms must be considered: immunization of people with a high risk of severe forms of the disease, but also immunization of people with a high risk of transmission, which would reduce the viral circulation and therefore further protect the population at risk of severe consequences. Therefore, in the context of COVID-19, in which these two mechanisms involve different age groups, the optimal boosting strategy might be to target these two populations rather than only people above a certain age. Moreover, one must also consider the level of SARS-CoV-2 transmissibility in the population, which could be affected by seasonality, mitigation measures, or new variants.

These results must be interpreted with caution, as several assumptions and simplifications had to be made. Regarding the vaccine efficacies, we chose to consider distinct values depending on the predominant VOC on the French territory to account for the reduced efficacies against the Delta variant,(10) and we considered two sets of vaccine efficacies against the Delta VOC, one in which the efficacies are slightly lower than against Alpha VOC, and one in which the efficacies are greatly decreased for the protection against infection. (5,6) We considered a first scenario in which the efficacies are decreased only for people aged 65 years and older, and a second scenario, in which the efficacies are decreased for all age groups. However, we did not model the progressive decrease of vaccine efficacy over time but rather changed the parameters at fixed time points. Finally, we assumed that the efficacies obtained after a booster dose were similar to the values with the Alpha variant, which seems to be consistent with some studies. (13)

The analyses were performed in the context of metropolitan France, however, we expect that the conclusions drawn here would hold in many similar settings. The recent emergence of the Omicron VOC(14) illustrates how low vaccination coverage in large parts of the world might increase the risk that new, more potent variants emerge, which could in turn threaten even fully vaccinated countries. This work provides important insight in the matter of booster vaccination, using the context of France as a case study, and can prove useful to inform these critical public health decisions. With a rapid surge of cases in Europe due to the Omicron variant and the threat of newly emerging variants, administering booster doses to active adults could be overall more effective in mitigating the impact on the health system, under expected seasonal conditions of viral transmissibility.

5 References

1. Agence nationale de sécurité du médicament. COVID-19 - Vaccins autorisés [Internet]. 2021. Available from: https://ansm.sante.fr/dossiers-thematiques/covid-19-vaccins/covid-19-vaccins-autorises

2. Santé Publique France. Données relatives aux personnes vaccinées contre la Covid-19 [Internet]. 2021. Available from: https://www.data.gouv.fr/fr/datasets/donnees-relatives-aux-personnes-vaccinees-contre-la-covid-19-1/

3. Santé Publique France. COVID-19 : point épidémiologique du 11 mars 2021 [Internet]. 2021. Available from: https://www.santepubliquefrance.fr/maladies-et-traumatismes/maladies-et-infections-respiratoires/infection-a-coronavirus/documents/bulletin-national/covid-19-point-epidemiologique-du-11-mars-2021

4. Santé Publique France. COVID-19 : point épidémiologique du 16 juillet 2021 [Internet]. 2021. Available from: https://www.santepubliquefrance.fr/maladies-et-traumatismes/maladies-et-infections-respiratoires/infection-a-coronavirus/documents/bulletin-national/covid-19-point-epidemiologique-du-16-juillet-2021

5. Tartof SY, Slezak JM, Fischer H, Hong V, Ackerson BK, Ranasinghe ON, et al. Effectiveness of mRNA BNT162b2 COVID-19 vaccine up to 6 months in a large integrated health system in the USA: a retrospective cohort study. The Lancet. 2021 Oct;398(10309):1407–16.

6. Goldberg Y, Mandel M, Bar-On YM, Bodenheimer O, Freedman L, Haas EJ, et al. Waning Immunity after the BNT162b2 Vaccine in Israel. N Engl J Med. 2021 Oct 27;NEJMoa2114228.

7. Haute Autorité de Santé. Avis n° 2021.0061/AC/SEESP du 23 août 2021 du collège de la HAS relatif à la définition des populations à cibler par la campagne de rappel vaccinal chez les personnes ayant eu une primovaccination complète contre la Covid-19 [Internet]. 2021. Available from: https://www.has-sante.fr/jcms/p_3283044/fr/avis-n-2021-0061/ac/seesp-du-23-aout-2021-du-college-de-la-has-relatif-a-la-definition-des-populations-a-cibler-par-la-campagne-de-rappel-vaccinal-chez-les-personnes-ayant-eu-une-primovaccination-complete-contre-la-covid-19

8. Roux J, Massonnaud C, Colizza V, Cauchemez S, Crépey P. Impact of national and regional lockdowns on COVID-19 epidemic waves: Application to the 2020 spring wave in France [Internet]. 2021 Jun [cited 2021 Nov 3] p. 2021.04.21.21255876. Available from: https://www.medrxiv.org/content/10.1101/2021.04.21.21255876v2

9. Bager, P. and Wohlfahrt, J. and Rasmussen, M. and Albertsen, M and Grove Krause, T. Hospitalisation associated with SARS-CoV-2 delta variant in Denmark. Lancet Infect Dis. 2021 Sep;

10. Sheikh A, McMenamin J, Taylor B, Robertson C. SARS-CoV-2 Delta VOC in Scotland: demographics, risk of hospital admission, and vaccine effectiveness. Lancet. 2021 Jun;397(10293):2461–2.

11. Collin A, Hejblum BP, Vignals C, Lehot L, Thiébaut R, Moireau P, et al. Using Population Based Kalman Estimator to Model COVID-19 Epidemic in France: Estimating the Effects of Non-Pharmaceutical Interventions on the Dynamics of Epidemic [Internet]. 2021 Jul [cited 2021 Oct 25]. Available from: https://www.medrxiv.org/content/10.1101/2021.07.09.21260259v3

12. Carrat F, de Lamballerie X, Rahib D, Blanché H, Lapidus N, Artaud F, et al. Antibody status and cumulative incidence of SARS-CoV-2 infection among adults in three regions of France following the first lockdown and associated risk factors: a multicohort study. Int J Epidemiol. 2021 Nov 10;50(5):1458–72.

13. Barda N, Dagan N, Cohen C, Hernán MA, Lipsitch M, Kohane IS, et al. Effectiveness of a third dose of the BNT162b2 mRNA COVID-19 vaccine for preventing severe outcomes in Israel: an observational study. The Lancet [Internet]. 2021 Oct 29 [cited 2021 Nov 8];0(0). Available from: https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(21)02249-2/abstract

14. WHO. Classification of Omicron (B.1.1.529): SARS-CoV-2 Variant of Concern [Internet]. 2021. Available from: https://www.who.int/news/item/26-11-2021-classification-of-omicron-(b.1.1.529)-sars-cov-2-variant-of-concern

Acknowledgments

The study was partially funded by the French national research agency (ANR) through the project SPHINX-17-CE36-0008-0. The authors thank Daniel Levy-Bruhl and Simon Cauchemez for fruitful discussions regarding this work.

Authors' contributions

CRM, JR, and PC designed and planned the study. PC developed the initial mathematical model. CRM, JR, and PC further developed the model for this study. CRM and JR performed the analyses. CRM, JR, and PC audited the code. CRM wrote the first draft of the manuscript. JR, PC, and VC critically commented on the primary and sensitivity analyses. All authors contributed to revisions of the manuscript.

Conflicts of interest

PC has received consulting fees from Sanofi-Pasteur for projects unrelated to COVID-19 and unrelated to this project. All other authors declare no potential conflicts of interest.

QUANTIFYING HORIZONTAL PLEIOTROPY IN HUMAN GENETIC VARIATION USING GWAS SUMMARY STATISTICS

Marie Verbanck [1, 2, @]

1 : EA 7537 - BioSTM (Biostatistique, Traitement et Modélisation des données biologiques)

Université de Paris

2 : Institute for Personalized Medicine, Icahn school of medicine at Mount Sinai, NYC

Invitation de la part de la Société Française de Biométrie

Horizontal pleiotropy, where one variant has independent effects on multiple traits, is crucial for our understanding of the genetic architecture of human traits. Here we have developed a method to model and quantify horizontal pleiotropy using publicly-available summary statistics from published genome-wide association studies. We developed a method to quantify horizontal pleiotropy through a variant-level pleiotropy quantification and perform validation using simulations. When applied to 1,564 medical phenotypes measured in 337,119 humans from the UK Biobank, our pleiotropy score detected a significant excess of horizontal pleiotropy. This signal of horizontal pleiotropy was pervasive throughout the human genome and across a wide range of phenotypes, but was especially prominent in regions of high linkage disequilibrium and among highly polygenic phenotypes. We identified thousands of loci with extreme horizontal pleiotropy, a majority of which had never been reported in any published GWAS. These findings suggest that horizontal pleiotropy is pervasive in human genetic variation, and has significant implications for our understanding of the genetic architecture of complex traits and disease.

SFdS - MALIA - SSFAM

Analyse statistique de la topologie de Mapper pour des filtres stochastiques

Mathieu Carrière[1] & Bertrand Michel[2]

[1] *mathieu.carriere@inria.fr, Inria centre at Université Côte d'Azur, 2004, route des Lucioles, BP 93, 06902 Sophia Antipolis Cedex, France*
[2] *bertrand.michel@ec-nantes.fr, Laboratoire de Mathématiques Jean Leray, UMR_C 6629, Ecole Centrale de Nantes, France*

Résumé. Mapper est un descripteur commun issu de l'analyse topologique de données, et utilisé dans une grande variété d'applications de la science des données, allant de la biologie algorithmique à la visualisation. Ceci dit, sa statistique, et plus précisément sa stabilité et vitesse de convergence vers sa version limite, appelée espace de Reeb, est encore mal comprise et constitue une question ouverte de la communauté. Dans cette présentation, nous proposons des bornes supérieures sur l'espérance de la distance entre Mapper et espace de Reeb. Notre approche inclut en particulier le cas où le filtre utilisé dans le calcul de Mapper est lui-même stochastique, comme c'est le cas pour les fonctions propres d'une ACP.

Mots-clés. Mapper, Espace de Reeb, Complexe Simplicial, Analyse Topologique de Données, Vitesse de Convergence, Région de Confiance.

Abstract. Mapper is a common descriptor used in Topological Data Analysis (TDA), with plenty of applications in various fields of science, such as computational biology and data visualization, among others. The stability and quantification of the rate of convergence of the Mapper to its limit version called Reeb space is a difficult yet important open question in the field of TDA. In this presentation, we provide risk bounds for estimating the Reeb space using Mapper. Our approach applies in particular to the setting where the filter function used to compute Mapper is potentially estimated from data, such as the eigenfunctions of PCA.

Keywords. Mapper, Reeb Space, Simplicial Complex, Topological Data Analysis, Convergence Rate, Confidence Region.

1 Introduction

Mapper est un outil (introduit par Singh, Mémoli et Carlsson (2007)) de clustering et de visualisation des données issu de la topologie algorithmique. Dans sa forme la plus générale, il se calcule en recouvrant le jeu de données par une famille d'ensembles, et

en caractérisant ce recouvrement au moyen de son *nerf*, c'est-à-dire du complexe simplicial, ou graphe généralisé, dont les sommets sont en bijection avec les éléments du recouvrement, et dont les k-simplexes sont donnés par les intersections de ces éléments. La représentation de ces intersections aboutit en effet à une visualisation des données qui préserve les phénomènes continus sous-jacents aux données, à l'opposé, par exemple, des méthodes traditionnelles type t-SNE ou UMAP, qui ne calculent qu'un plongement dans $\mathbb{R}^2$ ou $\mathbb{R}^3$ sans tenir compte de la connectivité, et donc de la géométrie et de la topologie, du jeu de données. Mettre en valeur la topologie du jeu de données au cours de la visualisation est particulièrement important en biologie algorithmique par exemple, pour laquelle de nombreux jeux de données sont dirigés par des processus biologiques sous-jacents continus (tels que la différentiation de cellules souches, le cycle cellulaire, ou les interactions inter- et intra-chromosomiques), qui sont pour la plupart détectables par la topologie. Voir par exemple Rizvi et coauteurs (2017).

Une difficulté associée au Mapper est le choix de ses paramètres. En effet, il est peu souvent clair de choisir le recouvrement à utiliser, en particulier si les données sont peu structurées, ou bien si elles ne sont fournies que par l'intermédiaire de leurs distances respectives. Pour ces raisons, le recouvrement utilisé pour le calcul du Mapper est très souvent paramétré par une fonction appelée *filtre*, à valeurs dans un espace métrique potentiellement différent de celui du jeu de données. Un recouvrement peut ainsi être défini de manière implicite en recouvrant l'image du filtre, ce qui est plus simple à priori car celle-ci est contrôlée par l'utilisateur, et en calculant les antécédents de chacun des éléments de ce nouveau recouvrement. Afin d'adapter le plus possible ce recouvrement implicite au jeu de données, il est aussi fréquent de séparer les composantes connexes de chacun des antécédents en différents éléments du recouvrement implicite. Cette séparation peut se calculer via un algorithme de clustering, dont le choix est laissé à la charge de l'utilisateur.

Même si l'utilisation de filtres permet de paramétrer implicitement et plus facilement le recouvrement à utiliser, cela ne permet toutefois pas de s'assurer de la pertinence des structures topologiques identifiées par Mapper. Ceci est d'autant plus problématique qu'il a été observé de nombreuses fois que Mapper souffrait d'une grande sensibilité vis-à-vis du filtre et du recouvrement de son image utilisés pour son calcul, au sens où leur choix peut parfois conduire à l'apparition de structures topologiques artificielles et non désirées. **Le but de cette présentation est de discuter la définition et le calcul de régions de confiance associées aux structures topologiques identifiées par Mapper.**

2 Définitions

Le Mapper est un complexe simplicial, dont la définition est rappelée ci-dessous.

Définition. Un *complexe simplicial* $\mathcal{C}$ est une famille de sous-ensembles d'un ensemble global $V(\mathcal{C}) = \{v_1, \ldots, v_n\}$, appelé l'ensemble des *sommets du complexe*, qui vérifie les propriétés suivantes:

- Pour tout $\sigma \subseteq V(\mathcal{C})$ appartenant à $\mathcal{C}$, on a $\tau \subseteq \sigma \Rightarrow \tau \in \mathcal{C}$,
- Pour tout $\sigma, \sigma' \in \mathcal{C}$, on a soit $\sigma \cap \sigma' = \emptyset$, soit $\sigma \cap \sigma' \in \mathcal{C}$.

Un complexe simplicial peut être construit de plusieurs manières différentes. L'une d'entre elles, particulièrement fréquente en topologie algébrique, implique de caractériser les intersections d'éléments d'un recouvrement.

Définition. Un *recouvrement* $\mathcal{U}$ d'un espace topologique X est une famille de sous-ensembles de X telle que $X \subseteq \bigcup_{U \in \mathcal{U}} U$. Le *nerf* d'un recouvrement $\mathcal{U}$, désigné par $\mathcal{N}(\mathcal{U})$, est le complexe simplicial dont les sommets sont en bijection avec les éléments de $\mathcal{U}$, et qui vérifie $\sigma = \{U_1, \ldots, U_m\} \in \mathcal{N}(\mathcal{U}) \Rightarrow \bigcap_{i=1}^{m} U_i \neq \emptyset$.

Finalement, le Mapper est défini comme l'adaptation du nerf dans le cas où $X = \hat{X}_n$ est un espace métrique fini et où le recouvrement est obtenu via les antécédents d'une fonction filtre continue $f : X \to Y$.

Définition. Soient X et Y deux espaces métriques, $\hat{X}_n = \{x_1, \ldots, x_n\} \subset X$ un nuage de points, et $f : \hat{X}_n \to Y$ une fonction continue appelée *filtre*. Soit $\delta > 0$, et $\mathcal{V}$ un recouvrement de image(f). Le *Mapper* associé à $(\hat{X}_n, f, \delta, \mathcal{V})$, désigné par $\mathrm{M}_{f,\delta}(\hat{X}_n, \mathcal{V})$, est le complexe simplicial défini comme le nerf du recouvrement $\mathcal{U}$, lui même défini par:

$$\mathcal{U} = \bigcup_{V \in \mathcal{V}} \mathrm{cc}_\delta(f^{-1}(V)),$$

où cc_δ désigne l'ensemble des composantes connexes, identifiées avec un clustering hiérarchique de paramètre δ.

Dans la section suivante, nous allons présenter des résultats permettant de caractériser la pertinence des structures topologiques identifiées par Mapper. Pour cela, nous allons les comparer à la version limite du Mapper, obtenue avec un recouvrement contenant des éléments infiniment fins, i.e., des singletons, et appelée *espace de Reeb*. Celui-ci peut se aussi se décrire mathématiquement en tant qu'espace quotient avec la définition suivante.

Définition. Soit X, Y deux espaces topologiques et $f : X \to Y$ une fonction continue. L'*espace de Reeb* associé à f, désigné par $\mathrm{R}_f(X)$, est l'espace quotient:

$$\mathrm{R}_f(X) = X/\sim_f,$$

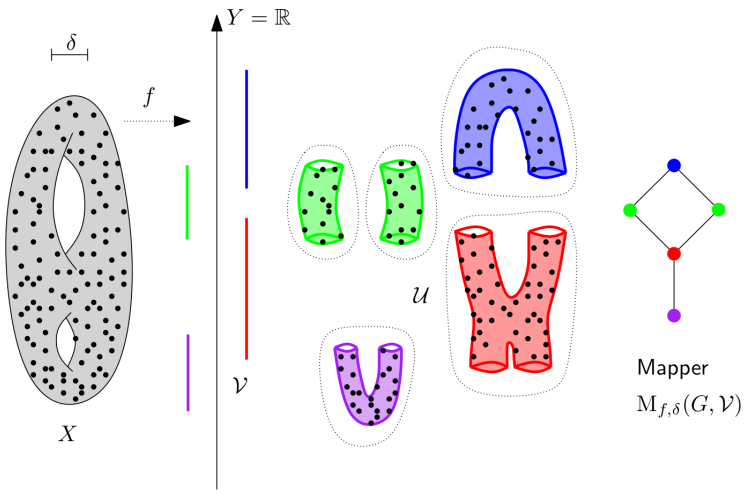


Figure 1: Exemple de calcul de Mapper sur l'échantillonnage d'un double tore avec la fonction hauteur.

où $\sim_f$ est la relation d'équivalence $x \sim_f x' \Leftrightarrow f(x) = f(x')$ et x, x' sont dans la même composante connexe de $f^{-1}(f(x)) = f^{-1}(f(x'))$.

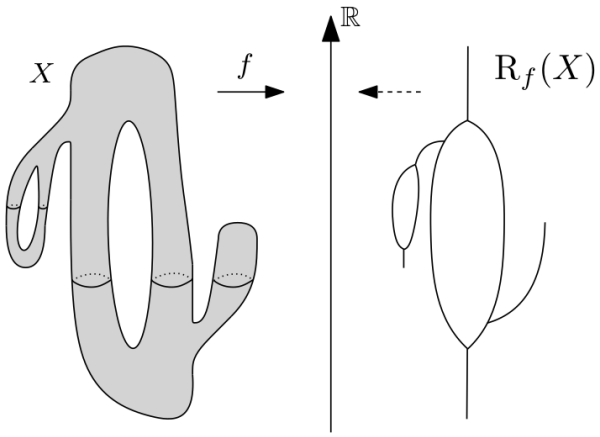


Figure 2: Exemple de calcul d'espace de Reeb.

3 Résultats

Dans cette section, nous présentons les deux résultats principaux de cette contribution, associés à Carrière, Oudot et Michel (2018) et Carrière et Michel (2019), qui bornent

supérieurement l'espérance de la distance séparant Mapper et espace de Reeb, pour un certain choix de paramètres, et sous l'hypothèse que les points sont tirés suivant une mesure de probabilité supposée (a, b)-standard. On suppose aussi dans la suite que X est une sous-variété Riemannienne suffisamment régulière plongée dans $\mathbb{R}^D$.

Définition. Une mesure de probabilité μ est appelée (a, b)-standard si $\mu(B(x, r)) \geq \min\{1, ar^b\}$, pour tout $x \in X$ and $r > 0$, où $B(x, r) = \{y \in \mathbb{R}^D \, : \, \|y - x\| \leq r\}$.

3.1 Cas $Y = \mathbb{R}$

Quand la fonction filtre est à valeurs réelles, on peut comparer Mapper et espace de Reeb au moyen de la *distance bottleneck* d_B. De plus, on peut définir le recouvrement de image(f) avec une famille d'intervalles ayant tous la même longueur, ou *résolution* $r > 0$, et le même pourcentage d'intersection, ou *gain* $g \in]0, 1[$.

Théorême. Soit $\hat{X}_n$ un nuage de points obtenu par échantillonnage d'une mesure (a, b)-standard μ. Soit $\delta_n = 4\left(\frac{2\log n}{an}\right)^{1/b}$, $g_n \in \left(\frac{1}{3}, \frac{1}{2}\right)$, $r_n = \frac{c\delta_n}{g_n}$, où c est la constante de Lipschitz de la fonction f. On a l'inégalité suivante:

$$\mathbb{E}\left[d_B\left(\mathrm{M}_{f,\delta_n}(\hat{X}_n, \mathcal{I}(g_n, r_n)),\ \mathrm{R}_f(X)\right)\right] \leq C\ \left(\frac{\log n}{n}\right)^{1/b},$$

où $C > 0$ est une constante qui ne dépend que de a, b et X.

3.2 Cas général

Dans le cas général, Mapper et espace de Reeb ne peuvent être comparés qu'au moyen d'une pseudo-métrique, la *distance de Gromov-Hausdorff* d_{GH}.

Théorême. Soit $\hat{X}_n$ un nuage de points obtenu par échantillonnage d'une mesure (a, b)-standard μ. Soit δ_n défini comme dans le théorême précédent, et $\mathcal{V}$ un recouvrement (potentiellement stochastique) de image(f). On a l'inégalité suivante:

$$\mathbb{E}\left[d_{GH}(\mathrm{M}_{f,\delta_n}(\hat{X}_n, \mathcal{V}), \mathrm{R}_f(X))\right] \leq 5 \cdot \mathbb{E}\left[\mathrm{res}(\mathcal{V})\right] + C\omega\left(\frac{\log(n)^2}{n}\right)^{1/b},$$

où $C > 0$ est une constante qui ne dépend que de a, b et X, et res désigne la résolution de $\mathcal{V}$, c'est-à-dire le diamètre de son plus grand élément.

3.3 Exemple numérique

Les résultats présentés précédemment permettent de calculer des régions de confiance pour Mapper. Nous présentons ci-dessous un exemple d'application, lorsque Mapper est calculé sur un ensemble d'images (vues comme des vecteurs Euclidiens), et lorsque ses paramètres sont estimés automatiquement via les résultats précédents. La confiance associée au cycle apparaissant dans le Mapper (dont le calcul est possible grâce aux théorêmes précédents) est suffisament élevée pour considérer la structure comme pertinente et non artificielle.

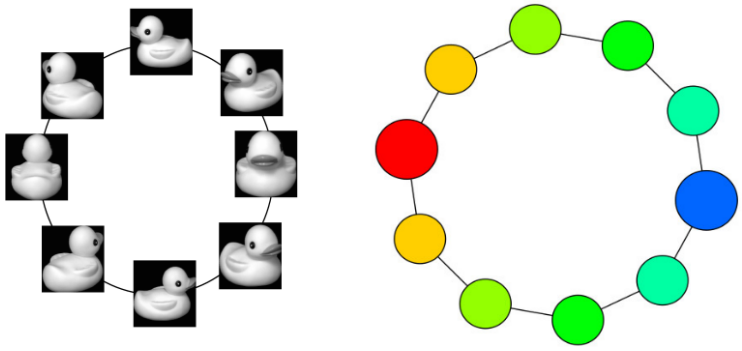


Figure 3: Exemple de calcul de Mapper sur un ensemble d'images.

Bibliographie

Mathieu Carrière, Bertrand Michel, and Steve Oudot. *Statistical analysis and parameter selection for Mapper.* Journal of Machine Learning Research, 19(12):1–39, 2018.

Mathieu Carrière and Bertrand Michel. *Statistical analysis of Mapper for stochastic and multivariate filters.* In CoRR. arXiv:1912.10742, 2019.

Gurjeet Singh, Facundo Mémoli, and Gunnar Carlsson. *Topological methods for the analysis of high dimensional data sets and 3D object recognition.* In 4th Eurographics Symposium on Point-Based Graphics (SPBG 2007), pages 91-100. The Eurographics Association, 2007.

Abbas Rizvi, Pablo Cámara, Elena Kandror, Thomas Roberts, Ira Schieren, Tom Maniatis, and Raúl Rabadán. *Single-cell topological RNA-seq analysis reveals insights into cellular differentiation and development.* Nature Biotechnology, 35:551–560, 2017.

Graph Neural Networks on Large Random Graphs

Samuel Vaiter [1]

[1]Laboratoire J.A. Dieudonné – Centre National de la Recherche Scientifique, Université Côte d'Azur (UCA) – France

Résumé

Graph Neural Networks (GNNs) are deep architectures defined over graph data that have garnered a lot of attention in recent years. In this talk, I will give some insight on how such architectures behave on random graphs. I will first give non-asymptotic convergence bounds of GNNs toward "continuous" equivalents as the number of nodes grows. Then, I will show their stability to small deformations of the underlying random graph model, a crucial property in traditional CNNs. Finally, I will discuss universality and approximation power with respect to traditional graph tools. This is joint work with Nicolas Keriven and Alberto Bietti.

Apprentissage pour l'amélioration de surfaces définies par des nuages de points

Julie Digne [1]

[1]Laboratoire d'InfoRmatique en Image et Systèmes d'information (LIRIS) – Université Claude Bernard Lyon 1, Centre National de la Recherche Scientifique : UMR5205 – France

Résumé

Les formes 3D acquises par un scanner laser sont souvent données sous la forme d'un ensemble de points qui peuvent être bruités ou avoir une résolution limitée et des données manquantes. Dans cet exposé, j'expliquerai comment la prise en compte de l'auto-similarité des surfaces peut aider à surmonter ces problèmes. En effet, de nombreuses surfaces présentent des textures et des structures géométriques répétées qui peuvent être exploitées pour améliorer l'acquisition de surface. En concevant des descripteurs de forme locaux et en agrégeant leurs informations, il est possible de débruiter une surface ou d'en calculer la super-résolution basée sur des données. Les nuages de points peuvent aussi comporter des parties non 2-variété : par exemple, les câbles de rue peuvent être considérés comme des courbes, en fonction de la précision d'acquisition. Nous définissons des Local Probing Fields adaptés à ce cas particulier et les analysons à l'aide d'algorithmes d'apprentissage de dictionnaires. Cela nous permet de revoir diverses tâches de traitement de forme telles que le débruitage et le rééchantillonnage de forme.

PLENIERE : Aurélien Bellet

Differentially Private Machine Learning

Aurélien Bellet

INRIA, Lille

Personal data is being collected at an unprecedented scale by businesses and public organizations, driven by the progress of data science and machine learning. While such data can be turned into useful knowledge about the global population by computing aggregate statistics or training machine learning models, this can also lead to undesirable disclosure of personal information. We must therefore deal with two conflicting objectives: maximizing the utility of data while protecting the privacy of individuals whose data is used in the analysis.

In this talk, I will present differential privacy (DP), a statistical definition of privacy which comes with rigorous guarantees as well as an algorithmic framework that allows the design of practical privacy-preserving algorithms. I will then discuss the application of DP to machine learning, and some related open questions.

Graphes – Réseaux

Learning common structures in a collection of networks

Saint-Clair Chabert-Liddell & Pierre Barbillon & Sophie Donnet

Université Paris-Saclay, AgroParisTech, INRAE, UMR MIA Paris-Saclay, 75005, Paris, France.
{*saint-clair.chabert-liddell,sophie.donnet,pierre.barbillon* }*@agroparistech.fr*

Résumé. Soit une collection de réseaux constituée d'un ensemble de réseaux qui ne partagent pas de nœuds mais qui décrivent le même type d'interactions observées dans des situations ou des contextes différents. Une hypothèse est que les réseaux de la collection partagent une structure commune puisque la nature des interactions est la même. Nous proposons de nous appuyer sur le modèle à blocs stochastiques (SBM) pour identifier la structure commune de la collection. Le SBM est un modèle probabiliste qui suppose l'existence de variables latentes représentant les groupes de nœuds (blocs) du réseau et dont les paramètres fournissent une description succincte de la structure du réseau à l'échelle mésoscopique. Nous appelons *colSBM* notre extension du SBM à une modélisation conjointe d'une collection de réseaux. Les réseaux de la collection sont supposés être des réalisations indépendantes de différents SBM, qui partagent - à travers des paramètres communs - la même structure de connectivité, éventuellement aux proportions de blocs et/ou un facteur de densité près.

Les paramètres du modèle sont estimés et les blocs latents sont retrouvés à l'aide d'un algorithme EM variationnel. Nous utilisons un critère ad-hoc, basé sur la vraisemblance de classification intégrée pour sélectionner le nombre de blocs et évaluer l'adéquation du consensus trouvé entre les structures des différents réseaux. Des applications sur des réseaux trophiques sont présentées afin d'illustrer l'intérêt de ces modèles.

Mots-clés. Réseaux d'interactions, modèles à blocs stochastiques, algorithmes variationnels.

Abstract. Let a collection of networks consists of a set of networks which do not share nodes but which describe the same kind of interactions observed in different situations or contexts. An hypothesis is that the networks in the collection share a common structure since the nature of the interactions is the same.

We propose to build on the popular stochastic block model (SBM) to identify the common structure in the collection. The SBM is a probabilistic model that assumes the existence of latent variables representing the groups of nodes (blocks) of the network and the parameters of which provide a succinct description of the structure of the network at the mesoscopic scale. We call *colSBM* our extension of the SBM to a joint modeling of a collection of networks. The networks in the collection are assumed to be independent realizations of different SBMs, which share –through common parameters– the same connectivity structure, possibly up to the block proportions and/or a density factor.

The model parameters are estimated and the latent blocks are retrieved using a variational EM algorithm. We use an ad-hoc criterion, based on the integrated classification likelihood (ICL) to select the number of blocks and to evaluate the adequacy of the consensus found between the structures of the different networks. Some applications are dealt with in order to illustrate the relevance of the proposed models.

Keywords. Interaction networks, stochastic block models, variational algorithms.

1 Data motivation and the stochastic block model

Consider a collection of M independent networks where each network indexed by m involves its own n_m nodes. The networks are encoded into their adjacency matrices $(X^m)_{m\in\{1,\dots,M\}}$ such that: $\forall m \in \{1,\dots,M\}$, $\forall (i \neq j) \in \{1,\dots,n_m\}^2$,

$$\begin{cases} X_{ij}^m = 0 & \text{if no interaction is observed between species } i \text{ and } j \text{ of network } m \\ X_{ij}^m \neq 0 & \text{otherwise.} \end{cases}$$

We consider that the networks consist of binary interactions: $X_{ij}^m \in \mathcal{K} = \{0,1\}$, $\forall (m,i,j)$. No self-interaction is considered. Besides, for the sake of simplicity, we assume that all the networks are directed. The extension to undirected networks i.e. such that $X_{ij}^m = X_{ji}^m$ for any $i \neq j$ is straightforward. $\mathbf{X} = (X^1,\dots,X^m)$ denotes the collection of adjacency matrices.

A first ecological example: three stream food webs As a first example, we consider the collection of three stream food webs from Thompson and Townsend (2003). The three networks collected respectively in Martins (Maine USA), Cooper and Herlzier (North-Carolina, USA) involve respectively 105, 58 and 71 species resulting in 343, 126 and 148 binary edges respectively. Classically, the food web edges represent directed trophic links showing the energy flow ie. $X_{ij}^m = 1$ if species j preys on species i, with no reciprocal interactions. When aiming at unraveling the structure of these networks the Stochastic Block Model (SBM) is an interesting tool which has proven its high flexibility by encompassing a large variety of structures (see Allesina and Pascual, 2009, for the particular case of food webs). When dealing with three networks, the standard strategy that we describe below, is to fit separately one SBM per network.

Separate SBM (sepSBM) The SBM introduces clusters of nodes and assumes that the interaction between two nodes is driven by the clusters the nodes belong to. More precisely, for network m, let the n_m nodes be divided into Q_m clusters. Let $Z^m =$

$(Z_i^m, \ldots, Z_{n_m}^m)$ be independent latent random variables such that $Z_i^m = q$ if node i of network m belongs to cluster q with $q \in \{1, \ldots, Q_m\}$ and

$$P(Z_i^m = q) = \pi_q^m \tag{1}$$

where $\pi_q^m > 0$ and $\sum_{q=1}^{Q_m} \pi_q^m = 1$. Given the latent variables Z^m, the X_{ij}^m's are assumed to be independent and distributed as

$$X_{ij}^m | Z_i^m = q, Z_j^m = r \sim \mathcal{B}(\cdot; \alpha_{qr}^m), \tag{2}$$

where $\mathcal{B}$ is the Bernoulli distribution. Equations (1) and (2) define the SBM model and we will now use the following short notation:

$$X^m \sim \text{SBM}_{n_m}(Q_m, \boldsymbol{\pi}^m, \boldsymbol{\alpha}^m). \qquad (sep\text{-}SBM)$$

where n_m is the number of nodes, Q_m is the number of blocks and $\boldsymbol{\pi}^m = (\pi_q^m)_{q=1,\ldots,Q_m}$ is the vector of their proportions. The $Q \times Q$ matrix $\boldsymbol{\alpha}^m = \left(\alpha_{qr}^m\right)_{q,r=1,\ldots,Q_m}$ denotes the connection parameters i.e. the parameters of the emission distribution: $\alpha_{qr}^m \in (0,1)$. In the *sep-SBM* model, each network m is modeled independently with its own parameters $(\boldsymbol{\pi}^m, \boldsymbol{\alpha}^m)$.

Application to the three stream food webs We fit the *sepSBM* on the 3 stream food webs, respectively referred to as Martins, Cooper and Herlzier. To do so, we use the `sbm` R-package (Chiquet et al., 2021; Leger et al., 2020) on each network, which implements a variational version of the EM algorithm to estimate the parameters and selects the number of clusters Q_m using a penalized likelihood criterion (ICL). We obtain respectively $\widehat{Q}_1 = 5$ blocks for Martins, $\widehat{Q}_2 = 3$ blocks for Cooper and $\widehat{Q}_3 = 4$ blocks for Herlzier. The adjacency matrices of the food webs reordered by block membership are plotted in Figure 1. The connectivity structures of these three networks seem to have a lot of similarities. We then present in the next section probabilistic models designed to introduce structure consensus.

2 Joint Modeling of a Collection of Networks

A collection of i.i.d. SBM The first model we propose is the most constrained one and assumes that the networks are independent realizations of the same Q-blocks SBM model with identical parameters. The so-called *iid-colSBM* states that:

$$X^m \sim \text{SBM}_{n_m}(Q, \boldsymbol{\pi}, \boldsymbol{\alpha}), \qquad \forall m = 1, \ldots M, \qquad (iid\text{-}colSBM)$$

where $\forall (q,r) \in \{1, \ldots, Q\}^2$, $\alpha_{qr} \in (0,1)$, $\pi_q \in]0,1]$ and $\sum_{q=1}^{Q} \pi_q = 1$. Assuming that the blocks are represented in the same proportions in each network is a strong assumption that may lead to the model being of little practical use. The following model relaxes this assumption.

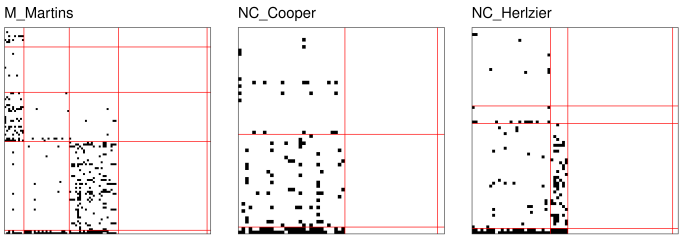


Figure 1: Matricial view of 3 stream food webs. The species are reordered by blocks and blocks are ordered by expected out-degrees to emulate the trophic levels (bottom to top and right to left). The blocks have been obtained by fitting a SBM on each network separately.

A collection of networks with varying block sizes π-$colSBM$ assumes that the networks share a common connectivity structure encoded in $\boldsymbol{\alpha}$, but that the proportions of the blocks are specific to each network. Moreover, by allowing some block proportions to be null, the model encompasses situations where some blocks may not be represented in all the realized networks. More precisely, for $m \in \{1, \dots, M\}$, the X^m are independent and

$$X^m \sim \text{SBM}_{n_m}(Q, \boldsymbol{\pi}^m, \boldsymbol{\alpha}) \qquad \forall m = 1, \dots M. \qquad (\pi\text{-}colSBM)$$

where $\alpha_{qr} \in (0, 1)$. Moreover, we assume that $\sum_{q=1}^{Q} \pi_q^m = 1, \forall m \in \{1, \dots, M\}$ where $\pi_q^m \in [0, 1]$ and if $\pi_q^m = 0$ then the block q is not represented in network m. In addition, we assume that for any $q \in \{1, \dots, Q\}$, $\exists m \in \{1, \dots, M\}$ such that $\pi_q^m > 0$, meaning that any block q is represented in at least one network. Let S be the $M \times Q$ support matrix such that $\forall (m, q)$ $S_{mq} = \mathbf{1}_{\pi_q^m > 0}$. Then, the set of admissible supports is $\mathcal{S}_Q := \left\{ S \in \mathcal{M}_{M,Q}(\{0, 1\}), \sum_{m=1}^{M} S_{mq} \geq 1 \quad \forall q = 1, \dots, Q \right\}$. The number of parameters is bounded by $M(Q-1) + Q^2$, this upper bound corresponding to the case where all the blocks are represented in all the networks, but with varying proportions.

A collection of networks with varying density (δ-$colSBM$) The iid-$colSBM$ can be relaxed in another direction, assuming that the M networks exhibit similar intra- and inter- blocks connectivity patterns but with different densities. More precisely, let $\delta_m \in \mathbb{R}$ be a density parameter for network m. The δ-$colSBM$ is defined as follows:

$$X^m \sim \text{SBM}_{n_m}(Q, \boldsymbol{\pi}, \delta_m \boldsymbol{\alpha}). \qquad (\delta\text{-}colSBM)$$

with $\pi_q > 0, \forall q = 1, \dots, Q$, $\sum_{q=1}^{Q} \pi_q = 1$. Moreover $\forall (m, q, r)$, $\delta_m \alpha_{qr} \in (0, 1)$ and one of the density parameter equal to one ($\delta_1 = 1$) for identifiability purpose. This model mimics different intensity of connections between networks.

Collection of networks with varying block sizes and density ($\delta\pi$-$colSBM$) Finally, we propose to mix the models π-$colSBM$ and δ-$colSBM$ to obtain a more complex one which allows each network to have its own block proportions $\boldsymbol{\pi}^m$ as well as a specific scale density parameter δ_m. Then, the $(X^m)_{m\in\{1,\dots,M\}}$ are independent and

$$X^m \sim \text{SBM}_{n_m}(Q, \boldsymbol{\pi}^m, \delta_m \boldsymbol{\alpha}), \qquad (\delta\pi\text{-}colSBM)$$

where $\forall (m,q,r)$, $\delta_m \alpha_{qr} \in (0,1)$, $\delta_1 = 1$, $\pi_q^m \geq 0$ and $\sum_{q=1}^{Q} \pi_q^m = 1$.

3 Inference and model selection

The likelihood corresponding to the joint model is not tractable in practice. A well-proven approach to handle this problem for the inference of the SBM is to rely on a variational version of the EM algorithm. This is done by maximizing a lower (variational) bound of the log-likelihood of the observed data (Daudin et al., 2008). There are two model selection issues. First, under a fixed $colSBM$, we aim to choose Q and determine the support matrix S for π-$colSBM$ and $\delta\pi$-$colSBM$. This task is tackled by introducing a penalized likelihood criterion (ICL). Second, the comparison of the the $colSBM$ models –each one introducing various degrees of consensus between the networks– with the sep-SBM – which assumes that each network has its own structure.

4 Application

In Figure 2, the inferred structures according to the four proposed joint models are presented. The ICL criterion favors any of these joint models over the separated SBM. They help unravel a common structure. Moreover, a more detailed structure is detected in the smaller networks by using the information from the others.

References

Allesina, S. and Pascual, M. (2009). Food web models: a plea for groups. *Ecology letters*, 12(7):652–662.

Chiquet, J., Donnet, S., and Barbillon, P. (2021). *sbm: Stochastic Blockmodels*. R package version 0.4.3.

Daudin, J.-J., Picard, F., and Robin, S. (2008). A mixture model for random graphs. *Statistics and computing*, 18(2):173–183.

Leger, J.-B., Barbillon, P., and Chiquet, J. (2020). *blockmodels: Latent and Stochastic Block Model Estimation by a 'V-EM' Algorithm*. R package version 1.1.4.

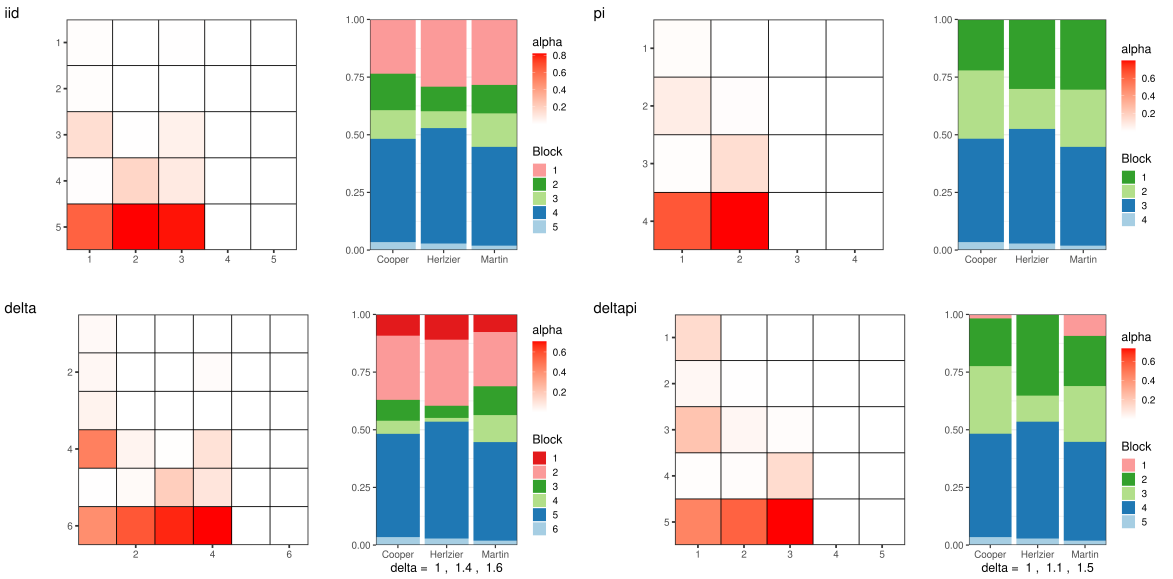


Figure 2: **Estimated structure of the collection of 3 stream food webs with the four *colSBM* models**. For each network, the matrix on the left is the estimated parameter connectivity $\widehat{\boldsymbol{\alpha}}$, the barplot on the right depicts $\widetilde{\boldsymbol{\pi}}^{(m)}$. The ordering is done by trophic level from bottom to top and right to left. For (δ-$\delta\pi$)*colSBM*s we give $\hat{\delta}$ below the barplot.

Thompson, R. M. and Townsend, C. R. (2003). Impacts on stream food webs of native and exotic forest: an intercontinental comparison. *Ecology*, 84(1):145–161.

Clustering networks with textual edges by combining the Embedded Topic Model and the Stochastic Block model. Variational inference and derivation of a model selection criterion.

Rémi Boutin [1] & Pierre Latouche [1] & Charles Bouveyron [2]

[1] *Université de Paris, MAP 5, UMR 8145, Paris, France*
remi.boutin@u-paris.fr *pierre.latouche@maths.cnrs.fr*
[2] *Université Côte d'Azur, Inria, CNRS, Laboratoire J.A. Dieudonné, Maasai team*
charles.bouveyron@univ-cotedazur.fr

Résumé. Les réseaux de communications (courriels, réseaux sociaux, objects connectés) sont de plus en plus présents autour de nous et leur analyse devient prépondérante pour le fonctionnement de notre modèle numérique. De nombreux réseaux de communication présentent des arêtes comportant du texte dont l'information n'est pas utilisée dans les méthodes classiques. Dans ce papier, nous proposons le ETSBM (embedded topic for the stochastic block model) pour détecter simultanément les partitions des nœuds et les thèmes présents dans les textes. Ce travail étend le SBM (stochastic block model) et le ETM (embedded topic model) grâce à un algorithme VBEM (variational-Bayes expectation-maximization) et une descente de gradient stochastique. Nous proposons également un critère de sélection de modèle. Notre méthodologie est évaluée sur des données simulées ainsi que sur des données réelles.

Mots-clés. embeddings, topic Models, clustering de nœuds, modèles génératifs, STBM, ETM, LDA, SBM.

Abstract. Communication networks (emails, social networks, IOT) are now ubiquitous and their analysis has become a strategic field to secure our numerical lives. Unfortunately, most of communication networks come with textual data on the edges which are not incorporated in classical model. In this paper, we introduce the embedded topics for the stochastic block model (ETSBM) in order to simultaneously perform clustering on the nodes while modelling the topics used between the different clusters. ETSBM extends both the stochastic bloc model and the embedded topic model. The main motivation of this work is to allow a simultaneous analysis of texts and node interactions. A variational-Bayes expectation-maximisation algorithm (VBEM) combined with a stochastic gradient descent (SGD) is used to perform inference. A model selection criterion is also derived. The methodology is evaluated on synthetic data and on a real world dataset.

Keywords. embeddings, topic models, graph clustering, generative models, STBM, ETM, LDA, SBM.

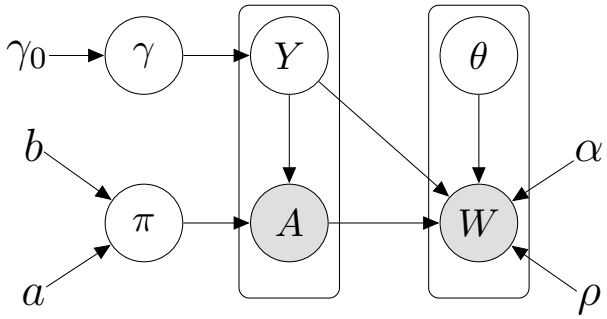


Figure 1: Graphical representation of the Embedded Topics for the Stochastic Block Model. The link between θ and δ is deterministic. In this figure, we give the dependency through θ.

1 Introduction and notations

The STBM model proposed in Bouveyron et al. (2016) aims at jointly analysing textual data and graph connections in order to perform simultaneously node clustering and topic modelling tasks for each detected cluster. The analysis of the textual data relies on word counts through LDA (Latent Dirichlet Allocation), see Blei et al. (2003) for more details. In LDA, the word "topic" refers to a distribution over the vocabulary. To overcome some of the limitations of LDA, e.g the bag-of-words representation of documents, ETM (Embedded Topic Model) has been introduced in Dieng et al. (2020). ETM makes use of a deep generative network that jointly learns word and topic representations in a vector space. It can also be used with pre-trained word embeddings, i.e vector representations of words, to incorporate semantic meaning of the words to improve the topic quality. We propose to use ETM instead of LDA within STBM to benefit from those advantages.

In the rest of this paper, we consider a graph with M nodes, for which $A \in \mathbb{R}^{M\times M}$ is the binary adjacency matrix of the graph such that there are no self loop, i.e $A_{ii} = 0$ for any i. Moreover, for any $i, j \in \{1, \dots, M\}$ with $i \neq j$, i is connected to j if and only if A_{ij} equals one. If so, they share a set of documents $W_{ij} := \{W_{ij}^1, \dots, W_{ij}^{D_{ij}}\}$ where D_{ij} denotes the number of documents shared by i and j such that for any $d = 1, \dots, D_{ij}$, W_{ij}^d is a collection of words with size N_{ij}^d, i.e $W_{ij}^d = \{w_{ij}^{d1}, \dots, w_{ij}^{dN_{ij}^d}\}$. V denotes the vocabulary size and the v-th word of the vocabulary will either be represented with a one hot encoded vector or directly with the vocabulary index of the word $w = v$. The embeddings are denoted $\rho \in \mathbb{R}^{d\times V}$ with d the dimension of the vector space and α_k refers to the vector representation of the topic k in the same space for all $k \in \{1, \dots, K\}$, with K the number of topics .

2 Generative model

This section presents our assumptions about the generation of the real data and are represented in Figure 1. The cluster memberships of the nodes are assumed to be independent and identically distributed following a multinomial distribution while the connections between the nodes are assumed to be independent and to only depend on the cluster of the nodes given each node's cluster:

$$p(Y \mid \gamma) = \prod_{i=1}^{M} \prod_{q=1}^{Q} \gamma_q^{Y_{iq}}, \qquad p(A \mid \pi, Y) = \prod_{i \neq j}^{M} \prod_{q,r}^{Q} \left(\pi_{qr}^{A_{ij}} (1-\pi_{qr})^{1-A_{ij}}\right)^{Y_{iq}Y_{jr}}. \tag{1}$$

We also assume that $\gamma \sim \mathcal{D}ir_Q(1, \ldots, 1)$ and $\pi_{qr} \sim \mathcal{B}eta(1,1)$ for any q, r.

For any pair of clusters (q, r), the prior on the topic proportions $\theta_{q,r}$ is a logistic normal distribution:

$$\delta_{qr} \sim \mathcal{N}(0_K, \mathrm{I}_K), \qquad \theta_{qr} = \mathrm{softmax}\left(\delta_{qr}\right).$$

with $\mathrm{softmax}(\delta) = \left(\frac{e^{\delta_1}}{\sum_k^K e^{\delta_k}}, \ldots, \frac{e^{\delta_K}}{\sum_k^K e^{\delta_k}}\right)$ for any $\delta \in \mathbb{R}^K$.

If two nodes i, j are connected and in the clusters q and r respectively, the n-th word of the d-th document is assumed to be distributed according to a mixture of topics:

$$p(w_{ij}^{dn} \mid \theta_{qr}, Y_{iq}Y_{jr}A_{ij} = 1) = \sum_{k=1}^{K} \theta_{qrk}\, \mathrm{softmax}(\rho^\top \alpha_k)_{|w_{ij}^{dn}}. \tag{2}$$

Finally, given each node's cluster membership Y_i, marginalizing the topic proportions gives the following:

$$p(W \mid Y, A, \alpha, \rho) = \prod_{i \neq j}^{M} \left\{ \prod_{q,r}^{Q} \left(\int p\left(\delta_{qr}\right) \prod_{d=1}^{D_{ij}} \prod_{n=1}^{N_{ij}^d} p\left(w_{ij}^{dn} \mid \delta_{qr}, Y_{iq}Y_{jr}A_{ij} = 1\right) d\delta_{qr} \right)^{Y_{iq}Y_{jr}} \right\}^{A_{ij}}. \tag{3}$$

3 Inference

We aim at maximising $\log p(A, W \mid \alpha, \rho)$ with respect to the parameters α, ρ. To evaluate it, it is necessary to marginalize over the set of parameters,

$$\log p(A, W \mid \alpha, \rho) = \log \left(\sum_Y \int_\delta \int_\gamma \int_\pi p(A, W, Y, \pi, \gamma, \delta \mid \alpha, \rho) d\pi d\delta d\gamma \right). \tag{4}$$

Unfortunately, the quantity in (4) is intractable since it requires computing the integral for all the Q^M configurations of Y. Moreover, the integral with respect to δ is not tractable because of the softmax function. To overcome this, we use a variational-bayes expectation-maximization (VBEM) approach to keep this method scalable to large datasets.

$$\log p(A, W \mid \alpha, \rho) = \mathscr{L}(R(\cdot)) + \mathrm{KL}(R(\cdot)||p(Y, \pi, \gamma, \delta \mid A, W)),$$

with

$$\mathscr{L}(R(\cdot)) = \sum_{Y} \int_{\pi,\gamma,\theta} R(Y, \pi, \gamma, \delta) \log \frac{p(A, W, Y, \pi, \gamma, \delta \mid \alpha, \rho)}{R(Y, \pi, \gamma, \delta)} d\pi d\delta d\gamma.$$

To make $\mathscr{L}(R(\cdot))$ tractable, we use the mean-field assumption,

$$R(Y, \pi, \gamma, \delta) = R(Y)R(\pi)R(\gamma)R(\delta).$$

We propose to use the same distributions as the one in the generative model. We use the amortized inference to avoid increasing the number of parameters linearly with the number of pairs of groups and to keep the training of the auto-encoder tractable, see Gershman and Goodman (2014) for more details.

$$R_\tau(Y) = \prod_{i=1}^{M} \mathcal{M}_Q(Y_i; \tau_i), \quad R_{\tilde{\pi}_{qr1}, \tilde{\pi}_{qr2}}(\pi) = \mathcal{B}eta(\pi_{qr}; \tilde{\pi}_{qr1}, \tilde{\pi}_{qr2}), \tag{5}$$

$$R_{\tilde{\gamma}}(\gamma) = \prod_{q=1}^{Q} \mathcal{D}ir(\gamma; \tilde{\gamma}), \quad R_{\mu,\sigma}(\delta) = \prod_{q,r}^{Q} \mathcal{N}\Big(\delta_{qr}; \mu^{\nu}_{qr}, \sigma^{\nu}_{qr}\Big), \tag{6}$$

with $\mu_{qr} = f(\tilde{w}_{qr}; \nu, \nu_m)$ and $\sigma_{qr} = f(\tilde{w}_{qr}; \nu, \nu_v)$, typically a neural network, ν denotes some parameters of the neural network shared by the function encoding μ and σ while ν_m and ν_v are parameters specific to each encoder. Finally, $\tilde{w}_{qr} = \sum_{i \neq j} \tau_{iq}\tau_{jr}A_{ij}W_{ij}$, are the expected documents under the approached posterior distribution, which simply is the variational distribution. Since $\tilde{w}_{qr}$ are constructed using $\tau = (\tau_i)_{1 \leq i \leq M}$, the parameters μ^{ν}_{qr} and σ^{ν}_{qr} also depend on τ. The Kullback-Leibler divergence between two Gaussian variables has a close form and is easy to compute. All the terms can be computed except for the expectation of $T_{ij}^{\theta_{qr}}$ that can be approximated using a Monte-Carlo estimator, by drawing S samples for each pair (q, r), such that:

$$\epsilon^s \sim \mathcal{N}(0, I_K), \quad \delta^s_{qr} = \mu^{\nu}_{qr} + \sigma^{\nu}_{qr} \odot \epsilon^s, \quad \theta^s_{qr} = \mathrm{softmax}(\delta^s_{qr}).$$

3.1 Optimisation & Algorithm

We aim at maximising the ELBO with respect to the variational parameters $\tilde{\pi}, \tilde{\gamma}, \tau$ and η and to the parameters ρ and α. Since τ lies on the simplex, the Lagrangian $\mathcal{L}$ of the

ELBO is given by:

$$\mathcal{L}(\gamma, \kappa, \tau, \lambda) = \mathscr{L}(\gamma, \kappa, \tau) + \sum_{i=1}^{M} \lambda_i \left(\sum_{q=1}^{Q} \tau_{iq} - 1 \right)$$

On the one hand, setting the derivatives of the Lagrangian with respect to $\tilde{\pi}, \tilde{\gamma}$ to zero gives the classical updates of the variational Bayes SBM, depending only on τ, see Latouche et al. (2012) for more details.

On the other hand, the optimisation of the ETM parameters is done by stochastic gradient descent based on Dieng et al. (2020). One should be careful about the fact that the meta-documents given to the encoder are based on τ and impact the ELBO also through the KL term which is very specific to our approach and allows regularization in the clustering task. Thus, contrary to existing variational auto-encoding strategy, the inputs of the encoder can here be optimized through τ and the allocations of nodes to clusters. In practice, we rely on the automatic differentiation technique to perform the optimization of the NLP term of the ELBO.

4 Conclusion and discussion

The embedded topics for the stochastic block model (ETSBM) introduced in this paper is well suited to perform clustering simultaneously on the nodes and edges and to sum up the textual edges content and the connections pattern. It can be used both on directed and undirected graphs and is suited for big datasets since it relies on a variational inference. The numerical experiments suggest the ELBO can be used as a criteria for selecting the number of cluster Q thanks to the Bayesian framework used. The model has been tested on a real dataset which showed both usefulness and its capacity in capturing information and performing meaningful clustering. Further work may be directed in the study of theoretical foundation of the selection model criteria, in incorporating dynamic networks and finally in using out of vocabulary words.

On going work This section is under current investigation. We will conduct a set of simulations based on the ones in Bouveyron et al. (2016). An application will be carried out on a dataset of academic papers and a co-authors network to detect both groups of authors and underlying topics. We will also conduct simulation to validate the use of Bayesian priors as a regularization within the model to penalize the ELBO and to make it a good model selection criterion for the number of clusters K in the network. The result will be presented at the conference if this paper is accepted.

References

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Bouveyron, C., Latouche, P., and Zreik, R. (2016). The stochastic topic block model for the clustering of vertices in networks with textual edges. *Statistics and Computing*, 28(1):11–31.

Dieng, A. B., Ruiz, F. J., and Blei, D. M. (2020). Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453.

Gershman, S. and Goodman, N. (2014). Amortized inference in probabilistic reasoning. In *Proceedings of the annual meeting of the cognitive science society*, volume 36.

Latouche, P., Birmele, E., and Ambroise, C. (2012). Variational bayesian inference and complexity control for stochastic block models. *Statistical Modelling*, 12(1):93–115.

Modèles génératifs de graphes
Application à la connectivité cérébrale

Clément Mantoux [1,2,3], Stanley Durrleman [1,2] & Stéphanie Allassonnière [4,5]

[1] *ARAMIS Project Team, Inria, Paris 75013, France*
[2] *ARAMIS Lab, Brain and Spine Institute, ICM, INSERM UMR 1127, CNRS UMR 7225, Sorbonne Université, Hôpital de la Pitié-Salpêtrière, Paris 75013, France ; clement.mantoux@inria.fr, stanley.durrleman@inria.fr*
[3] *CMAP, École polytechnique, Palaiseau 91120, France*
[4] *Centre de Recherche des Cordeliers, Université de Paris, INSERM UMR 1138, Sorbonne Université, Paris 75006, France ; stephanie.allassonniere@parisdescartes.fr*
[5] *HeKA Project Team, Inria, Paris 75006, France*

Résumé La connectivité fonctionnelle est l'étude de réseaux d'interactions entre les régions du cerveau humain. Pour chaque individu, on peut ainsi mesurer une matrice de connectivité, qui varie beaucoup d'un sujet à l'autre. En raison du faible nombre d'observations et de leur grande dimension, de nombreux modèles statistiques classiques ne peuvent être employés en pratique pour décrire cette variabilité. Nous présentons un modèle hiérarchique simple pour contourner ces difficultés en s'appuyant sur la structure de rang faible des matrices de connectivité. Nous montrons que notre modèle est identifiable, et que son estimateur est consistant et asymptotiquement normal. Enfin, nous appliquons cet algorithme à une base de données de matrices de connectivité cérébrale. Nous montrons que notre modèle donne une description précise, compacte et interprétable de la distribution des données.

Mots-clés. Modèle hiérarchique, Ensembles de graphes, Connectivité cérébrale, Algorithme EM, Variétés de Stiefel, Distributions de von Mises-Fisher.

Abstract. Functional connectivity is the study of interactions networks between regions of the human brain. For each individual, a connectivity matrix can be measured, which varies greatly from subject to subject. Due to the small number of observations and their high dimension, many classical statistical models cannot be used in practice to describe this variability. We present a simple hierarchical model to overcome these difficulties by exploiting the low-rank structure of connectivity matrices. We show that our model is identifiable, and that its estimator is consistent and asymptotically normal. Finally, we apply this algorithm to a database of brain connectivity matrices. We show that our model gives an accurate, compact and interpretable description of the data distribution.

Keywords. Hierarchical model, Populations of networks, Brain connectivity, EM algorithm, Stiefel manifolds, von Mises-Fisher distributions.

1 Un modèle pour les ensembles de graphes

1.1 Motivation

Lorsqu'on étudie la connectivité fonctionnelle du cerveau d'un individu, on mesure via une Imagerie par Résonance Magnétique (IRM), pour chaque paire de régions du cerveau (i, j), un coefficient d'interaction $A_{ij} = A_{ji}$. La matrice symétrique $A \in \mathbb{R}^{n\times n}$ ainsi obtenue résume donc comment ces régions "fonctionnent" ensemble : si $A_{ij} > 0$, les régions sont souvent activées ensemble, et si $A_{ij} < 0$ leurs activations sont en moyenne anti-corrélées.

En neurosciences, on cherche à expliquer la variabilité de la connectivité d'un individu à l'autre. On dispose typiquement de N matrices symétriques $A^{(1)}, \dots, A^{(N)} \in \mathbb{R}^{n\times n}$ mesurées sur autant de sujets. Les IRM fonctionnelles nécessitant un matériel important, la plupart des études disposent de quelques centaines de sujets. Les observations obtenues sont de dimension comparable au nombre d'individus, par exemple $n(n+1)/2 = 231$ si le cerveau est divisé en $n = 21$ régions. Dans ce contexte, les modèles les plus simples sont inefficaces. Par exemple, une distribution Gaussienne requiert $O(n^4)$ paramètres, trop pour être estimés en pratique ; ces paramètres sont en outre difficiles à interpréter par un neurologue.

Il est donc nécessaire d'exploiter la structure particulière des matrices de connectivité $A^{(i)}$. Ces matrices ont en pratique un rang faible devant n, ce qui permet une description compacte de leurs propriétés. Cette propriété est notamment exploitée dans des auto-encodeurs de graphes ou des modèles de dictionnaire. Le modèle hiérarchique que nous présentons dans cet exposé, ainsi que son application aux données cérébrales, ont récemment fait l'objet d'une publication (Mantoux *et al.*, 2021). Ce modèle rassemble des propriétés des auto-encodeurs de graphes et des modèles de dictionnaires. Il propose un cadre génératif plus souple, adapté pour modéliser les ensembles de matrices de connectivité et plus généralement les ensembles de matrices d'adjacence.

1.2 Description du modèle

Le modèle que nous proposons s'appuie sur la décomposition spectrale des matrices de connectivité. Chaque matrice $A^{(i)} \in \mathbb{R}^{n\times n}$ peut ainsi s'écrire

$$A^{(i)} = X\mathrm{Diag}(\lambda)X^\top = \lambda_1 x_1 x_1^\top + \cdots + \lambda_n x_n x_n^\top ,$$

avec n le nombre de régions du cerveau, $\lambda_1 \geq \cdots \geq \lambda_n$ les valeurs propres de $A^{(i)}$ et $X = (x_1, \dots x_n) \in \mathbb{R}^{n\times n}$ ses vecteurs propres. Dire que la matrice $A^{(i)}$ est de rang faible revient alors à faire l'approximation $A^{(i)} = \sum_{k=1}^{p} \lambda_k x_k x_k^\top + \varepsilon$: on ne modélise que les p ($\ll n$) premiers termes de cette décomposition, et on traite le résidu ε comme un terme de bruit.

Nous proposons de décrire la distribution des matrices $A^{(i)}$ en spécifiant la distribution des p premiers vecteurs propres $X^{(i)} \in \mathbb{R}^{n\times p}$ et valeurs propres $\lambda^{(i)} \in \mathbb{R}^p$, ainsi que celle

du bruit. On modélise ainsi les valeurs propres par une distribution Gaussienne $\lambda^{(i)} \sim \mathcal{N}(\mu, \sigma_\lambda^2 I_p)$, et les vecteurs propres par une distribution spécifique que nous détaillons ci-dessous. Le bruit $\varepsilon^{(i)}$ est modélisé comme une matrice symétrique de coefficients Gaussiens indépendants, de variance σ_ε^2 : $\varepsilon^{(i)} \sim \mathcal{N}_{\text{sym}}(0, \sigma_\varepsilon^2 I_{n(n+1)/2})$. Enfin, on fait l'hypothèse que $X^{(i)}$, $\lambda^{(i)}$ et $\varepsilon^{(i)}$ sont indépendants. Cette hypothèse forte permet de conserver un nombre de paramètres restreint, et produit néanmoins des résultats cohérents en pratique.

La distribution des $X^{(i)}$ est contrainte par la propriété d'orthogonalité des vecteurs propres : ceux-ci satisfont en effet d'après le théorème spectral, pour tout i, $(X^{(i)})^\top X^{(i)} = I_p$, et sont donc des éléments de la variété de Stiefel $\mathcal{V}_{np}$. On modélise donc les variables $X^{(i)}$ par une distribution de von-Mises Fisher (vMF) sur la variété de Stiefel (Khatri et Mardia, 1977). Cette distribution peut être considérée comme l'analogue sur $\mathcal{V}_{np}$ d'une Gaussienne à composantes indépendantes. Sa densité de probabilité s'exprime simplement :

$$p_{\text{vMF}}(X) = \frac{1}{\mathcal{C}_{np}(f_1, \dots, f_p)} \exp(\langle f_1, x_1 \rangle + \dots \langle f_p, x_p \rangle) \,.$$

Pour chaque vecteur x_k, la direction $f_k/|f_k|$ donne le point de plus forte probabilité, et la norme $|f_k|$ contrôle la concentration autour de ce point. Dans la limite où $F = (f_1, \dots f_p)$ devient grand, la distribution devient concentrée en un point, et quand F devient petit la distribution devient uniforme sur $\mathcal{V}_{np}$. La constante de normalisation $\mathcal{C}_{np}(f_1, \dots, f_p)$ s'exprime par des fonctions hypergéométriques matricielles. Elle peut être approximée en pratique par une méthode numérique proposée par Kume *et al.* (2013).

Le modèle peut enfin être résumé par les équations suivantes :

$$\begin{cases} X^{(i)} \sim \text{vMF}(F) \\ \lambda^{(i)} \sim \mathcal{N}(\mu, \sigma_\lambda^2 I_p) \\ \varepsilon^{(i)} \sim \mathcal{N}_{\text{sym}}(0, \sigma_\varepsilon^2 I_{n(n+1)/2}) \\ A^{(i)} = (X^{(i)})^\top \text{Diag}(\lambda^{(i)}) X^{(i)} + \varepsilon^{(i)} \,. \end{cases}$$

Les paramètres du modèle sont $\theta = (F, \mu, \sigma_\lambda, \sigma_\varepsilon)$. Nous reviendrons plus en détail sur leur interprétation dans le paragraphe d'application à la connectivité cérébrale.

1.3 Estimation par Maximum A Posteriori

Dans un article en cours de revue, nous examinons les propriétés du modèle ainsi que son estimation. Nous montrons que notre modèle est identifiable dès lors que les colonnes de F sont non nulles et triées selon un ordre quelconque. Nous montrons que l'estimateur du maximum de vraisemblance peut théoriquement diverger vers des variances infinies pour les variables latentes. Afin de le régulariser, nous considérons donc le Maximum A Posteriori (MAP), qui ajoute une probabilité *a priori* sur θ et garantit que l'estimateur est bien défini. Nous montrons que cet estimateur est fortement consistant, et nous retrouvons le classique résultat de normalité asymptotique pour une version restreinte du modèle.

En pratique, les paramètres du modèle sont estimés numériquement en utilisant l'algorithme MCMC-SAEM (Kuhn et Lavielle, 2004; Allassonnière *et al.*, 2010), qui adapte l'algorithme EM aux modèles exponentiels complexes, où l'étape E ne peut être calculée.

2 Application à la connectivité fonctionnelle

Nous appliquons enfin notre modèle aux données de l'UK Biobank (Sudlow *et al.*, 2015). Nous disposons de $N = 1000$ matrices de connectivité fonctionnelle, donnant les corrélations entre $n = 21$ régions du cerveau. La figure 1 montre cinq matrices du jeu de données, où l'on peut voir la diversité des interactions d'un individu à l'autre. Nous estimons les paramètres du modèle pour $p = 5$ composantes de la décomposition spectrale.

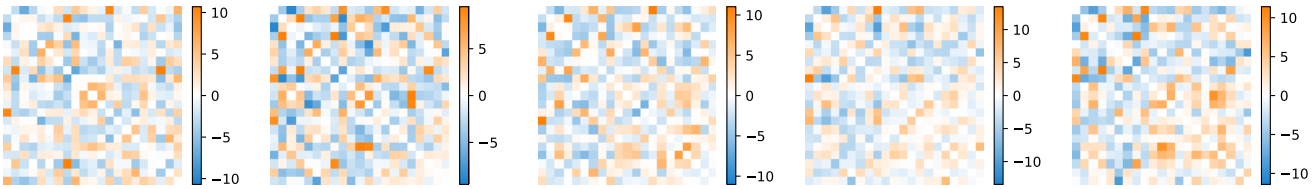


FIGURE 1 – Cinq matrices de connectivité de l'UK Biobank.

Interprétation Sur la figure 2, nous visualisons les termes $f_k f_k^\top / |f_k|^2$, qui correspondent à la décomposition spectrale de plus forte probabilité. La décomposition spectrale de chaque individu est ainsi une variation de cette décomposition centrale. L'interprétation des résultats est simple : chaque terme $f_k f_k^\top$ est une matrice de rang 1, où la participation de chaque région est résumée par un coefficient. La concentration $|f_k|$ donne la variabilité de chaque terme de la décomposition au sein de la population. Le terme (5) de la décomposition implique un nombre très restreint de régions du cerveau, toutes participant aux fonctions de la vision. Le terme (3) fait principalement figurer l'anti-corrélation entre les régions 1 et 3. La région 1 contient plusieurs parties du cerveau impliquées dans l'activité au repos, tandis que la région 3 comporte certaines parties liées à la perception active de sensations. Le terme (4) comporte principalement une corrélation des zones de la vision et une corrélation des zones liées au contrôle moteur, ainsi qu'une anti-corrélation entre ces deux groupes de régions.

Prédiction *a posteriori* Enfin, nous évaluons l'adéquation de notre modèle aux données en effectuant un test de prédiction. Pour chaque matrice, on masque un ensemble de coefficients, et on estime la valeur la plus probable pour les coefficients cachés sachant les coefficients visibles et les paramètres du modèle. On compare cette prédiction à celle d'un modèle linéaire entraîné pour prédire les valeurs cachées à partir des valeurs visibles.

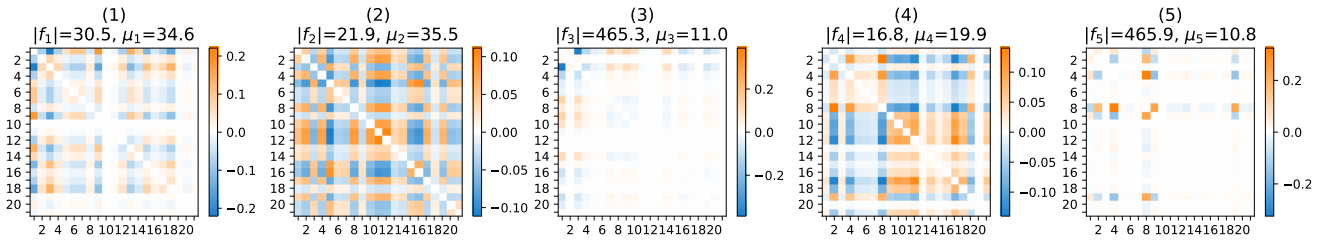


FIGURE 2 – Termes de la décomposition centrale. On visualise $f_k f_k^\top / |f_k|^2$, multiplié par signe(μ_k) pour avoir le bon signe des coefficients.

Notre modèle réalise une erreur relative moyenne de 65% (±15%), contre 58% (±14%) pour le modèle linéaire. À titre de comparaison, notre modèle utilise 112 paramètres, contre 26.640 pour le modèle linéaire, qui a spécifiquement été entraîné sur cette tâche. Le modèle proposé offre ainsi une représentation compacte et informative du jeu de données. Un exemple de prédiction est montré dans la figure 3.

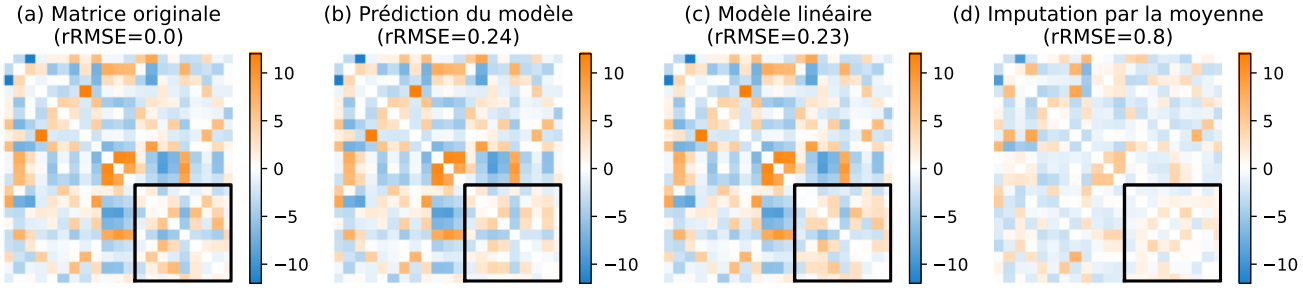


FIGURE 3 – Exemple de prédiction de lien. On compare les erreurs relatives (rRMSE) de prédictions de notre modèle et du modèle linéaire, par rapport à la matrice originale. On affiche également la moyenne empirique des données pour comparaison.

3 Conclusion

Le modèle que nous avons introduit produit une description fidèle et interprétable de la distribution des matrices de connectivités cérébrales. Prendre en compte la structure de rang faible de ces matrices permet d'utiliser un nombre restreint de paramètres, et garantit ainsi une estimation efficace. Ce travail ouvre par ailleurs plusieurs perspectives. Notre modèle pourrait être appliqué à d'autres ensembles de graphes, comme les réseaux de transport. Le modèle hiérarchique introduit peut sans peine être adapté pour prendre en compte des covariables comme l'âge ou le sexe des individus ; il pourrait par exemple s'intégrer comme brique élémentaire d'un modèle longitudinal de suivi de population.

Des extensions de notre modèle pourraient être considérées pour prendre en compte la dépendance entre valeurs propres et vecteurs propres.

Remerciements Les auteurs remercient Baptiste Couvy-Duchesne, Federica Cacciamani et Stéphane Epelbaum pour leur contribution aux travaux présentés.

Références

Stéphanie Allassonnière, Estelle Kuhn et Alain Trouvé : Construction of Bayesian deformable models via a stochastic approximation algorithm : A convergence study. *Bernoulli*, 16(3):641–678, août 2010. ISSN 1350-7265.

Chinubhai. G. Khatri et Kantilal V. Mardia : The von Mises–Fisher Matrix Distribution in Orientation Statistics. *Journal of the Royal Statistical Society : Series B (Methodological)*, 39(1):95–106, 1977. ISSN 2517-6161.

Estelle Kuhn et Marc Lavielle : Coupling a stochastic approximation version of EM with an MCMC procedure. *ESAIM : Probability and Statistics*, 8:115–131, août 2004. ISSN 1292-8100, 1262-3318.

Alfred Kume, Simon P. Preston et Andrew T. A. Wood : Saddlepoint approximations for the normalizing constant of Fisher–Bingham distributions on products of spheres and Stiefel manifolds. *Biometrika*, 100(4):971–984, décembre 2013. ISSN 0006-3444.

Clément Mantoux, Baptiste Couvy-Duchesne, Federica Cacciamani, Stéphane Epelbaum, Stanley Durrleman et Stéphanie Allassonnière : Understanding the Variability in Graph Data Sets through Statistical Modeling on the Stiefel Manifold. *Entropy*, 23(4):490, avril 2021.

Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, Bette Liu, Paul Matthews, Giok Ong, Jill Pell, Alan Silman, Alan Young, Tim Sprosen, Tim Peakman et Rory Collins : UK Biobank : An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Medicine*, 12(3):e1001779, mars 2015. ISSN 1549-1676.

Analyse statistique de graphes, via des processus de diffusion de la chaleur.

Etienne Lasalle [1,2]

[1] *Université Paris-Saclay, CNRS, Laboratoire de Mathématiques d'Orsay, Rue Michel Magat Bâtiment 307, 91405 Orsay, France.*
[2] *DataShape, Inria Saclay.*
etienne.lasalle@universite-paris-saclay.fr

Résumé. Ce travail porte sur la comparaison de données de graphes, potentiellement pondérés et de tailles différentes. Lorsqu'on travaille avec des graphes pondérés, on peut interpréter les poids des arêtes comme des conductivités thermiques. Dès lors, on peut comparer les graphes en comparant leur répartition de chaleur après un temps de diffusion t. Ce paramètre d'échelle t doit être minutieusement choisi pour s'assurer des comparaisons pertinentes. A l'opposé de précédents travaux considérant un temps de diffusion fixé arbitrairement ou choisi à partir des données, on propose de prendre en compte tout le processus de diffusion. Pour cela, on définit des processus à valeurs réelles indexés par tous les temps de diffusion dans $[0, T]$, en concaténant les comparaisons faites aux différentes échelles. Dans cet exposé, nous commencerons par présenter ces processus de comparaison de graphes et leurs propriétés statistiques. Puis, nous montrerons comment en dériver des tests à deux échantillons consistants. Nous présenterons quelques applications sur des jeux de données synthétiques et réels. On s'intéressera notamment aux données provenant de graphes d'activations de réseaux de neurones.

Mots-clés. Graphes, processus empiriques, diffusion de la chaleur, test à deux échantillons.

Abstract. This work deals with the comparison of graph data, potentially with weighted edges and of different sizes. When working with weighted graphs, one can interpret weights as the thermal conductivity of edges. Therefore, graphs can be characterized and compared using their heat distribution after a diffusion time t. This scale parameter t needs to be carefully chosen to ensure relevant comparisons. While previous works considered either arbitrarily fixed or data dependent diffusion times, our approach takes into account the whole diffusion process. For that, we define real-valued processes indexed by all the diffusion times in $[0, T]$, corresponding to the comparisons at all the different scales. In this talk, we will start by introducing these comparison processes and their statistical properties. Then, we will show how these results can be applied to derive consistent two-sample tests. Our work will be illustrated on synthetic and real data sets. A particular attention will be devoted to activation graphs from neural networks.

Keywords. Graphs, empirical processes, heat diffusion, two-sample tests.

1 Introduction

De part leurs capacités de modélisation et grâce aux moyens d'acquisition et de stockage de plus en plus importants, les données de graphes reçoivent une attention croissante. Une effort particulier a été fourni pour développer des notions de distance entre graphes. Cependant, ces notions de distances sont extrêmement dépendantes du contexte et des caractéristiques des graphes considérés. Plus particulièrement, différentes informations peuvent être disponibles : les graphes peuvent être orientés, pondérés, posséder le même nombre de sommets, ou non. Une autre information clé peut être la connaissance d'une correspondance entre les sommets des graphes.

Dans ce travail [Lasalle, 2021], nous considérons ces différentes variations pour des graphes non orientés. Pour cela, nous définissons deux processus de comparaisons basés sur la diffusion de la chaleur. Le premier suppose la correspondance entre les sommets connue, le second s'affranchit de ces hypothèses. Leur analyse nous permet de développer de méthodes de bootstrap pour déterminer des bandes de confiances ainsi que des tests à deux échantillons.

2 Notions de distance, via la diffusion de la chaleur

Considérons $\mathcal{G}_n$ l'ensemble des graphes pondérés de taille n partageant les mêmes sommets $\{1, \dots, n\}$ et tels que les poids des arêtes sont positifs. La diffusion de la chaleur sur un graph $G \in \mathcal{G}_n$ peut se caractériser par le *noyau de la chaleur* au temps $t \geq 0$, défini par une exponentielle de matrice :

$$K_t(G) := e^{-tL}, \tag{1}$$

où L est le laplacien de G. $K_t(G)$ est une matrice symétrique, positive, de taille $n \times n$. Sa j-ème colonne représente la distribution de chaleur des différents nœuds au temps t, lorsque seule une unité de chaleur se trouvait au nœud j au temps $t = 0$ (voir Figure 1). Pour deux graphes G, G' dans $\mathcal{G}_n$, on considère leur *Heat Kernel Distance* (HKD) [Hammond et al., 2013] au temps t, définie par

$$D_t((G, G')) = \|K_t(G) - K_t(G')\|_F, \tag{2}$$

où $\|\cdot\|_F$ désigne la norme de Frobenius.

Pour comparer des graphes sans correspondance entre les sommets ou de tailles différentes, on travaille dans $\mathcal{G}^n$, l'ensemble des graphes de taille *au plus* n. On définit alors une nouvelle notion de distance, en utilisant la théorie de la persistance étendue [Cohen-Steiner et al., 2009], un outil provenant de l'analyse topologique des données. Pour un graphe G ayant pour sommets $V = \{1, \dots, n'\}$, la Heat Kernel Signature (HKS) [Hu et al., 2014] au temps t se définit comme la fonction $h_t(G) : V \to \mathbb{R}$, vérifiant $h_t(G)(i) = K_t(G)(i, i)$. Une fois les graphes de $\mathcal{G}^n$ munis de leur HKS, on peut définir des descripteurs topologiques appelés diagrammes de persistance $Dg : G \to Dg(G, h_t(G))$

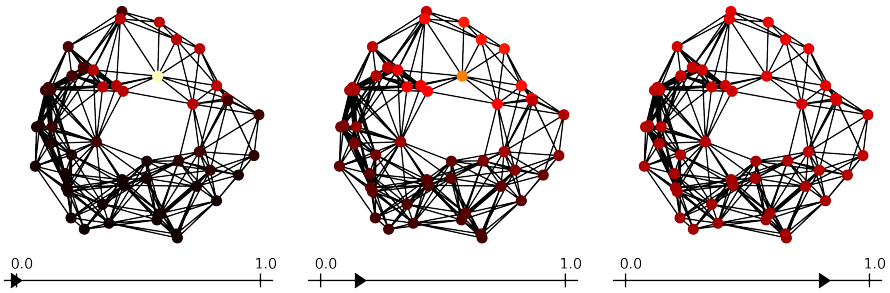


Figure 1: Représentations de répartitions de chaleur pour différents temps de diffusion (indiqués sur les axes horizontaux), pour une condition initiale élémentaire, *i.e.* où une seule unité de chaleur est placée en un sommet du graphe.

[Carrière et al., 2020]. Les diagrammes de persistance sont des multi-ensembles de points de $\mathbb{R}^2$. Ils encodent la manière dont les composantes topologiques, comme les composantes connexes ou les trous, évoluent au sein des familles de sous-graphes associés aux sous-niveaux et sur-niveaux des HKS. Les diagrammes de persistances peuvent ensuite être comparés via la distance Bottleneck d_B [Edelsbrunner and Harer, 2010, Section VIII.2]. Il s'agit d'une distance de transport entre les deux diagrammes utilisant la norme infini sur $\mathbb{R}^2$. Ces comparaisons de diagrammes se font sans aucune hypothèse sur la taille des graphes ou l'identification de sommets. Ainsi, on peut définir la Heat Persistence Distance (HPD) au temps t entre deux graphes G, G' dans $\mathcal{G}^n$ par

$$H_t((G, G')) = d_B(Dg(G, h_t(G)), Dg(G', h_t(G'))). \tag{3}$$

A la fois pour la HKD et la HPD, le choix d'un paramètre t pertinent et informatif est crucial. Nous proposons d'adopter un point de vue processus, et de considérer l'ensemble des distances pour des t dans $[0, T]$.

3 Processus empiriques

Soit P une mesure de probabilité sur $\mathcal{G}_n \times \mathcal{G}_n$ et considérons un N-échantillon $((G_1, G'_1), \ldots, (G_N, G'_N))$ tiré sous P. On s'intéresse alors aux propriétés statistiques du processus empirique associé à la HKD :

$$\left\{ \sqrt{N} \left(\frac{1}{N} \sum_{i=1}^{N} D_t((G_i, G'_i)) - \mathbb{E}_P \left[D_t((G, G')) \right] \right), \ t \in [0, T] \right\} \tag{4}$$

De même,pour un N-échantillon $((G_1, G'_1), \ldots, (G_N, G'_N))$ tiré sous une mesure de probabilité P sur $\mathcal{G}^n \times \mathcal{G}^n$, on étudie le processus empirique associé à la HPD :

$$\left\{ \sqrt{N} \left(\frac{1}{N} \sum_{i=1}^{N} H_t((G_i, G'_i)) - \mathbb{E}_P \left[H_t((G, G')) \right] \right), \ t \in [0, T] \right\} \tag{5}$$

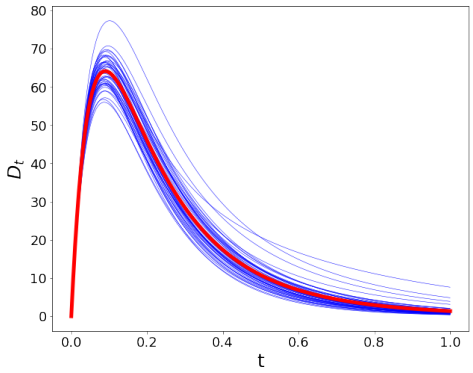


Figure 2: En bleu, les processus HKD associés à des paires de graphes de Erdös-Rényi indépendants. En rouge, la moyenne empirique de ces processus.

Grâce au caractère lipschitzien des fonctions $t \to D_t((G, G'))$ et $t \to H_t((G, G'))$, nous sommes en mesure de prouver la convergence faible de ces processus vers des processus gaussiens indexés par $[0, T]$. Ce résultat supporte théoriquement l'utilisation de la moyenne empirique des processus (voir Figure 2) dans des procédures statistiques. Notamment, il assure la validité asymptotique des méthodes bootstrap construisant des bandes de confiance contenant l'espérance des processus, ou construisant des tests à deux échantillons.

4 Applications

Pour évaluer les performances de ces tests à deux échantillons, on les applique à des données simulées en utilisant des modèles génératifs de graphes aléatoires tels que le modèle d'Erdös-Rényi, le Stochastic Block Model ou encore des modèles de graphes géométriques. On évalue la puissance et l'erreur de première espèce en fonction de la taille des échantillons.

On propose également d'appliquer ces tests à deux échantillons sur des graphes d'activations de réseaux de neurones. Un graphe d'activation encode la manière dont une donnée traverse un réseau de neurones entrainé (voir Figure 3). Ainsi, afin de s'assurer que l'on demande au réseau de faire une prédiction sur des données cohérentes avec les données d'entrainement, on peut effectuer un test à deux échantillons entre les graphes d'activations associés au données d'entrainement, et ceux associés aux données à tester.

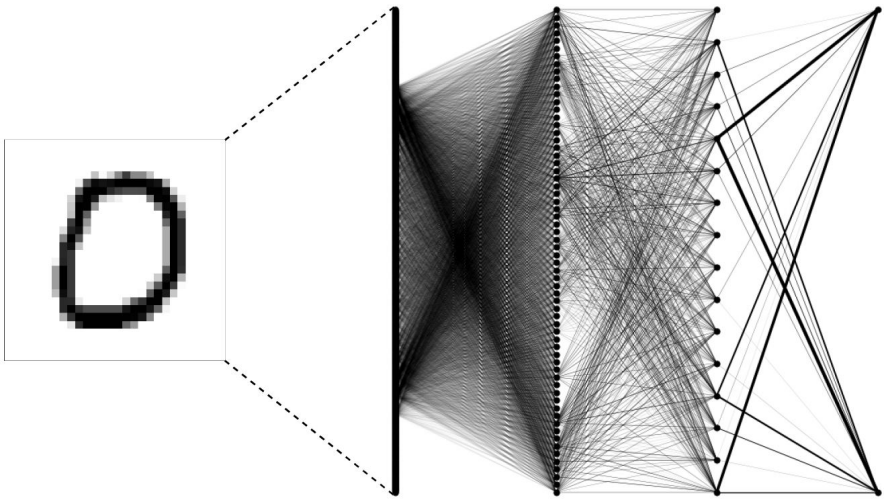


Figure 3: Exemple d'une donnée issue de MNIST [LeCun et al., 2010] et de son graphe d'activation, pour un réseaux dense avec deux couches cachées entrainé sur deux classes (chiffre 0 ou 1).

Bibliographie

[Carrière et al., 2020] Carrière, M., Chazal, F., Ike, Y., Lacombe, T., Royer, M., and Umeda, Y. (2020). Perslay: a neural network layer for persistence diagrams and new graph topological signatures. In *International Conference on Artificial Intelligence and Statistics*, pages 2786–2796. PMLR.

[Cohen-Steiner et al., 2009] Cohen-Steiner, D., Edelsbrunner, H., and Harer, J. (2009). Extending persistence using poincaré and lefschetz duality. *Foundations of Computational Mathematics*, 9(1):79–103.

[Edelsbrunner and Harer, 2010] Edelsbrunner, H. and Harer, J. (2010). *Computational topology: an introduction*. American Mathematical Soc.

[Hammond et al., 2013] Hammond, D. K., Gur, Y., and Johnson, C. R. (2013). Graph diffusion distance: A difference measure for weighted graphs based on the graph laplacian exponential kernel. In *2013 IEEE Global Conference on Signal and Information Processing*, pages 419–422. IEEE.

[Hu et al., 2014] Hu, N., Rustamov, R. M., and Guibas, L. (2014). Stable and informative spectral signatures for graph matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2305–2312.

[Lasalle, 2021] Lasalle, E. (2021). Heat diffusion distance processes: a statistically founded method to analyze graph data sets. *arXiv preprint arXiv:2109.13213*.

[LeCun et al., 2010] LeCun, Y., Cortes, C., and Burges, C. (2010). Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2.

Convergence en loi des U-statistiques sur une matrice exchangeable ligne-colonne

Tâm Le Minh [1]

[1] *Université Paris-Saclay, AgroParisTech, INRAE, UMR MIA-Paris, 75005, Paris, France, tam.le-minh@inrae.fr*

Résumé. Les U-statistiques sont utilisées pour estimer les paramètres d'une population en moyennant une fonction d'un sous-ensemble sur tous les sous-ensembles de cette population. Notre travail porte sur une population formée par les valeurs d'une matrice échangeable ligne-colonne. On considère les U-statistiques issues de fonctions sur des quadruplets, c'est-à-dire des sous-matrices de taille 2×2. On démontre un résultat de convergence faible pour ces U-statistiques et on établit un Théorème Central Limite dans le cas où la matrice est dissociée. Les matrices échangeables ligne-colonne sont une représentation naturelle des réseaux bipartites échangeables, nous appliquons donc les résultats à l'inférence statistique pour les réseaux.

Mots-clés. U-statistiques, échangeabilité ligne-colonne, Théorème Central Limite, réseaux bipartites

Abstract. U-statistics are used to estimate a population parameter by averaging a function of a subsample over all the subsamples of the population. In our work, the population we are interested in is formed by the entries of a row-column exchangeable matrix. We consider U-statistics derived from functions of quadruplets, i.e. submatrices of size 2×2. We prove a weak convergence result for these U-statistics in the general case and we establish a Central Limit Theorem when the matrix is also dissociated. Since row-column exchangeable matrices are an actual representation for exchangeable bipartite networks, we apply these results to statistical inference in network analysis.

Keywords. U-statistics, row-column exchangeability, Central Limit Theorem, bipartite network

1 Introduction

1.1 U-statistiques

Les U-statistiques sont une classe de statistiques qui généralisent le concept de moyenne empirique pour des fonctions à plusieurs variables. Par exemple, soit $X_1, X_2, ...$ une série

de variables aléatoires indépendantes et identiquement distribuées. Soit h une fonction de r variables. Alors sur un échantillon de taille n, la U-statistique associée à h est

$$U_n = \left[r! \binom{n}{r} \right]^{-1} \sum_{1 \leq i_1 \neq ... \neq i_r \leq n} h(X_{i_1}, ..., X_{i_r}).$$

Si $\mathbb{E}[h(X_1, ..., X_r)] = \theta$, alors U_n est un estimateur sans-biais et de variance réduite de θ sur cet échantillon.: $\mathbb{E}[U_n] = \theta$ et $\mathbb{V}[U_n] \leq \mathbb{V}[h(X_1, ..., X_r)]$. Hoeffding (1948) a montré que la loi de $\sqrt{n}(U_n - \theta)$ convergeait vers une loi normale centrée de variance connue, ce qui en fait un estimateur asymptotiquement normal. Les U-statistiques peuvent donc être utilisées pour l'estimation et la construction d'intervalles de confiance de paramètres.

Nandi et Sen (1963), puis Eagleson et Weber (1978) étendent ce résultat aux cas où les variables $X_1, X_2, ...$ sont échangeables, c'est-à-dire que pour toute permutation finie σ,

$$(X_1, X_2, ...) \overset{\mathcal{L}}{=} (X_{\sigma(1)}, X_{\sigma(2)}, ...).$$

Ils montrent que les U-statistiques de variables échangeables convergent en loi vers un mélange infini de gaussiennes, puis identifient les cas où ce mélange se réduit à une gaussienne simple, ce qui établit un Théorème Central Limite (TCL).

1.2 Matrices RCE

Certains types de données, tels que les réseaux d'interaction bipartites et les tables de présence-absence s'inscrivent souvent dans un cadre d'échangeabilité plus complexe. En effet, ces données peuvent se présenter sous la forme d'une matrice rectangulaire Y de taille $m \times n$, où les lignes et les colonnes représentent deux types d'entités différentes en interaction, comme m insectes et n plantes dans les réseaux plantes-pollinisateurs, ou m espèces et n sites dans les tableaux de présence-absence. Y_{ij} désigne donc l'interaction entre l'entité i de type 1 et l'entité j de type 2. Si les entités sont échangeables au sein de chaque type, alors la matrice est dite échangeable ligne-colonne (row-column exchangeable, RCE). Elles vérifient la propriété suivante, pour tout couple de permutations finies $\Phi = (\sigma_1, \sigma_2)$,

$$\Phi Y \overset{\mathcal{L}}{=} Y,$$

avec $\Phi Y := (Y_{\sigma_1(i)\sigma_2(j)})_{1 \leq i \leq m, 1 \leq j \leq n}$.

1.3 Cadre d'étude

Nous considérons des fonctions de quadruplets, c'est-à-dire des matrices de taille 2×2 que nous notons

$$Y_{\{i_1,i_2;j_1,j_2\}} := \begin{pmatrix} Y_{i_1 j_1} & Y_{i_1 j_2} \\ Y_{i_2 j_1} & Y_{i_2 j_2} \end{pmatrix}.$$

Sans perte de généralité, nous pouvons considérer que ces fonctions h sont symétriques, on parle alors de noyaux, telles que $h(Y_{\{1,2;1,2\}}) = h(Y_{\{2,1;1,2\}}) = h(Y_{\{1,2;2,1\}}) = h(Y_{\{2,1;2,1\}})$. La U-statistique associée au noyau h sur la matrice Y s'écrit alors

$$U_{m,n}^h = \binom{m}{2}^{-1}\binom{n}{2}^{-1} \sum_{\substack{1\le i_1<i_2\le m\\ 1\le j_1<j_2\le n}} h(Y_{\{i_1,i_2;j_1,j_2\}}). \tag{1}$$

La structure de symétrie des quadruplets est plus complexe que celles des variables qui sont étudiées dans la littérature des U-statistiques. Nous démontrons donc que les U-statistiques de quadruplets convergent bien en loi vers un mélange de gaussiennes et nous identifions précisément les cas où ce résultat devient un TCL.

2 Résultats

À présent, nous considérons Y une matrice RCE infinie, i.e. pour tout (m, n), la sous-matrice formée par les m premières lignes et les n premières colonnes de Y est RCE suivant la définition du paragraphe 1.2. Afin de déduire un résultat asymptotique, nous construisons une suite de dimensions $m_N \times n_N$ définie par

$$\begin{cases} m_N = 2 + \lfloor c(N+1) \rfloor \\ n_N = 2 + \lfloor (1-c)(N+1) \rfloor \end{cases}$$

La suite de U-statistiques considérée est $U_N^h := U_{m_N,n_N}^h$ définie par la formule (1).

De cette manière, U_N^h est calculée sur un nombre m_N de lignes et un nombre n_N de colonnes croissant à la même vitesse $O(N)$, et $c \in]0, 1[$ est un facteur de forme qui rend la matrice rectangulaire.

2.1 Théorème général

Théorème 2.1. *Soient* $\mathcal{F}_N = \sigma\big((U_{kl}^h, k \ge m_N, l \ge n_N)\big)$ *et* $\mathcal{F}_\infty := \bigcap_{N=1}^\infty \mathcal{F}_N$. *On pose* $U_\infty^h := \mathbb{E}[h(Y_{\{1,2;1,2\}})|\mathcal{F}_\infty]$ *et*

$$V = \frac{4}{c}\mathrm{Cov}\big(h(Y_{\{1,2;1,2\}}), h(Y_{\{1,3;3,4\}})\big|\mathcal{F}_\infty\big) + \frac{4}{1-c}\mathrm{Cov}\big(h(Y_{\{1,2;1,2\}}), h(Y_{\{3,4;1,3\}})\big|\mathcal{F}_\infty\big).$$

Si $\mathbb{E}[h(Y_{\{1,2;1,2\}})^2] < \infty$ *et* $V \neq 0$, *alors*

$$\sqrt{N}(U_N^h - U_\infty^h) \xrightarrow[N\to\infty]{\mathcal{D}} W,$$

où W *est une variable aléatoire de fonction caractéristique* $\phi(t) = \mathbb{E}[\exp(-\frac{1}{2}t^2V)]$ *(mélange infini de gaussiennes).*

Ce théorème découle du résultat de convergence des sommes de différences de martingales inverses d'Eagleson et Weber (1978). Ici, les différences de martingales inverses sont les $Z_{NK} := \sqrt{N}(U_K^h - U_{K+1}^h)$ et leur somme donne bien $\sum_{K=N}^{\infty} Z_{NK} = \sqrt{N}(U_K^h - U_\infty^h)$. La démonstration comporte trois étapes :

1. d'abord, on montre que Z_{NK} est bien une différence de martingales inverses, i.e. U_N^h est une martingale inverse par rapport à $\mathcal{F}_N$, i.e. $\mathbb{E}[U_N^h|\mathcal{F}_{N+1}] = U_{N+1}^h$ pour tout N,

2. puis, on montre l'existence et l'expression d'une variable aléatoire V ("variance asymptotique") telle que
$\sum_{K=N}^{\infty} \mathbb{E}[Z_{NK}^2|\mathcal{F}_{K+1}] \xrightarrow[N\to\infty]{\mathbb{P}} V$,

3. enfin, on vérifie la condition de Lindeberg conditionnelle, i.e. pour tout $\epsilon > 0$,
$\sum_{K=N}^{\infty} \mathbb{E}[Z_{NK}^2 \mathbb{1}_{\{|Z_{NK}|>\epsilon\}}|\mathcal{F}_{K+1}] \xrightarrow[N\to\infty]{\mathbb{P}} 0$.

2.2 Un Théorème Central Limite

On remarque que si V est constant, alors on obtient un TCL. On a démontré que c'était le cas si la matrice infinie Y est dissociée, c'est-à-dire que pour tout (m, n), $(Y_{ij})_{1\le i\le m, 1\le j\le n}$ est indépendante de $(Y_{ij})_{m<i<\infty, n<j<\infty}$.

Théorème 2.2. *En plus des hypothèses du Théorème 2.1, si Y est dissociée, alors U_∞^h et V sont constants et*

$$\sqrt{N}(U_N^h - U_\infty^h) \xrightarrow[N\to\infty]{\mathcal{L}} \mathcal{N}(0, V),$$

Plus précisément,

1. $U_\infty^h = \mathbb{E}[h(Y_{\{1,2;1,2\}})]$,
2. $V = \frac{4}{c}\mathrm{Cov}\big(h(Y_{\{1,2;1,2\}}), h(Y_{\{1,3;3,4\}})\big) + \frac{4}{1-c}\mathrm{Cov}\big(h(Y_{\{1,2;1,2\}}), h(Y_{\{3,4;1,3\}})\big)$.

De manière plus générale, d'après le théorème de représentation d'Aldous-Hoover (Aldous, 1981), les modèles RCE dissociés correspondent aux modèles qui peuvent s'écrire :

$$Y \overset{\mathcal{L}}{=} Y^*,$$

où pour tout i, j, $Y_{ij}^* = \phi(\xi_i, \eta_j, \zeta_{ij})$ avec ϕ une fonction quelconque et les $\xi_i, \eta_j, \zeta_{ij}$ sont des variables aléatoires i.i.d.. Ces modèles sont ceux qui peuvent être représentés par un graphon (ou $W-$graphe). Cela identifie donc la classe de modèles pour lesquels le Théorème 2.2 s'applique.

3 Application à l'analyse de réseaux

Le TCL est notamment intéressant quand U_N est un estimateur. Les applications les plus directes rentrant dans le cadre des matrices RCE sont les réseaux bipartites. Nous étudions un exemple d'inférence à l'aide de U-statistiques sur des réseaux d'interaction.

3.1 Le modèle WBEDD

Nous considérons le modèle WBEDD (weighted bipartite expected degree distribution), qui est une extension échangeable du modèle à suite de degrés attendus de Chung et Lu (2002) pour les graphes pondérés bipartites. Le modèle EDD suppose que les espérances des poids (nombres de liens) des nœuds sont tirés dans une distribution. Il s'écrit ainsi :

$$\begin{aligned} \xi_i, \eta_j &\overset{iid}{\sim} \mathcal{U}[0,1] \\ Y_{ij} \mid \xi_i, \eta_j &\sim \mathcal{L}(\lambda f(\xi_i) g(\eta_j)), \end{aligned}$$

où f et g sont des fonctions définies sur $[0,1]$ avec $\int f = \int g = 1$ régissant respectivement la distribution des espérances des poids des nœuds en ligne et en colonne, λ est l'espérance de la densité du réseau et $\mathcal{L}(\mu)$ est une loi de probabilité d'espérance μ. On remarque que ce modèle est effectivement RCE.

Le modèle est adapté aux réseaux écologiques bipartites. En effet, pour l'étude des réseaux, les écologues utilisent souvent des modèles nuls, c'est-à-dire des modèles générant des réseaux aléatoires sous une hypothèse nulle $\mathcal{H}_0$ à tester. Pour réaliser leur test, ils comparent une statistique d'intérêt calculée sur le réseau observé par rapport à sa distribution dans les réseaux générés par les modèles nuls. Parmi les modèles nuls les plus répandus, beaucoup génèrent des réseaux avec une distribution de degrés fixée (souvent celle qui est observée). Le modèle WBEDD ressemble à ces modèles dans la mesure où il génère des réseaux à partir d'hypothèses sur les poids des nœuds, les poids étant analogues aux degrés pour les réseaux pondérés. La différence principale est que l'on ne tire pas directement les poids, mais en fait l'espérance de ces poids dans une distribution, indépendemment pour chaque nœud.

3.2 Un test statistique

La fonction f détermine donc la distribution des poids des nœuds en ligne. Par exemple, pour des réseaux d'interaction écologiques, une fonction f non constante signifie qu'il y a un déséquilibre dans les rôles des espèces en ligne, en particulier entre des espèces généralistes (interagissant avec beaucoup d'espèces) et des espèces spécialistes (interagissant avec peu d'espèces). Pour identifier cet effet dans les réseaux observés, on peut réaliser un test de l'hypothèse $\mathcal{H}_0 : f \equiv 1$ contre $\mathcal{H}_1 : f \not\equiv 1$. Pour cet exemple, on dira que la distribution $\mathcal{L}(\mu)$ du modèle WBEDD est la loi de Poisson $\mathcal{P}(\mu)$, adaptée aux données de comptage.

Sous $\mathcal{H}_0$, $F_2 := \int f^2 = 1$ donc $\lambda^2 F_2 = \lambda^2$. Comme précédemment annoncé, en posant $h_1(Y_{\{i_1,i_2;j_1,j_2\}}) = (Y_{i_1j_1}Y_{i_1j_2} + Y_{i_2j_1}Y_{i_2j_2})/2$, on a $\mathbb{E}[h_1(Y_{\{i_1,i_2;j_1,j_2\}})] = \lambda^2 F_2 = \lambda^2$. L'application du théorème donne pour la U-statistique associée :

$$\sqrt{\frac{N}{V^{h_1}}}(U_N^{h_1} - \lambda^2) \xrightarrow[N\to\infty]{\mathcal{L}} \mathcal{N}(0,1)$$

où $V^{h_1} = \lambda^4 c^{-1}(F_4 - F_2^2) + 4\lambda^4(1-c)^{-1}F_2(G_2 - 1) = 4\lambda^4(1-c)^{-1}(G_2 - 1)$ sous $\mathcal{H}_0$, avec les notations $F_4 := \int f^4$, $G_2 := \int g^2$.

On estime λ^2 par la U-statistique associée à $h_2(Y_{[i_1,i_2;j_1,j_2]}) = (Y_{i_1j_1}^2 Y_{i_2j_2}^2 + Y_{i_1j_2}^2 Y_{i_2j_1}^2)/2$ et $\lambda^2 G_2$ par celle qui est associée à $h_3(Y_{[i_1,i_2;j_1,j_2]}) = \frac{1}{2}(Y_{i_1j_1}Y_{i_2j_1} + Y_{i_1j_2}Y_{i_2j_2})$. En définissant

$$\hat{V}_N^{h_1} := \frac{4}{1-c}(U_N^{h_2})^2 \left[\frac{U_N^{h_3}}{U_N^{h_2}} - 1\right]$$

puis en appliquant Slutsky, on obtient

$$T_N := \sqrt{\frac{N}{\hat{V}_N^{h_1}}}(U_N^{h_1} - U_N^{h_2}) \xrightarrow[N\to\infty]{\mathcal{L}} \mathcal{N}(0,1).$$

Ce résultat définit des intervalles de confiance asymptotiques pour T_N qui sera la statistique de test. Ces intervalles de confiances sont alors utilisés en guise d'intervalles d'acceptation du test.

Bibliographie

Aldous, D. J. (1981). Representations for partially exchangeable arrays of random variables. *Journal of Multivariate Analysis*, 11(4), 581-598.

Chung, F. and Lu, L. (2002). The average distances in random graphs with given expected degrees. *Proceedings of the National Academy of Sciences*, 99,15879–15882.

Eagleson, G. K. and Weber, N. C. (1978). Limit theorems for weakly exchangeable arrays. In *Mathematical Proceedings of the Cambridge Philosophical Society*, 84, 123–130. Cambridge University Press.

Hoeffding, W. (1948). A Class of Statistics with Asymptotically Normal Distribution. *The Annals of Mathematical Statistics*, 293–325.

Le Minh, T. (2021). Weak convergence of U-statistics on a row-column exchangeable matrix. *arXiv preprint arXiv: 2103.12597.*

Nandi, H. and Sen, P. (1963). On the properties of U-statistics when the observations are not independent: Part two unbiased estimation of the parameters of a finite population. *Calcutta Statistical Association Bulletin*, 12, 124–148.

Séries Temporelles

Annulé

Modèle de régression et prévision de séries temporelles pour la demande en électricité à Djibouti

Mohamed Abdillahi Isman[1,2] & Anne-Françoise Yao[2] & Paul-Marie Grollemund[2] & Julien Ah Pine[3]

[1] *Université de Djibouti, Djibouti, magouleh@gmail.com*
[2] *Laboratoire de Mathématiques Blaise Pascal, Université Clermont Auvergne, 63178 Aubière, France, magouleh@gmail.com, anne.yao@uca.fr, paul_marie.grollemund@uca.fr*
[3] *Laboratoire ERIC, Université Lyon 2, 69676 Bron, France, julien.ah-pine@univ-lyon2.fr*

Résumé. Le fonctionnement du système électrique repose sur la nécessité d'assurer l'équilibre entre la production et la demande en électricité à chaque instant, pour une grande variété de situations (pannes, maintenance des groupes de production et des infrastructures de réseau etc). L'offre et la demande en électricité à Djibouti n'est toujours pas satisfaite à cause de différents facteurs comme l'augmentation du nombre de ménages, les conditions climatiques et les habitudes de consommation. A Djibouti, aucune étude n'a été réalisée jusqu'à présent. Aussi, notre objectif est de proposer des modèles d'aide à la maîtrise de l'équilibre entre production et demande pour la société nationale d'Électricité De Djibouti (EDD). Dans cet exposé, nous présenterons les résultats de la première étape de l'étude où nous avons testé différentes méthodes permettant de comprendre l'évolution de la demande à Djibouti sur les dix dernières années. Nous nous intéressons ici aux performances de quelques modèles classiques de séries temporelles en amont d'étapes intégrant des co-variables qui seront par la suite définies en accord avec la compagnie d'électricité de Djibouti.

Mots-clés. Modélisation de série temporelle, prévision, Modèles SARIMA, TBATS, NNAR, STL

Abstract. The operation of the electrical system is based on the need to ensure the balance between production and demand for electricity at all times, for a wide variety of situations (breakdowns, maintenance of production units and network infrastructures, etc.). Electricity supply and demand in Djibouti is still not satisfied due to various factors such as the increase in the number of households, climatic conditions and consumption habits. In Djibouti, no study has been carried out so far. Also, our objective is to propose models to help control the balance between production and demand for the National Electricity Company of Djibouti (EDD). In this presentation, we will present the results of the first stage of the study where we tested different methods to understand the evolution of demand in Djibouti over the last ten years. We are interested here in the performances of some classic models of time series upstream of stages integrating co-variables which will be defined thereafter in agreement with the electricity company of Djibouti.

Keywords. Time Series Modeling, Forecasting, SARIMA Models, TBATS, NNAR, STL

1 Introduction

L'électricité a atteint une place très importante dans le monde et en particulier à Djibouti. Il est devenu un contributeur majeur à l'amélioration du niveau de vie des habitants, du développement économique et industriel. Sa demande dépend fortement de l'endroit où elle est utilisée mais la régularité de son offre pose des défis particuliers puisqu'il s'agit d'une énergie difficile à stocker à grande échelle. Elle doit, en permanence, faire l'objet d'un équilibre entre la production et la demande pour répondre au besoin d'approvisionnement d'électricité. De ce fait, le gestionnaire de réseau doivent anticiper la demande en réalisant des prévisions sur

la demande d'électricité. A l'échelle Djiboutienne, la production d'électricité a atteint 627.111 Mwh en 2020 contre 605.150 Mwh en 2019, soit une progression de 3.63% (Djibouti 2020). La part de l'énergie importée de l'Ethiopie en 2017 s'est établie à 508.198 Mwh, soit un accroissement de 13% par rapport à 2016 où l'énergie importée avait atteint 449.510 Mwh. En proportion, la part de l'énergie importée représente 89.6% en 2017, contre 83.6% en 2016 du volume de production. La consommation électrique était, pour sa part, fixée à 517.035 Mwh en 2020, en hausse de 2.9% par rapport à l'énergie consommée en 2019 qui se situait à 502.244 Mwh.

Il en résulte que Djibouti, petit pays situé dans la Corne de l'Afrique avec une superficie de 23 200 km^2 et avec une population de 962 451 habitants, connaît une insuffisance de production d'électricité qui provoque régulièrement de pannes d'électricité et une haute dépendance vis-à-vis des importations énergétique. Les deux tiers de la population vivent en zone urbaine, essentiellement dans la capitale. Le reste de la population vit principalement dans les villes secondaires. Son climat se classe parmi les plus chauds au monde avec une température moyenne annuelle de 30.1° C. En été, de mai à septembre, les températures extérieures atteignent entre 35° à 45° et par conséquent entraînent une forte consommation d'électricité. Pendant cette période, on observe souvent un pic de consommation d'électricité qui est dû à l'utilisation des équipements domestiques liés à la réfrigération (climatisation et ventilation). Ces derniers représentent le principal usage de l'électricité et correspondent à 78.6% de la consommation d'électricité des ménages ((INSD) 2004). Durant ces périodes, le gestionnaire de réseau est fortement sollicité pour faire face aux défis de la forte demande.

Le secteur de l'électricité à Djibouti est un secteur en constante évolution et la demande en électricité est croissante d'une année à l'autre. Aujourd'hui malgré les efforts (l'élargissement du parc de production d'électricité) prises par les responsables de la société de production et de distribution de l'électricité à Djibouti, la demande d'électricité n'est toujours pas satisfaite complètement. Ceci est dû à l'insuffisance des productions en énergie électrique et par conséquent, on voit une panne d'électricité régulière programmée pendant les périodes de pic de consommation d'électricité. Le nombre de client de la société nationale d'électricité (EDD) ne cesse de croître ainsi que la consommation d'électricité au fil du temps. De ce fait, En vue d'assurer une adéquation entre les capacités de production d'électricité et la demande de sa clientèle, il est important pour l'EDD de développer des stratégies d'anticipation de la demande d'électricité. Dans ce contexte, Il est donc fondamental de construire une modélisation de la demande d'électricité à la fois pour décrire finement le présent (demande actuelle) mais aussi pour être en mesure de prédire leur évolution sur un plan technique (une meilleure gestion des données énergétiques et d'analyses, ainsi que dans l'obtention de prévisions plus précises).

2 Méthodes de modélisation et prévision

Nous décrivons dans cette partie quelques méthodes de prévision que nous avons utilisé dans la prévision de la demande d'électricité de Djibouti.

2.1 Modèle SARIMA

Le Modèle SARIMA est une extension du modèle ARIMA qui prend en charge les séries chronologiques saisonnières. Le modèle ARIMA est une classe de modèles de série chronologique linéaire non stationnaire proposés par Box et Jenkins en 1976. Le principe du modèle ARIMA est que les valeurs futures sont des combinaisons linéaires de valeurs passées.

Une série chronologique saisonnière peut être détectée par le graphique temporel de la série ainsi que par le graphique ACF. La présence de la saisonnalité dans la série donne au graphique temporel de la série une composition ondulatoire. On peut aussi déduire la saisonnalité à partir du comportement de l'ACF qui semble être décroissance sinusoïdal.

Un modèle SARIMA peut s'écrire mathématiquement comme suivant :

$$\phi(L)\Phi(L^S)\Delta^d\Delta_S^D Y_t = \theta(L)\Theta(L^S)\epsilon_t$$

où $\phi(L)$ et $\theta(L)$ sont des polynômes caractéristiques d'ordre p et q respectivement pour la partie autorégressif (AR) et moyenne mobile (MA) non-saisonnière. $\Phi(L^S)$ et $\Theta(L^S)$ sont des polynômes caractéristiques d'ordre P et Q pour la partie autorégressive et moyenne mobile saisonnière. $\Delta^d = (1-L)^d$ et $\Delta_S^D = (1-L^s)^D$ répresentent respectivement les opérateurs de différenciations réguliers et saisonniers s avec d le nombre de fois que la série est différenciée pour éliminer l'effet de la tendance et D le nombre de fois que la série est différenciée pour éliminer l'effet de la saisonnalité sur la série. Y_t est la série étudiée et ϵ_t est l'erreur aléatoire.

La procédure itérative de la modélisation ARIMA proposé par Box et Jenkins sont (i) l'identification du modèle (ii) Estimation des paramètres du modèle choisi et (iii) validation du modèle/vérifications de diagnostic. Ces trois étapes ont été effectuées dans cette étude.

2.2 Méthode STL

La méthode STL (Seasonal-Trend decomposing using Loess) (Mishra, Sriharsha and Zhong 2021) est un algorithme de decomposition d'une série chronologique basé sur une séquence de procédure de lissage. La méthode a été présenté par Cleveland et al. au Journal Officiel de Statistique en 1990. L'algorithme STL utilise la méthode LOESS pour lisser la série chronologique en composants. Si la série chronologique Y_v a la partie tendance, la partie saisonnière et la partie reste notée respectivement T_v, S_v, R_v, pour $v = 1$ à N, alors la forme mathématique de la méthode est donnée par

$$Y_v = T_v + S_v + R_v$$

L'algorithme de décomposition STL consiste en deux procédures récursives.

Une boucle interne qui met à jour la composante tendance et la composante saisonnière de manière itérative à l'aide du lissage LOESS.

- $1^{ère}$ étape : L'enlèvement de la tendance. $Y_v - T_v^{(k)}$. Pour le premier passage $T_v^{(0)} = 0$.
- $2^{ème}$ étape : Cette étape utilise la méthode de LOESS pour lisser les valeurs de chaque cycle de sous-série de la série sans tendance avec $q = n_s$ et $d = 1$. Le résultat est défini comme C_v^{k+1}.
- $3^{ème}$ étape : Un filtre passe-bas composé de filtres de moyenne mobile et de LOESS, est utilisé pour capturer tous les mouvements à basse fréquence depuis C_v^{k+1}. Le résultat est marqué comme L_v^{k+1}.
- $4^{ème}$ étape : Décomposition de cycle de sous-série lissés en detrending : $S_v^{k+1} = C_v^{k+1} - L_v^{k+1}$ pour $v = 1$ à N.
- $5^{ème}$ étape : Désaisonnalisation à partir de la série initiale : $Y_v - S_v^{k+1}$.
- $6^{ème}$ étape : Cette étape utilise la méthode de LOESS pour lisser la série désaisonnalisée de l'étape 5 pour obtenir la composante de tendance avec $q = n_t$ et $d = 1$. (q est la fenêtre de lissage et d est le degré du polynôme).

Une boucle externe qui calcule les coefficients de robustesse pour minimiser l'effet des valeurs aberrantes (elle remplacera les LOESS à la deuxième et sixième étapes de la boucle intérieure avec le robuste LOESS).

2.3 Modèle TBATS

TBATS (Trigonometric seasonality, Box-Cox transformation, ARMA errors, Trend and Seasonal components) (BROŻYNA et al. 2018) est un modèle de série chronologique qui est utile pour traiter des données qui affichent un comportement saisonnier complexe. La modélisation de la saisonnalité est faite à l'aide d'une représentation trigonométrique basée sur les séries de Fourier.

Modèle :

$$y_t^{(\lambda)} = l_{t-1} + \phi b_{t-1} + \sum_{i=1}^{T} S_{t-m_i} + d_t$$

$$l_t = l_{t-1} + \phi b_{t-1} + \alpha d_t$$

$$b_t = b_{t-1} + \beta d_t$$

$$dt = \sum_{i=1}^{p} \varphi_i d_{t-i} + \sum_{i=1}^{q} \theta_i e_{t-i} + e_t$$

Où

- $y_t^{(\lambda)}$: série temporelle au moment t (transformé Box-Cox)
- $S_t^{(i)}$: i-ème composante saisonnière
- l_t : niveau local
- b_t : tendance avec amortissement
- d_t : processus $ARMA(p, q)$ pour les résidus
- e_t : gaussien avec du bruit

Partie saisonnière :

$$S_t^{(i)} = \sum_{j=1}^{k_i} S_{j,t}^{(i)}$$

$$S_{j,t}^{(i)} = S_{j,t-1}^{(i)} cos(\omega_i) + S_{j,t-1}^{*(i)} sin(\omega_i) + \gamma_1^{(i)} d_t$$

$$S_{j,t}^{*(i)} = -S_{j,t-1}^{(i)} sin(\omega_i) + S_{j,t-1}^{*(i)} cos(\omega_i) + \gamma_2^{(i)} d_t$$

$$\omega_i = \frac{2\pi}{m_i}$$

Paramètres du modèle :

- T : quantité de saisonnalités
- m_i : durée de la i-ème période saisonnière
- k_i : quantité d'harmoniques pour la i-ème période saisonnière
- λ : transformation Box-Cox
- α, β : lissage
- φ_i, ϕ_i : $ARMA(p, q)$ coefficients
- ϕ : tendance amortie
- $\gamma_1^{(i)}$, $\gamma_2^{(i)}$: lissage saisonnier (deux pour chaque période)

2.4 Modèle NNAR

Le modèle NNAR (Neural Network Auto-Regression) (Yu et al. 2021) est un réseau de neurones Feed-Forward avec une seule couche cachée et des valeurs décalées du série chronologique utilisées comme entrées. Le modèle est noté comme $NNAR(p, P, k)_T$ où p est le nombre d'entrées des valeurs décalées de la série, P les décalages saisonniers, k le nombre de nœuds dans la couche cachée et T la longueur de la période saisonnière. La formule du modèle s'écrit comme

$$y_t = f(y_{t-1}, y_{t-2}, ..., y_{t-p}, y_{t-T}, y_{t-2T}, ..., y_{t-PT}) + \epsilon_t$$

où f est la fonction d'activation sigmoïde avec une seule couche cachée de k neurone et ϵ_t est la série résiduelle.

Les valeurs p, P et k sont trouvées automatiquement par la fonction si elles ne sont pas spécifiés. Pour les séries temporelles saisonnières, la valeur par défaut est $P = 1$, et p est choisi à partir du modèle linéaire optimal ajusté aux données désaisonnalisées. Si k n'est pas spécifié, alors il est défini sur $k = (p + P + 1)/2$ (arrondi à l'entier le plus proche).

2.5 Métriques d'évaluations des performances de cinq modèles

Les modèles ajustés sur l'ensemble d'entraînement ont été utilisés pour prévoir à l'avance 36 mois. Trois indicateurs ont été calculés pour mesurer l'exactitude de la prévision pour les cinq modèles : l'erreur quadratique moyenne (RMSE), erreur absolue moyenne (MAE) et erreur absolue moyenne en pourcentage (MAPE). Ces indicateurs sont exprimés comme suit :

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2} \quad , \qquad MAE = \frac{1}{n}\sum_{t=1}^{n}|y_t - \hat{y}_t| \qquad \text{et} \qquad MAE = \frac{1}{n}\sum_{t=1}^{n}\frac{|y_t - \hat{y}_t|}{y_t}$$

où y_t est la série observée au temps t et $\hat{y}_t$ la série ajustée.

3 Résultats

3.1 Données de l'étude

Les données utilisées dans cette étude sont la demande d'électricité collectés à partir de différentes postes de transformation installées dans la capitale de Djibouti. Ces données sont celles enregistrées à chaque mois de l'année sur la période de 2012 à 2020 par l'Électricité de Djibouti (EDD), qui est la plus grande compagnie de production et de distribution d'électricité à Djibouti. Nous nous sommes intéressés particulièrement les clients qui sont connectés au réseau électrique à basse tension (BT) avec une puissance souscrite entre 1kva et 9kva. Les données de l'ensemble d'apprentissage utilisées pour la construction de modèles sont de la période de janvier 2012 à décembre 2017 et celles de l'ensemble de test sont de la période de janvier 2018 à décembre 2020.

3.2 Comparaison de performance des modèles

Nous avons comparé pour l'instant cinq modèles dont les modèles SARIMA, NNAR, TBATS, STL et la combination SARIMA-NNAR-TBATS-STL qui est le modèle donnant la moyenne des prévisions par les modèles précédents. Les métriques d'erreur de prévision montrent que le modèle SARIMA est plus performant (en sens de RMSE, MAE ou MAPE) pour prédire la demande en électricité au moins pour la période 2018-2020. Ces mesures sont calculées à partir de l'ensemble de test, donc uniquement pour les observations jusqu'à la ligne noire en pointillés dans le graphique ci-dessous. Les observations après cette ligne pointillée noire sont les prévisions futures de la demande d'électricité.

Table 1: Comparaisons de performance des cinq modèles

	RMSE	MAE	MAPE
STL	11.066197	9.807536	13.728329
SARIMA	2.608943	2.177854	3.267496
NNAR	4.846155	4.001450	5.732787
TBATS	4.531631	3.905998	5.745926
Combination	4.791855	3.886614	5.486039

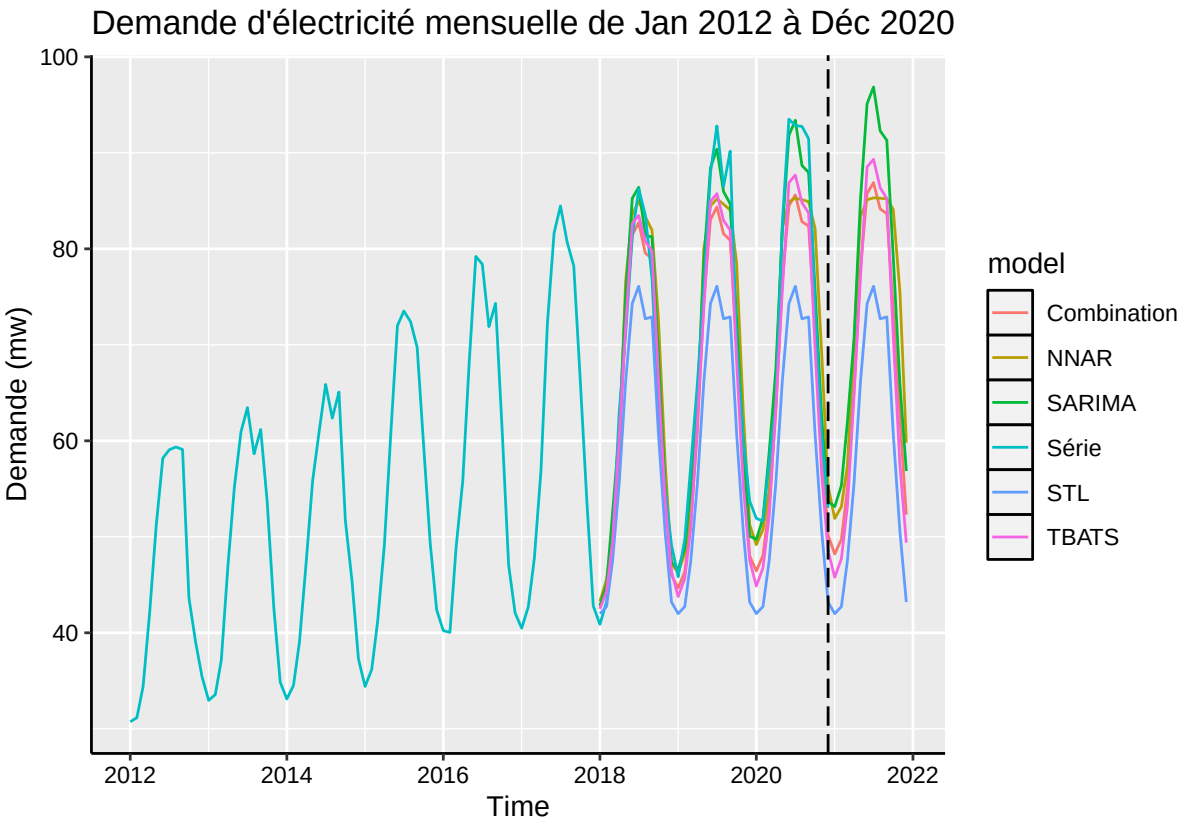


Figure 1: Prévision avec cinq méthodes

Remerciements : Ce travail a été soutenu par la compagnie d'électricité de Djibouti en collaboration avec le Laboratoire de Mathématiques Blaise Pascal (LMBP) de l'Université Clermont Auvergne en France.

References

Aydemir, P. (2018) *Performance of ensemble forecasting tools for analysis turkish consumer price index.* Master's thesis, Middle East Technical University.

BROŻYNA, J., Mentel, G., Szetela, B. and Strielkowski, W. (2018) Multi-seasonality in the TBATS model using demand for electric energy as a case study. *Economic Computation & Economic Cybernetics Studies & Research*, 52 (1).

Djibouti, É. de (2020) *Rapport annuel de la compagnie d'éléctricité de djibouti.*

Godahewa, R., Bergmeir, C., Webb, G.I. and Montero-Manso, P. (2020) A strong baseline for weekly time series forecasting. *arXiv preprint arXiv:2010.08158*,.

(INSD), I.N. de la S. de D. (2004) *Enquête djiboutienne auprès des ménages - energie (EDAM-energie).* INSD.

Kalimoldayev, M., Drozdenko, A., Koplyk, I., Marinich, T., Abdildayeva, A. and Zhukabayeva, T. (2020) Analysis of modern approaches for the prediction of electric energy consumption. *Open Engineering*, 10 (1), pp. 350–361.

Khan, F.M. and Gupta, R. (2020) ARIMA and NAR based prediction model for time series analysis of COVID-19 cases in india. *Journal of Safety Science and Resilience*, 1 (1), pp. 12–18.

Mishra, A., Sriharsha, R. and Zhong, S. (2021) OnlineSTL: Scaling time series decomposition by 100x. *arXiv preprint arXiv:2107.09110*,.

Nwokike, C.C., Offorha, B.C., Obubu, M., Ugoala, C.B. and Ukomah, H.I. (2020) Comparing SANN and SARIMA for forecasting frequency of monthly rainfall in umuahia. *Scientific African*, 10, p. e00621.

Ollech, D. (2021) Seasonal adjustment of daily time series. *Journal of Time Series Econometrics*, 13 (2), pp. 235–264.

Perone, G. (2021) Comparison of ARIMA, ETS, NNAR, TBATS and hybrid models to forecast the second wave of COVID-19 hospitalizations in italy. *The European Journal of Health Economics*, pp. 1–24.

Yu, G., Feng, H., Feng, S., Zhao, J. and Xu, J. (2021) Forecasting hand-foot-and-mouth disease cases using wavelet-based SARIMA–NNAR hybrid model. *Plos one*, 16 (2), p. e0246673.

Zhou, K., Wang, W., Huang, L. and Liu, B. (2021) Comparative study on the time series forecasting of web traffic based on statistical model and generative adversarial model. *Knowledge-Based Systems* [Post-print], 213, p. 106467. Available from https://www.sciencedirect.com/science/article/pii/S0950705120305967.

Estimation Adaptative des Variances dans un Modèle Espace-État

Joseph de Vilmarest [1] & Olivier Wintenberger [2]

[1] *LPSM, Sorbonne Université and EDF R&D,*
joseph.de_vilmarest@sorbonne-universite.fr
[2] *LPSM, Sorbonne Université,* olivier.wintenberger@sorbonne-universite.fr

Résumé. Nous considérons la prévision adaptative de séries temporelles. Nous étudions un modèle espace-état linéaire gaussien dans lequel les variances sont inconnues, et évoluent potentiellement au cours du temps. Nous représentons les variances comme des variables latentes auxiliaires dans un modèle espace-état augmenté en mode *tracking*. L'inférence repose sur le cadre *online Variational Bayes*, qui consiste à approcher la loi jointe a posteriori par la meilleure distribution produit, la meilleure au sens de la divergence de Kullback-Leibler (KL). Nous obtenons l'algorithme Viking, qui optimise la divergence KL à chaque étape de manière récursive. Cet algorithme est robuste car il fonctionne sur des données mal spécifiées, au sens où elles ne sont pas générées par le modèle espace-état considéré.

Mots-clés. Modèles espace-état, séries temporelles

Abstract. We consider the problem of time series forecasting in an adaptive setting. We focus on the inference of state-space models under unknown and potentially time-varying noise variances. We introduce an augmented model in which the variances are represented as auxiliary latent variables in a tracking mode. The inference relies on the online Variational Bayes methodology. It consists in estimating the posterior distribution with the best product distribution in the sense of the Kullback-Leibler (KL) divergence. We derive Viking, a recent algorithm minimizing the KL divergence at each step recursively. Viking is robust to misspecification, in the sense the data-generating process does not need to be the state-space model.

Keywords. State-space models, time series

1 Introduction

On se place dans le cadre de la prévision adaptative de série temporelle. On cherche à prévoir de façon itérative $y_t \in \mathbb{R}$ à l'aide de variables explicatives $x_t \in \mathbb{R}^d$ ainsi que des valeurs passées $(x_s, y_s)_{s<t}$. L'objectif est de produire une prévision $\hat{y}_t$ à chaque étape en vue de minimiser l'erreur quadratique $(y_t - \hat{y}_t)^2$.

Algorithm 1: Filtre de Kalman

1. *Initialisation*: $\hat{\theta}_1 \in \mathbb{R}^d$, P_1 positive semi-définie.
2. *Itération*: pour tout t,
   (a) Prévision: $\hat{y}_t = \hat{\theta}_t^\top x_t$.
   (b) Mise à jour de P: $P_{t|t} = P_t - \frac{P_t x_t x_t^\top P_t}{x_t^\top P_t x_t + \sigma_t^2}$.
   (c) Mise à jour de $\hat{\theta}$: $\hat{\theta}_{t+1} = \hat{\theta}_t + P_{t|t} x_t (y_t - \hat{\theta}_t^\top x_t)$.
   (d) Mise à jour de P: $P_{t+1} = P_{t|t} + Q_t$.

1.1 Modèle espace-état

La régression linéaire est un modèle statique qui permet de prévoir y_t sachant x_t. Cela consiste à estimer y_t par $\theta^\top x_t$ pour un certain $\theta \in \mathbb{R}^d$. Le paramètre θ peut être obtenu, entre autres, par moindres carrés ordinaires qui donnent l'optimum de la perte empirique, ou bien par régression LASSO ou Ridge qui régularisent les moindres carrés ordinaires.

Dans un cadre dynamique, on considère le modèle espace-état suivant:

$$\begin{aligned} \text{Equation d'espace:} \quad & y_t = \theta_t^\top x_t + \varepsilon_t, \\ \text{Equation d'état:} \quad & \theta_{t+1} = \theta_t + \eta_t, \end{aligned}$$

où $(\varepsilon_t)_t$ et $(\eta_t)_t$ sont des bruits blancs gaussiens centrés de variance/covariance respectives σ_t^2 et Q_t. Lorsque le modèle est bien spécifié, et si les paramètres σ_t^2, Q_t ainsi qu'un prior $\theta_1 \sim \mathcal{N}(\hat{\theta}_1, P_1)$ sont connus, le filtre de Kalman (Algorithme 1) permet d'obtenir de façon récursive les valeurs de

$$\hat{\theta}_t = \mathbb{E}[\theta_t \mid (x_s, y_s)_{s<t}], \qquad P_t = \mathbb{E}[(\theta_t - \hat{\theta}_t)(\theta_t - \hat{\theta}_t)^\top \mid (x_s, y_s)_{s<t}].$$

On en déduit alors $y_t \sim \mathcal{N}(\hat{\theta}_t^\top x_t, \sigma_t^2 + x_t^\top P_t x_t)$. Voir par exemple Durbin et Koopman (2012) pour la preuve des formules de Kalman.

1.2 Estimation des variances

Cependant, dans la plupart des applications pratiques, les variances σ_t^2, Q_t ne sont pas connues, et le modèle espace-état est souvent mal-spécifié, au sens où les données ne sont pas généré par un modèle espace-état linéaire gaussien. Il n'y a pas de consensus pour estimer ces variances, mais les méthodes existantes se divisent essentiellement en deux parties. Une méthode consiste à supposer que les variances sont constantes au cours du temps, et l'on cherche à maximiser la vraisemblance (voir par exemple Durbin et Koopman

2012, ou Brockwell et Davis, 2016). Un autre paradigme s'intéresse à l'estimation en ligne des variances, et ces méthodes sont souvent appelées *filtres de Kalman adaptatif* (Mehra, 1972).

Une motivation à l'utilisation de variances dynamiques est la crise covid. En effet, lors de périodes non stationnaires telles que celle que nous vivons, l'hypothèse de variances constantes est moins satisfaisante, et il est intéressant de considérer des variances qui évoluent au cours du temps.

2 Viking: Variational Bayesian Variance Tracking

On présente ici une approche d'estimation des variances introduite dans (de Vilmarest et Wintenberger, 2021).

2.1 Modèle espace-état augmenté

Une manière d'estimer les variances σ_t^2 et Q_t dynamiquement est de les introduire comme de nouvelles variables latentes dans un modèle espace-état augmenté. Nous choisissons un prior gaussien pour les variances, de même que pour l'état. Cependant une variance est positive, donc nous appliquons des transformations aux variable latentes gaussiennes afin de retrouver les variances. Précisément nous modélisons $\sigma_t^2 = \exp(a_t)$ et $Q_t = f(b_t)$, où a_t, b_t suivent des distributions gaussiennes. Finalement, notre modèle dynamique s'écrit de la manière suivante:

$$\begin{aligned}
&\theta_0 \sim \mathcal{N}(\hat{\theta}_0, P_0)\,, \quad a_0 \sim \mathcal{N}(\hat{a}_0, s_0)\,, \quad b_0 \sim \mathcal{N}(\hat{b}_0, \Sigma_0)\,,\\
&a_t - a_{t-1} \sim \mathcal{N}(0, \rho_a)\,, \quad b_t - b_{t-1} \sim \mathcal{N}(0, \rho_b I)\,,\\
&\theta_t - \theta_{t-1} \sim \mathcal{N}(0, f(b_t))\,,\\
&y_t - \theta_t^\top x_t \sim \mathcal{N}(0, \exp(a_t))\,.
\end{aligned}$$

Nous supposons implicitement que:

$$p(\theta_t, a_t, b_t \mid \theta_{t-1}, a_{t-1}, b_{t-1}) = p(\theta_t \mid \theta_{t-1}, b_t) p(a_t \mid a_{t-1}) p(b_t \mid b_{t-1})\,.$$

2.2 Variational Bayes: l'algorithme Viking

Nous appliquons l'approche bayésienne pour estimer la loi jointe de l'état θ_t ainsi que celle des variances représentées par les variables latentes a_t, b_t. Cependant la distribution a posteriori n'admet pas de forme analytique, et nous suivons l'approche Variational Bayes introduite par Smidl et Quinn (2006).

On cherche à estimer la distribution a posteriori par une distribution factorisée de la forme $\mathcal{N}(\hat{\theta}_{t|t}, P_{t|t})\mathcal{N}(\hat{a}_{t|t}, s_{t|t})\mathcal{N}(\hat{b}_{t|t}, \Sigma_{t|t})$. On estime la meilleure forme factorisée au

sense de la divergence de Kullback-Leibler (KL). Précisément, on cherche les valeurs de $\hat{\theta}_{t|t}, P_{t|t}, \hat{a}_{t|t}, s_{t|t}, \hat{b}_{t|t}, \Sigma_{t|t}$ qui minimisent

$$KL\Big(\mathcal{N}(\hat{\theta}_{t|t}, P_{t|t})\mathcal{N}(\hat{a}_{t|t}, s_{t|t})\mathcal{N}(\hat{b}_{t|t}, \Sigma_{t|t}) \;||\; p(\cdot \mid \mathcal{F}_t)\Big),$$

avec $KL(p \;||\; q) = \int \log(dp/dq)dp$. À chaque étape, l'approche variationnelle est un problème d'optimisation couplé entre les trois distributions.

L'optimisation en $\hat{\theta}_{t|t}, P_{t|t}$ admet une forme analytique:

Théorème 1. *À $\hat{a}_{t|t}, s_{t|t}, \hat{b}_{t|t}, \Sigma_{t|t}$ fixés, le minimum de la divergence de Kullback-Leibler est atteint en*

$$\begin{aligned}
A_t &= \mathbb{E}_{b_t \sim \mathcal{N}(\hat{b}_{t|t}, \Sigma_{t|t})}[(P_{t-1|t-1} + f(b_t))^{-1}]\,, \\
P_{t|t} &= A_t^{-1} - \frac{A_t^{-1} x_t x_t^\top A_t^{-1}}{x_t^\top A_t^{-1} x_t + \exp(\hat{a}_{t|t} - \frac{1}{2} s_{t|t})}\,, \\
\hat{\theta}_{t|t} &= \hat{\theta}_{t-1|t-1} + \frac{P_{t|t} x_t}{e^{\hat{a}_{t|t} - s_{t|t}/2}}(y_t - \hat{\theta}_{t-1|t-1}^\top x_t)\,.
\end{aligned}$$

Notons que ces formules de mises à jour sont proches de celles du filtre de Kalman à variances connues σ_t^2 et Q_t. La différence est que l'on a remplacé σ_t^2 par $\exp(\hat{a}_{t|t} - \frac{1}{2} s_{t|t})$ qui est égal à $\mathbb{E}_{a_t \sim \mathcal{N}(\hat{a}_{t|t}, s_{t|t})}[\exp(a_t)^{-1}]^{-1}$. De la même manière, on a remplacé $P_{t-1|t-1} + Q_t$ par $\mathbb{E}_{b_t \sim \mathcal{N}(\hat{b}_{t|t}, \Sigma_{t|t})}[(P_{t-1|t-1} + f(b_t))^{-1}]^{-1}$. Dans le cas où $s_{t|t} = 0, \Sigma_{t|t} = 0$ alors on connait les variances et on ré-obtient bien le filtre de Kalman avec $\sigma_t^2 = \exp(\hat{a}_{t|t})$ et $Q_t = f(\hat{b}_{t|t})$.

Dans le cas contraire, si $\Sigma_{t|t} \neq 0$, cela montre qu'appliquer le filtre de Kalman avec des estimateurs non biaisés des variances est sous-optimal au sens de la divergence de Kullback-Leibler.

Le théorème 1 donne des formes closes pour l'optimum de la divergence de KL en $\hat{\theta}_{t|t}, P_{t|t}$. Nous n'obtenons pas de formes analytiques pour l'optimisation en $\hat{a}_{t|t}, s_{t|t}$ ni en $\hat{b}_{t|t}, \Sigma_{t|t}$. Pour obtenir un algorithme récursif nous utilisons des bornes d'ordre deux, puis les deux premiers moments de la distribution gaussienne. Nous optimisons des bornes supérieures de la KL, ce qui ne garantit pas de minimiser la divergence KL mais au moins de réduire la divergence KL instantanée à chaque étape.

Nous référons à (de Vilmarest et Wintenberger, 2021) pour le détail de l'algorithme nommé Viking (Variational Bayesian Variance Tracking).

Bibliographie

Brockwell, Peter J. and Richard A. Davis (2016). *Introduction to time series and forecasting*. Springer.

Durbin, J. and Koopman, S. J. (2012). *Time series analysis by state space methods*, Oxford university press.
Mehra, Raman (1972). "Approaches to adaptive filtering". *IEEE Transactions on Automatic Control* 17.5, pp. 693–698.
Smidl, Vaclav and Anthony Quinn (2006). *The variational Bayes method in signal processing*. Springer Science & Business Media.
de Vilmarest, Joseph, and Olivier Wintenberger (2021). "Viking: Variational Bayesian Variance Tracking." *arXiv preprint* arXiv:2104.10777 (2021).

Annulé

On multi-change-points tests based on multi-samples U-Statistics for weakly dependent observations

Echarif Elharfaoui [1] & Michel Harel [2] & Joseph Ngatchou-Wandji [3]

[1] *Université Chouaib Doukkali, Faculté des Sciences, 24000 El Jadida, Maroc*
elharfaoui.e@ucd.ac.ma
[2] *Université de Limoges, France et* [3] *Université de Lorraine, France*

Abstract. In this paper, we study multi-change-points detection test using multi-sample tests based on U-statistics for absolutely regular observations. Our results extend the results of Ngatchou-Wandji et al. (2022) based on U-statistics for absolutely regular observations only in the case of one change-point detection. The asymptotic distributions of the test statistics under the null hypothesis and under the local alternatives are given explicitly and the tests are shown to be consistent.

Keywords. U-statistics, non-parametric change-point tests, weak invariance, weak dependence, Wiener process..

1 Introduction

We are interested in detecting possible differences between the distributions of real-valued random variables $X_1, X_2, ..., X_n$. In practice, this issue can be of primary importance for data from industrial quality control, financial market, epidemic, medical diagnostics, hydrology, climatology etc.
Some models are built to detect a given number of breaks, at unknown positions.

U-statistics in change-point analysis have been introduced by Csörgő and Horváth. They use U-statistics to detect a change in the distribution in a sequence of independent real valued random variables. Further details can be found in Csörgő and Horváth (1997).

These last years there is a growing interest in change-points study in time series data. Most of the techniques and approaches used are mainly based either on testing for the existence of changes or for their locations, or on estimating the locations. Some relevant references on change-points estimation are Döring (2010, 2011), Chen et al. (2011), Dehling et al. (2015).

Detecting multiple change points in a very long sequence has emerged as an important problem that has attracted more and more attention recently.

This issue is considered in this paper where the tests studied are derived from basic processes of the general form

$$Z_n^*(\lambda^1, \lambda^2, ..., \lambda^k) = n^{-3/2} \sum_{l=1}^{k} \sum_{i=[n\lambda^{l-1}]}^{[n\lambda^l]} \sum_{j=[n\lambda^l]+1}^{[n\lambda^{l+1}]} h(X_i, X_j),\ 0 \le \lambda^l \le 1,\ 1 \le l \le k+1,$$

where by convention $[n\lambda^0] = 1$ and $[n\lambda^{k+1}] = n$, with $h : \mathbb{R}^2 \to \mathbb{R}$ a kernel function, which is a generalization of the process defined in Ngatchou-Wandji et al. (2021) from the one-change-point to the multi-change-points.

Our conditions are described here by our mixing assumption on the data, and our study is done in a Skorohod space. Furthermore, besides the Kolmogorov-Smirnov (KS) type test usually studied in the literature, we study a Cramer-von Mises (CM) version which has the advantage that its theoretical limiting distribution under the null hypothesis can be approximated for any kernel h. Ngatchou-Wandji et al. (2022) studied to the classical case of one change-point detection. We generalize their results for the case of multi-change-points. The paper is organized as follows. In Section 2, we define useful quantities such as the test statistics, and we list some assumptions. In Section 3 we study the asymptotic properties of our tests statistics under the null hypothesis, under a sequence of local alternatives and under fixed alternatives.

2 General definitions and assumptions

For cumulative distribution functions Q and R, denote by $\theta(Q, R)$ the following real number

$$\theta(Q, R) = \int\int h(x, y)dQ(x)dR(y).$$

For any $i = 1, 2, ..., n$, let F_i be the cumulative distribution function of X_i. We aim to check all possible differences between the F_i's. We study this problem by testing the hypothesis $\mathcal{H}_0$ against the alternative $\mathcal{H}_{1,k}$, defined respectively by

$\mathcal{H}_0$: The distribution function of X_i is F_1 for all $i \in \{1, 2, ..., n\}$

$\mathcal{H}_{1,k}$: There exists $k + 2$ real values $: 0 = \lambda_0^0 < \lambda_0^1 < \lambda_0^2 < ... < \lambda_0^k < \lambda_0^{k+1} = 1$ such that the distribution function of X_i is F_l if $[n\lambda_0^{l-1}] \leq i < [n\lambda_0^l]$, $1 \leq l \leq k + 1$ and the distribution of X_n is F_{k+1}, there exists some l, $1 \leq l \leq k + 1$ such as $\theta(F_1, F_1) \neq \theta(F_1, F_l)$.

In order to evaluate the capacity of the tests to detect weak changes, we also consider the local alternatives $\mathcal{H}_{1,k,n}$ of the form

$\mathcal{H}_{1,k,n}$: There exists $k + 2$ real values $: 0 = \lambda_0^0 < \lambda_0^1 < \lambda_0^2 < ... < \lambda_0^k < \lambda_0^{k+1} = 1$ such that $F_{[n\lambda^l]+1}, F_{[n\lambda^l]+2}, ..., F_{[n\lambda^{l+1}]} = F_{l+1}^{(n)}$, $0 \leq l \leq k$, the distribution function of X_n is $F_{k+1}^{(n)}$, there exists some l, $1 \leq l \leq k + 1$ such as $\theta(F_1, F_l^{(n)}) = \theta(F_1, F_1) + n^{-1/2}[A_l + o(1)]$, where A_l is some constant.

Remark 1 *Particular examples of local alternatives $\mathcal{H}_{1,k,n}$ are those for which there exists a constant γ_l such that $F_l^{(n)}(x) = F_1(x + n^{-1/2}\gamma_l)$ and the kernel function h is twice differentiable with finite integral $\int\int(\partial h(x, y)/\partial y)dF_1(x)dF_1(y)$ and bounded second-order de-*

rivatives $\partial^2 h(x,y)/\partial^2 y$. This can be checked easily by a suitable application of the Taylor-Young formula. With this example, one finds, for any l

$$A_l = -\gamma_l \int\int \frac{\partial h}{\partial u}(x,u) dF_1(x) dF_1(u).$$

In the purpose of solving our testing problem, the tests we are going to use are based on the following KS and CM statistics for the multi-change-points

$$T_{1,n} = \max_{1<m_1<m_2<...<m_k<n} \left| n^{-3/2} \sum_{l=1}^{k} \sum_{i=m_{l-1}}^{m_l} \sum_{j=m_l+1}^{m_{l+1}} \left\{ h(X_i, X_j) - \theta(\widehat{F}_1, \widehat{F}_1) \right\} \right|, \tag{1}$$

$$T_{2,n} = \frac{1}{n} \sum_{1<m_1<m_2<...<m_k<n} \left\{ n^{-3/2} \sum_{l=1}^{k} \sum_{i=m_{l-1}}^{m_l} \sum_{j=m_l+1}^{m_{l+1}} \left\{ h(X_i, X_j) - \theta(\widehat{F}_1, \widehat{F}_1) \right\} \right\}^2 \tag{2}$$

where by convention $m_0 = 1$ and $m_{k+1} = n$ and where $\widehat{F}_1$ stands for any consistent estimator of F_1, a simple example being the empirical cumulative distribution function. Denote by $[x]$ the integer part of any real number x. Noting that for any $k \in \{1, ..., n-1\}$, there exists $\lambda_*^l \in [0,1]$ $(1 \leq l \leq k)$ such that $k = [n\lambda_*^l]$, one can write, asymptotically,

$$T_{1,n} = \sup_{0 \leq \lambda^1 < \lambda^2 < ... < \lambda^k \leq 1} |Z_n(\lambda^1, ..., \lambda^k)| \quad \text{and} \quad T_{2,n} = \int_0^1 ... \int_0^1 Z_n^2(\lambda^1, ..., \lambda^k) d\lambda^1 ... d\lambda^k$$

where Z_n stands for the following stochastic process

$$Z_n(\lambda^1, ..., \lambda^k) = n^{-3/2} \sum_{l=1}^{k} \sum_{i=[n\lambda^{l-1}]}^{[n\lambda^l]} \sum_{j=[n\lambda^l]+1}^{[n\lambda^{l+1}]} \{h(X_i, X_j) - \theta(F_1, F_1)\}, \; 0 \leq \lambda^l \leq 1, \; 1 \leq l \leq k+1 \tag{3}$$

where by convention $[n\lambda^0] = 1$ and $[n\lambda^{k+1}] = n$.
Define the following U-statistic U_n with kernel h, and the following functions

$$U_n = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} h(X_i, X_j)$$

$$h_{F_1,1}(x) = \int h(x,y) dF_1(y) - \theta(F_1, F_1), \quad h_{F_1,2}(y) = \int h(x,y) dF_1(x) - \theta(F_1, F_1)$$

$$h_{F_l,1}(x) = \int h(x,y) dF_l(y) - \theta(F_1, F_l), \quad h_{F_l,2}(y) = \int h(x,y) dF_l(x) - \theta(F_1, F_l)$$

$$g_{F_1,F_1}(x,y) = h(x,y) - h_{F_1,1}(x) - h_{F_1,2}(y) + \theta(F_1, F_1),$$

$$g_{F_l,F_l}(x,y) = h(x,y) - h_{F_l,1}(x) - h_{F_l,2}(y) + \theta(F_1, F_l).$$

Consider the Hoeffding decomposition of U_n under $\mathcal{H}_0$

$$U_n = \theta(F_1, F_1) + U_{n,1}^{(1)} + U_{n,2}^{(1)} + U_n^{(2)} \tag{4}$$

where

$$U_{n,1}^{(1)} = n^{-1}\sum_{i=1}^{n} h_{F_1,1}(X_i), \quad U_{n,2}^{(1)} = n^{-1}\sum_{i=1}^{n} h_{F_1,2}(X_i)$$

$$U_n^{(2)} = \frac{2}{n(n-1)} \sum_{1\le i<j\le n} \{h(X_i, X_j) - h_{F_1,1}(X_i) - h_{F_1,2}(X_j)\} - \theta(F_1, F_1).$$

Also, define the following real numbers

$$\sigma_{p,r} = \mathbb{E}\big[h_{F_1,p}(X_1)h_{F_1,r}(X_1)\big] - 2\sum_{i=1}^{\infty} \mathbb{C}ov\big(h_{F_1,p}(X_1)h_{F_1,r}(X_{i+1})\big),\ p, r = 1, 2.$$

We make the following assumptions :

(A1) The sequence $\{X_i\}_{i\in\mathbb{N}}$ is absolutely regular with the rate

$$\beta(n) = \mathcal{O}(\tau^n), \quad 0 < \tau < 1, \tag{5}$$

(A2) $\{X_i\}_{i\in\mathbb{N}}$ is stationary.

(A3) We consider $(Y_i^l)_{1\le i\le n}$ a sequence of stationary and absolute regular random variables with rate (5) and the cumulative distribution function of the Y_i^l's is F_l.

(A4) We consider $(Y_{ni}^l)_{1\le i\le n, n\ge 1}$ a sequence of stationary and absolute regular random variables with cumulative distribution function $F_l^{(n)}(x) = F_1(x + n^{-1/2}\gamma_l)$. We assume the cumulative distribution functions $F_{l,ij}^{(n)}$ and $F_{l,ij}^{*(n)}$ of the (Y_{ni}^l, Y_{nj}^l)'s and (X_i^l, Y_{nj}^l)'s respectively satisfy

$$\lim_{n\to\infty} F_{l,ij}^{(n)}(x, y) = F_{ij}(x, y) \text{ and } \lim_{n\to\infty} F_{l,ij}^{*(n)}(x, y) = F_{ij}(x, y),\ 1 \le i < j \le n$$

where F_{ij} is the cumulative distribution function of (X_i, X_j).

• From Harel and Puri (1994) we recall that a non-necessarily stationary triangular sequence $\{\mathcal{V}_{ni},\ 1 \le i \le n,\ n \ge 1\}$ is absolutely regular if $\beta(m) \to 0$, as $m \to \infty$, where

$$\beta(m) = \sup_{n\in\mathbb{N}} \sup_{m\le n} \max_{1\le j\le n-m} \mathbb{E}\left\{ \sup_{A\in\mathcal{A}_{n,j+m}^{\infty}} \big|\mathbb{P}(A \mid \mathcal{A}_{n,0}^{j}) - \mathbb{P}(A)\big| \right\},$$

with $\mathcal{A}_{n,i}^{j}$ standing for the σ-algebra generated by $\mathcal{V}_{ni}, ..., \mathcal{V}_{nj},\ i, j \in \mathbb{N} \cup \{\infty\}$.

• We consider $(Y_i^l)_{1\le i\le n}$ is for any $l \in \{1, ..., k, k+1\}$ a sequence of absolute regular random variables with the same rate as the sequence $(X_i)_{1\le i\le n}$, under $\mathcal{H}_0$ and the distribution function of Y_i^l is F_l. For any $i, j \in \mathbb{N}$, the absolute regular dependence between Y_i^l and Y_j^l is the same as the dependence between X_i and X_j. The random variables Y_i^1 have the same law as the random variables X_i.

3 Theoretical result

In this section, we state all our theoretical results.

Theorem 1 *Assume the assumptions (A1)-(A3) hold. Then, under $\mathcal{H}_0$, if* $\max\left\{\sup_{ij}\mathbb{E}\left(|h(X_i,X_j)|^{2+\delta}\right), \int\int_{\mathbb{R}^2}|h(x,y)|^{2+\delta}\,dF_1(x)dF_1(y)\right\}<\infty$ *for some $\delta>0$ and the absolute regularity condition (5) is satisfied, then for any $p,r=1,2$, $\sigma_{pr}<\infty$. If in addition $\sigma_{pr}>0$, $1\leq p,r\leq 2$, then $\{Z_n(\lambda^1,\lambda^2,...,\lambda^k);0<\lambda^1<\lambda^2<...<\lambda^k<1\}_{n\in\mathbb{N}}$ converges in distribution towards a mean-zero Gaussian process with representation*

$$Z(\lambda^1,\lambda^2,...,\lambda^k)=\sum_{l=1}^{k}\left\{(\lambda^{l+1}-\lambda^l)W_l(\lambda^l)+\lambda^l(W_{l+1}(\lambda^{l+1})-W_{l+1}(\lambda^l))\right\}$$

where $\{W_l(\lambda^l),W_{l+1}(\lambda^{l+1})\}_{0\leq\lambda^l<\lambda^{l+1}\leq 1}$ is a two-dimensional zero-mean Brownian motion with covariance kernel matrix with entries $\mathbb{C}ov\big(W_l(s),W_{l+l}(t)\big)=\min(s,t)\sigma_{pr}$, $p,r=1,2$.

Theorem 2 *Assume (A1), (A2) and (A4) hold, h is twice differentiable with bounded second-order derivatives $\partial^2h(x,y)/\partial x\partial y$, and the integral $\int\int(\partial h(x,y)/\partial y)dF_1(x)dF_1(y)$ is finite. Then, under $\mathcal{H}_{1,k,n}$, if*

$$\sup_{1\leq i,j\leq n}\mathbb{E}\left(|h(X_i,X_j)|^{2+\delta}\right),\quad \sup_{n\geq 1}\sup_{i,j}\mathbb{E}\left(\left|h(Y_{ni}^l,Y_{nj}^l)\right|^{2+\delta}\right),$$

$$\sup_{n\geq 1}\sup_{1\leq i,j\leq n}\mathbb{E}\left(\left|h(X_i,Y_{nj}^l)\right|^{2+\delta}\right),\quad \int\int_{\mathbb{R}^2}|h(x,y)|^{2+\delta}\,dF_1(x)dF_1(y),$$

$$\sup_{n\geq 1}\int\int_{\mathbb{R}^2}|h(x,y)|^{2+\delta}\,dF_l^{(n)}(x)dF_l^{(n)}(y),\quad \sup_{n\geq 1}\int\int_{\mathbb{R}^2}|h(x,y)|^{2+\delta}\,dF_1(x)dF_l^{(n)}(y)$$

are finite for some $\delta>0$, $1\leq l\leq k+1$, if for any $p,r=1,2$, $\sigma_{pr}>0$, then the sequence of processes $\{Z_n(\lambda^1,\lambda^2,...,\lambda^k);\ 0<\lambda^1<\lambda^2<...<\lambda^k<1\}_{n\in\mathbb{N}}$ converges in distribution towards a Gaussian process $\widetilde{Z}$ with mean $\sum_{l=1}^{k}(\lambda^{l+1}-\lambda^l)(\lambda^l-\lambda^{l-1})A_l$ and representation

$$\widetilde{Z}=Z(\lambda^1,\lambda^2,...,\lambda^k)+\sum_{l=1}^{k}(\lambda^{l+1}-\lambda^l)(\lambda^l-\lambda^{l-1})A_l$$

where by convention $\lambda^0=0$ and where $\{Z(\lambda^1,\lambda^2,...,\lambda^k),\ 0<\lambda^1<\lambda^2<...<\lambda^k<1\}$ is the zero-mean Gaussian process defined in Theorem 1.

Corollary 1 *Suppose that the distribution function of X_i is F_l if $[n\lambda_0^{l-1}]\leq i<[n\lambda_0^l]$, $1\leq l\leq k+1$ and the distribution of X_n is F_{k+1} for all $x\in\mathbb{R}$ and there exists constant B_l, $1\leq l\leq k+1$ such that $F_l(x)=F_1(x+n^{-1/2}B_l)$ for all $x\in\mathbb{R}$ and suppose also the kernel function h is twice differentiable and $\int\int(\partial h(x,y)/\partial y)dF_1(x)dF_l(y)<\infty$ and $\partial^2h(x,y)/\partial^2y$ is bounded, then we are under the alternative $\mathcal{H}_{1,k,n}$.*

Theorem 3 *Assume (A1)-(A3) hold and that under $\mathcal{H}_{1,k}$, the integrability conditions in Theorem 2 hold. Then,*

$$n^{-1/2}Z_n^*(t_1,t_2,...,t_k) \xrightarrow{a.s.}_{n\to\infty} \sum_{l=1}^{k} \left\{\theta(F_l,F_l)t_l(\lambda_0^1-t_l)+\theta(F_l,F_{l+1})t_l(\lambda_0^{l+1}-\lambda_0^l)\right\} \quad (6)$$

if $\lambda_0^{l-1} \leq t_l \leq \lambda_0^l,\ l=1,2,...,k$ *and*

$$n^{-1/2}Z_n^*(t_1,t_2,...,t_k) \xrightarrow{a.s.}_{n\to\infty} \sum_{l=1}^{k} \left\{\theta(F_{l+1},F_{l+1})(t_l-\lambda_0^l)(\lambda_0^{l+1}-t_l)+\theta(F_l,F_{l+1})\lambda_0^l(\lambda_0^{l+1}-t_l)\right\} \quad (7)$$

if $\lambda_0^l \leq t_l \leq \lambda_0^{l+1},\ l=1,2,...,k$, *where by convention* $\lambda_0^0=0$.

We easily deduce the following corollary.

Corollary 2 *Under $\mathcal{H}_0$ and the conditions of Theorem 2,*

$$n^{-1/2}Z_n^*(t_1,t_2,...,t_k) \xrightarrow{a.s.}_{n\to\infty} \theta(F_1,F_1)\sum_{l=1}^{k} t_l(\lambda_0^{l+1}-t_l).$$

Bibliography

Chen KM., Cohen A. and Sackrowitz H. (2011). Consistent multiple testing for change points. *J Multvar Anal*, 102, pp. 1339-1343.

Csörgő M. and Horváth L. (1997). Limit Theorems in Change-Point Analysis, Willey, Chichester, UK.

Dehling H., Fried R., Garcia I. and Wendler M. (2015). Change-Point Detec-tion Under Dependence Based on Two-Sample U-Statistics. Asymptotic Laws and Methods in Stochastics, Springer New York NY.

Döring M. (2010). Multiple change-point estimation with U-statistics. *J Stat Plann Inference*, 140, pp. 2003-2017.

Döring M. (2011). Convergence in distribution of multiple change point estimators. *J Stat Plann Inference*, 141(7), pp. 2238-2248.

Harel M. and Puri M. (1994). Law of the iterated logarithm for perturbed empirical distribution functions evaluated at a random point for nonstationary random variables. *J Theor Probab*, 4, pp. 831-855.

Ngatchou-Wandji J., Elharfaoui E. and Harel M. (2022). On change-points tests based on two-samples U-Statistics for weakly dependent observations. *Stat Pap*, 63(1), pp. 287-316.

Yoshihara, Y. (1976). Limiting behavior of U-statistics for stationary absolutely regular processes. *Z Wahrscheinlichkeitstheorie Verw Gebiete*, 35, pp. 237-252.

On estimating in generalized integer-valued GARCHX models with structural breaks

Mohamed Djemaà SADOUN[1,2] & Abderaouf KHALFI[1,2]

mo-hamedsadoun@outlook.fr *khalfi.doc@gmail.com*

[1] *RECITS Laboratory, Operational Research Department, USTHB, Algiers, Algeria*

[2] *Research Center in Applied Economics for Development (CREAD), Algiers, Algeria*

Résumé. Nous abordons le problème de l'estimation des paramètres dans un modèle $INGARCH$ généralisé à changements structurels incluant des co-variables exogènes (ci-après noté $GCP - INGARCHX$). Cette classe de modèles appartient aux modèles de type "observation-driven" avec changement de régime où le changement de régime est entraîné par certains points de rupture survenant dans le temps. Les estimateurs des moindres carrés conditionnels (MCC) et du maximum de vraisemblance conditionnelle (MVC) des paramètres du modèle sous-jacent sont obtenus dans les deux cas: lorsque les points de rupture sont connus ou non. L'estimation de ces points de rupture est réalisée par une méthode d'estimation hors ligne. Une étude de simulation et une application sur données réelles sont fournies pour évaluer la qualité du modèle.

Mots-clés. Processus de comptage à valeurs entières, modèle $INGARCH$ non stationnaire, estimateurs du (MCC) et (MVC), méthode d'estimation hors ligne.

Abstract. We deal with the estimation problem in a generalized $INGARCH$ model with structural changes including exogenous covariates (hereafter referred to as $GCP - INGARCHX$). This class of models belongs to the observation-driven type models with regimes change where the regime-switching is driven by certain failure points occuring in the time. The conditional least squares (CLS) and the conditional maximum likelihood (CML) estimators of the underlying parameters are obtained for both the cases that the break points are known or not. An off-line estimation method is used to estimate the breaks points. A simulation study and an application on real data set are provided to assess the performance of the model.

Keywords. Integer-valued process of counts, no-stationary $INGARCH$ model, (CLS) estimators, (CML) estimators, off-line estimation method.

1 Introduction and model formulation

In count time series analysis, Poisson distribution is frequently used and provides a classical framework (see Fokianos (2012)). However, due to the fact that many time series of counts encountered in practice exhibits the nonlinearity, the over and/or the under-dispersion, the multimodality, and the non-stationarity, a simple counts model with a regime-switching of change point-type including generalized and more flexible distributions is then needed to capture these characteristics often found in real-life examples. We

propose a generalized $INGARCHX$ regime-switching model, where the regime-switching is driven by a certain latent random variable defined by pieces in time, and where "X" refers to specific exogenous covariates. In this contribution, we consider a simple version of the first-order Generalized Change-Point $INGARCHX$ model (hereafter referred to as $GCP-INGARCHX\,(1,1,q)$). The model is formulated by equation (1) below involving some observable count variable Z_t seen as a mixture of generalized $INGARCHX$ components $Y_{k,t}$. Here the mixture is realized using a sequence of time-segmented variables.

$$\begin{cases} Z_t = \sum_{k=1}^{m+1} \mathbf{1}\,(S_t = k)\, Y_{S_t,t} \\ \mathbb{E}\,(Y_{S_t,t}|\mathcal{F}_{t-1}) = \lambda_{S_t,t}, \\ var\,(Y_{S_t,t}|\mathcal{F}_{t-1}) = \upsilon_{S_t,0}\lambda_{S_t,t} + \upsilon_{S_t,1}\lambda_{S_t,t}^2 \\ \lambda_{S_t,t} = \alpha_{S_t,0} + \alpha_{S_t,1}\, Z_{t-1} + \beta_{S_t}\, \lambda_{S_t,t-1} + \sum_{l=1}^{q} \gamma_{S_t,l} X_{t-l}. \end{cases} \tag{1}$$

where $\mathcal{F}_t$ indicates the information given up to time t. $\upsilon_{S_t,i} \geq 0, i = 0,1$, but not simultaneously equal to zero, $\alpha_{S_t,0} > 0$, $\alpha_{S_t,1} \geq 0$, $\beta_{S_t} \geq 0$, and $\gamma_{S_t,l} \geq 0$ for $l = 1,...,q$. $\mathbf{1}\,(.)$ denotes the indicator function. $\{S_t\}_{t\in\mathbb{Z}}$ is a sequence of independent and identically distributed random variables defined by pieces in time, which is defined as follows:

$$S_t = \begin{cases} 1, & if\ t \leq c_1, \\ 2, & if\ c_1 < t \leq c_2. \\ \vdots & \vdots \\ m & if\ c_{m-1} < t \leq c_m \\ m+1 & if\ t > c_m \end{cases}$$

where $(c_1, c_2, ..., c_m)'$ stand the vector of unknown break points. Furthermore, it is assumed that Z_{t-1} and S_t are independent. Note that $GCP-INGARCHX\,(1,1,q)$ model could be extended to $GCP-INGARCHX\,(P,Q,q)$ model, where the change occurs only on the link function $\lambda_{S_t,t}$ by increasing the order of lags as follows:

$$\lambda_{S_t,t} = \alpha_{S_t,0} + \sum_{i=1}^{P} \alpha_{S_t,i}\, Z_{t-i} + \sum_{j=1}^{Q} \beta_{S_t,j}\, \lambda_{S_t,t-j} + \sum_{l=1}^{q} \gamma_{S_t,l} X_{t-l}. \tag{2}$$

Where $\alpha_{S_t,0} > 0$, $\alpha_{S_t,i} \geq 0$, $\beta_{S_t,j} \geq 0$, and $\gamma_{S_t,l} \geq 0$ for $i = 1,...,P$, $j = 1,...,Q$, and $l = 1,...,q$. While supposing that Z_{t-i} and S_t are independent for all t and $i > 0$.

2 Estimation when break-points are known

For the problem of parameters estimation, we shall focus on the change-point Negative Binomial $INGARCHX$ model and the change-point Generalized Poisson $INGARCHX$ model with following conditional distribution specifications:

$$Y_{S_t,t}|\mathcal{F}_{t-1} \rightsquigarrow \mathcal{NB}\,(\lambda_{S_t,t}, \upsilon_{S_t}) \quad (A.1)\,; \qquad Y_{S_t,t}|\mathcal{F}_{t-1} \rightsquigarrow \mathcal{GP}\left(\left(1-\phi_{S_t}\right)\lambda_{S_t,t}, \phi_{S_t}\right) \quad (A.2)\,.$$

Two methods of estimation will be considered in this paragraph, namely the CLS method and the CML method. Although the first one is less performant than the second one, it remains frequently used nowadays. Indeed, it is simple and it produces consistent estimators which can be used as initial values in sophisticated methods as the maximum likelihood estimation. We note that for the CLS method we used the $ARMA$ presentation of our model to estimate the parameters.

2.1 Conditional least squares method

Assume that $P = Q = 1$, we follow Ferland *et al* (2006) (equation (14)), and Zhu (2010) (equation (16)) to represent the process $\{\lambda_t; t \in Z\}$ under the representation of a segmented $ARMAX\,(1,1,q)$ process. Obtain the CLS estimates imply found the soultion the following minimization problem

$$Q_n\left(\Theta\right) = \sum_{t=2}^{n}\left(Z_t - \widehat{Z}_t\right)^2 = \sum_{t=2}^{n}\left(Z_t - \left(\sum_{k=1}^{m+1}\left(\alpha_{k,0} + \varphi_k Z_{t-1} + \delta_k \varepsilon_{t-1} + \textstyle\sum_{l=1}^{q}\gamma_{k,l}X_{t-l}\right)\mathbf{1}_{t,k}\right)\right)^2$$

with the $(m+1)\times(3+q)$ dimension parameter $\Theta = \left(\theta_1', \theta_2', ..., \theta_{m+1}'\right)'$, such as $\theta_k = \left(\alpha_{k,0}, \alpha_{k,1}, \beta_k, \left(\gamma_{k,1}, ..., \gamma_{k,q}\right)\right)' \in \mathbb{R}^{3+q}$ for $k = 1, ..., m+1$.

The CLS procedure is done in two different steps as follows:

1 Let $W_t = Z_t - \frac{\sum_{i=1}^{n} Z_i}{n}$, we fit the data by using a higher-oder $AR\,(p)$ model, then obtain the CLS estimators for the autoregressive coefficients $\widehat{a}_i$ and define $\widehat{\varepsilon}_t = W_t - \sum_{i=1}^{p}\widehat{a}_i W_{t-i}$.

2 We minimize the quantity

$$Q_n\left(\Theta\right) = \sum_{t=2}^{n}\left(Z_t - \left(\sum_{k=1}^{m+1}\left(\alpha_{k,0} + \varphi_k\, Z_{t-1} + \delta_k\, \widehat{\varepsilon}_{t-1} + \textstyle\sum_{l=1}^{q}\gamma_{k,l}X_{t-l}\right)\mathbf{1}_{t,k}\right)\right)^2$$

to obtain the CLS estimates $\widetilde{\alpha}_{k,0}, \widetilde{\varphi}_k$ and $\widetilde{\delta}_k$, then $\widetilde{\alpha}_{k,1} = \widetilde{\varphi}_k + \widetilde{\delta}_k$ and $\widetilde{\beta}_k = -\widetilde{\delta}_k$ and also $\widetilde{\gamma}_{k,l}, l = 1, ..., q$ with $k = 1, ..., m+1$.

2.2 Conditional maximum likelihood method

Then, we considered two kinds of $GCP-INGARCHX\,(1,1,q)$ models based on different distributions as mentioned above.

For the first case of Negative Binomial $INGARCHX$ model with

$$Y_{S_t,t}|\mathcal{F}_{t-1} \rightsquigarrow \mathcal{NB}\left(\upsilon_{S_t}, \frac{\upsilon_{S_t}}{\upsilon_{S_t} + \lambda_{S_t,t}}\right),$$

the conditional likelihood function of the n observations $Z_1, ..., Z_n$ conditionally on the pre-sample values, is given by

$$L(\Theta) = \prod_{t=1}^{n} \sum_{k=1}^{m+1} \left(\frac{\Gamma(\upsilon_k + Y_{k,t})}{\Gamma(\upsilon_k)\Gamma(Y_{k,t}+1)} \left(\frac{\upsilon_k}{\upsilon_k + \lambda_{k,t}} \right)^{\upsilon_k} \left(\frac{\lambda_{k,t}}{\upsilon_k + \lambda_{k,t}} \right)^{Y_{k,t}} \right) \mathbf{1}_{t,k}$$

$$\Theta = \left(\theta_1', \theta_2', ..., \theta_{m+1}'\right)',$$

with $\theta_k = \left(\alpha_{k,0}, \alpha_{k,1}, \beta_k, \left(\gamma_{k,1}, ..., \gamma_{k,q}\right), \upsilon_k\right)'$ for $k = 1, ..., m+1$

For the second case of Generalized Poisson $INGARCHX$ model with

$$Y_{S_t,t} | \mathcal{F}_{t-1} \rightsquigarrow \mathcal{GP}\left(\left(1 - \phi_{S_t}\right)\lambda_{S_t,t}, \phi_{S_t}\right),$$

the conditional likelihood function of the n observations $Z_1, ..., Z_n$ conditionally on the pre-sample values, is given by

$$L(\Theta) = \prod_{t=1}^{n} \sum_{k=1}^{m+1} \left(\frac{(1-\phi_k)\lambda_{k,t}\left[(1-\phi_k)\lambda_{k,t} + \phi_k Y_{k,t}\right]^{Y_{k,t}-1} e^{-\left[(1-\phi_k)\lambda_{k,t} + \phi_k Y_{k,t}\right]}}{Y_{k,t}!} \right) \mathbf{1}_{t,k}$$

$$\Theta = \left(\theta_1', \theta_2', ..., \theta_{m+1}'\right)',$$

with $\theta_k = \left(\alpha_{k,0}, \alpha_{k,1}, \beta_k, \left(\gamma_{k,1}, ..., \gamma_{k,q}\right), \phi_k\right)'$ for $k = 1, ..., m+1$

To obtain the optimal value of Θ, we suggest some numerical optimization routine while using the (CLS) estimators $\widetilde{\Theta}$ as initial values. When the vector of break or change points $(c_1, c_2, ..., c_m)'$ is unknown, the CLS and CML methods mentioned above cannot be applied immediately since the change points parameter lies in an indicator function. Therefore, we need to estimate $(c_1, c_2, ..., c_m)'$ first. A commonly used approach to estimate the multiple breakpoints is the so-called search minimum least squares $(SMLS)$ algorithm proposed by Bai and Perron (1998). Motivated by this work, we proposed an adapted version of their algorithm based on the principle of the $(SMLS)$ to estimate the m change points in our formulation.

3 Estimation when break-points are unknown

Two approaches are usually considered in the detection and estimation of the break dates: the **on-line** and **off-line** detection. In this work, an **off-line** change point estimation method is used like in Davis *et al* (2006). The purpose is to estimate the unknown coefficients together with the break points when n observations on Z_t are available.

Description of the method:

1. For each m-partition $(c_1, ..., c_m)$, denoted $\{c_k\}$, the associated least squares estimates of θ_k' are obtained by minimizing the sum of squared residuals

$$\sum_{i=1}^{m+1} \sum_{t=c_{i-1}+1}^{c_i} \left(Z_t - \left(\alpha_{i,0} + \varphi_i Z_{t-1} + \delta_i \widehat{\varepsilon}_{t-1} + \textstyle\sum_{l=1}^{q} \gamma_{i,l} x_{t-l}\right)\right)^2$$

2. Let $\widehat{\theta}'(\{c_k\}) = \left(\widehat{\theta}_1', \widehat{\theta}_2', ..., \widehat{\theta}_{m+1}'\right)'$ denote the resulting estimates.
Substituting them in the objective function and denoting the resulting sum of squared residuals as $S_n(c_1, ..., c_m)$, the estimated break points $(\widehat{c}_1, ..., \widehat{c}_m)$ are such that

$$(\widehat{c}_1, ..., \widehat{c}_m) = \arg \min_{c_1,...,c_m} S_n(c_1, ..., c_m)$$

Where the minimization is taken over all partitions $(c_1, ..., c_m)$ such that $c_i - c_{i-1} \geq q$. Thus the change-point estimators are global minimizers of the objective function. Note that the break points need not be obtained via an exhaustive grid search. We develop here an adapted version of an efficient algorithm known as the triangular matrix of sums of squared residuals, which was introduced by Bai and Perron (2003) by implementing the dynamic programming principle. This algorithm allows global minimizers to be obtained using a number of sums of squared residuals that is of order $O(n^2)$ for any $m \geq 2$. In practice, for the choice of m, we generally use a usual (AIC, BIC, and the modified Schwarz criterion (LWZ)) information criterion to estimate the number of breaks.

4 Numerical illustration

Simulation results Assume that $P = Q = 1$ and $q = 0$. We have evaluated the (CLS), and (CML) on 2 time series, of small, moderate and relatively large sizes ($n = 100, 200, 300, 500$), generated from two different $GCP - INGARCH(1,1)$ models. A Negative Binomial conditional distribution and a Generalized Poisson conditional distribution are considered in Models I and II. We report the results for the size $n = 300$. The true parameter values of these models are given as follows :

$$I : \underline{\theta} = [\underline{\theta}_1'; \underline{\theta}_2'; t^*]' = \left[(\alpha_{1,0}, \alpha_{1,1}, \beta_1, \upsilon_1)'; (\alpha_{2,0}, \alpha_{2,1}, \beta_2, \upsilon_2)'; 0.3n\right]'$$
$$\underline{\theta} = \left[(0.3, 0.2, 0.35, 2)'; (0.4, 0.5, 0.2, 3)'; 0.3n\right]'.$$
$$II : \underline{\theta} = [\underline{\theta}_1'; \underline{\theta}_2'; t^*]' = \left[(\alpha_{1,0}, \alpha_{1,1}, \beta_1, \phi_1)'; (\alpha_{2,0}, \alpha_{2,1}, \beta_2, \phi_2)'; 0.3n\right]'$$
$$\underline{\theta} = \left[(0.3, 0.2, 0.35, 0.1)'; (0.4, 0.5, 0.2, 0.3)'; 0.3n\right]'.$$

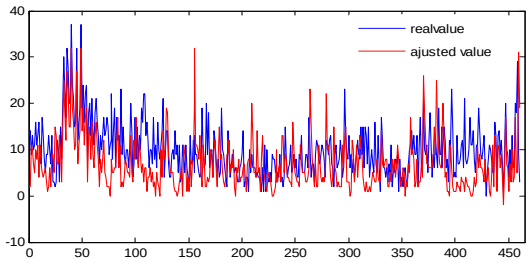
Simulation results for Model I

	k		$\alpha_{k,0}$	$\alpha_{k,1}$	β_k	υ_k
CLS	1	TV	0.3	0.2	0.35	2
		Mean	.2071	.1954	.2653	2.2779
		MSE	.0268	.0174	.0324	.4608
	2	TV	0.4	0.5	0.2	3
		Mean	.4591	.4230	.1573	3.3642
		MSE	.0279	.0241	.0243	.5199
CML	1	TV	0.3	0.2	0.35	2
		Mean	.2395	.2093	.3057	2.0388
		MSE	.0182	.0098	.0191	.3011
	2	TV	0.4	0.5	0.2	3
		Mean	.4321	4395	.1819	3.1010
		MSE	.0186	.0149	.0138	.3125

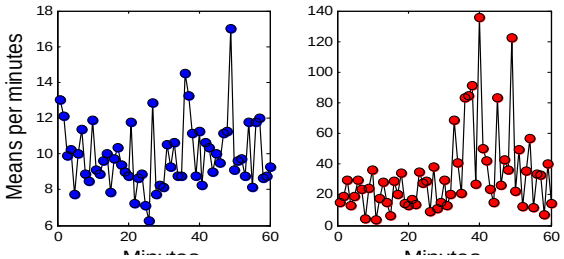
Simulation results for Model II

$\alpha_{k,0}$	$\alpha_{k,1}$	β_k	ϕ_k
0.3	0.2	0.35	0.1
.2806	.1895	.2842	.0936
.0062	.0104	.0176	.0028
0.4	0.5	0.2	0.3
.3745	.4846	.1638	.2900
.0169	.0171	.0183	.0032
0.3	0.2	0.35	0.1
.2907	.1965	.3100	.0943
.0058	.0075	.0161	.0025
0.4	0.5	0.2	0.3
.3825	.5013	.1711	.2984
.0155	.0166	.0167	.0027

Real data example We consider the data set of size 460 observations, consisting on the intra-day counts of number of transactions per minute for the stock Ericsson **B** during first July, 2002. This time series was considered by Diop and Kengne (2017) which represents the transaction of approximately 8 hours (from 09 : 35 until 15 : 14). A change point in the parameter is found at $t^* = \widehat{t} = 143$.



Number of transactions per minute



Sample means and variances

Fit results of some $NB - INGARCH\,(1,1)$ *model*

	$INARCH$	$INGARCH$	$GCP - INARGH$	$GCP - INGARCH$
AIC	709.3	693.2	686.5	650.7
BIC	714.4	701.1	693.8	662.4

Bibliographie

Bai, J. and Perron. P. (1998). Estimation and testing linear models with multiple structural changes. Econoometrica, Vol. 66, No. 1. $47 - 78$.

Bai, J. and Perron. P. (2003). Computation and analysis of multiple structural change models. Journal of Applied Econometrics. (18) $1 - 22$.

Davis, R. A., Lee, T. C. M. and Rodriguez-Yam, J. A. (2006). Structural break estimation for nonstationary time series models. *Journal of the American Statistical Association.* Vol. 101, No. 473. DOI: 10.1198/016214505000000745.

Diop, M. L. and Kengne. W. (2017). Testing Parameter Change in General Integer-Valued Time Series. *Journal of Time Series Analysis.* DOI: $10.1111/jtsa.12240$.

Ferland, R., Latour, A. and Oraichi, D. (2006). Integer-Valued GARCH Process. *J. Time Ser. Anal.*, Vol. 27. No. 6, $923 - 942$.

Fokianos, K. (2012). Count time series models. *In Handbook of Statistics-Time Series Analysis: Methods and Applications, Amsterdam: Elsevier*, $315 - 347$.

Zhu, Fukang. (2010). A negative binomial integer-valued GARCH model. *Journal of Time Series Analysis.* $32 : 54 - 67$. DOI: $10.1111/j.1467 - 9892.2010.00684.x$.

Tight Risk Bound for High Dimensional Time Series Completion

Amélie Rosier [1]

[1] *Laboratoire MODAL'X, Université Paris Nanterre, Nanterre, France et ESME Sudria, Paris, France, amelie.rosier@esme.fr*

Résumé. La complétion de matrices de faible rang a longuement été étudiée au cours des 10 dernières années, mais la théorie sur ce sujet n'a été jusque là développée que pour des observations indépendantes. L'application de ces méthodes aux données dépendantes n'a été que très peu investiguée. Dans cet exposé, nous proposerons une méthode de complétion pour une matrice dont les lignes sont des séries temporelles observées partiellement et dont les termes présentent effectivement une dépendance temporelle. En utilisant la structure de série temporelle, nous établirons une borne de risque plus fine que celle de [1]. Il s'agit d'un travail en collaboration avec Pierre Alquier et Nicolas Marie.
Mots-clés. Complétion de matrices ; Séries temporelles ; Estimation ; Séléction de modèle.

Abstract. Initially designed for independent data, low-rank matrix completion was successfully applied in many domains to the reconstruction of partially observed high-dimensional time series. However, there is a lack of theory to support the application of these methods to dependent datas. In this presentation, we propose a general model for multivariate, partially observed time series. We show that the least-square method with a rank penalty leads to reconstruction error of the same order as for independent datas. Moreover, when the time series has some additional properties such as periodicity, the rate can actually be faster than in the independent case. This is a joint work with Pierre Alquier and Nicolas Marie.
Keywords. Matrix completion ; Time series ; Estimation ; Model selection.

1 Résumé détaillé

La complétion de matrice a longuement été étudiée au cours des dernières années. Une référence incontournable dans ce domaine est le travail de Koltchinskii et al. [1] dans lequel les auteurs ont établi une borne de risque pour l'estimateur de la matrice de rang k et de taille $d \times T$ en $O(k(d+T)log(n)/n)$, pour n entrées bruitées indépendantes. Au cours de cette présentation, nous étendrons les résultats connus au cas de données temporelles partiellement observées et présentant effectivement de la dépendance. Considérons la

matrice aléatoire $\mathbf{M}$ de taille $d \times T$, $d, T \in \mathbb{N}^*$. On suppose que les lignes $\mathbf{M}_{1,.}, \ldots, \mathbf{M}_{d,.}$ de la matrice $\mathbf{M}$ sont des séries temporelles et on dispose de n entrées bruitées $Y_1, \ldots, Y_n$ de la matrice, $n \in \{1, \ldots, d \times T\}$. Ces dernières satisfont le modèle suivant :

$$Y_i = \text{trace}(\mathbf{X}_i^* \mathbf{M}) + \xi_i.$$

Pour exploiter la structure de série temporelle, on peut réécrire la matrice $\mathbf{M}$ sous la forme $\mathbf{M} = \mathbf{\Theta}^0 + \varepsilon = \mathbf{T}^0 \mathbf{\Lambda} + \varepsilon$, où $\mathbf{\Lambda}$ est une matrice de $\mathcal{M}_{\tau,T}(\mathbb{C})$ reflétant la structure de série temporelle potentiellement présente dans les lignes de $\mathbf{M}$, $\mathbf{T}^0$ est une matrice inconnue de $\mathcal{M}_{d,\tau}(\mathbb{R})$, ε est une matrice aléatoire centrée dont les lignes sont indépendantes et $(X_1, \xi_1), \ldots, (X_n, \xi_n)$ sont n variables aléatoires i.i.d à valeurs dans $M_{d,T}(\mathbb{R}) \times \mathbb{R}$ et indépendantes de ε. Enfin, X_i est indépendante de ξ_i pour tout $i \in \{1, \ldots, n\}$. Le paramètre k est le rang de $M_{k,\tau}$ et τ est un paramètre caractérisant une propriété spécifique à la tendance d'une série temporelle comme la τ-périodicité. Notons que si les séries dans $\mathbf{M}_{1,.}, \ldots, \mathbf{M}_{d,.}$ sont τ-périodiques, $\mathbf{\Lambda} = (\mathbf{I}_\tau | \cdots | \mathbf{I}_\tau)$. Enfin, notons que les ξ_i sont les erreurs d'observation liées à l'instrument de mesure, tandis que ε est la composante stochastique de $\mathbf{M}$.

Considérons l'estimateur des moindres carrés suivant:

$$\left\{ \begin{array}{rcl} \widehat{\mathbf{\Theta}}_{k,\tau} & = & \widehat{\mathbf{T}}_{k,\tau} \mathbf{\Lambda} \\ \widehat{\mathbf{T}}_{k,\tau} & \in & \arg\min_{\mathbf{T} \in \mathcal{S}_{k,\tau}} r_n(\mathbf{T}\mathbf{\Lambda}) \end{array} \right. ,$$

où $\mathcal{S}_{k,\tau}$ est un sous-ensemble de

$$\mathcal{M}_{d,k,\tau} := \left\{ \mathbf{UV} \ ; \ (\mathbf{U}, \mathbf{V}) \in \mathcal{M}_{d,k}(\mathbb{R}) \times \mathcal{M}_{k,\tau}(\mathbb{R}) \text{ tel que } \sup_{j,\ell} |\mathbf{U}_{j,\ell}| \leqslant \sqrt{\frac{\mathfrak{m}_0}{k \mathfrak{m}_{\mathbf{\Lambda}}(\tau)}} \right.$$
$$\left. \text{et } \sup_{\ell,t} |\mathbf{V}_{\ell,t}| \leqslant \sqrt{\frac{\mathfrak{m}_0}{k \mathfrak{m}_{\mathbf{\Lambda}}(\tau)}} \right\},$$

et

$$r_n(\mathbf{A}) := \frac{1}{n} \sum_{i=1}^{n} (Y_i - \langle \mathbf{X}_i, \mathbf{A} \rangle_{\mathcal{F}})^2 \ ; \ \forall \mathbf{A} \in \mathcal{M}_{d,T}(\mathbb{R}).$$

Rappelons que le produit scalaire de Frobenius $\langle ., . \rangle_{\mathcal{F}}$ est défini par

$$\langle \mathbf{A}, \mathbf{B} \rangle_{\mathcal{F}} = \text{trace}(\mathbf{A}^* \mathbf{B}) = \sum_{j,t} \mathbf{A}_{j,t} \mathbf{B}_{j,t} . \ ; \ \forall (\mathbf{A}, \mathbf{B}) \in \mathcal{M}_{d,T}(\mathbb{R})^2.$$

1.1 Bornes de risque pour l'estimateur $\widehat{\mathbf{\Theta}}_{k,\tau}$

Sous l'hypothèse que les lignes de ε sont issues d'un processus ϕ-mélangeant, nous établirons un contrôle en $O(k(d+\tau)\log(n)/n)$ du risque quadratique de l'estimateur des moindres carrés $\widehat{\mathbf{\Theta}}_{k,\tau}$:

$$\|\widehat{\mathbf{\Theta}}_{k,\tau} - \mathbf{\Theta}^0\|^2_{\mathcal{F},\Pi} \leqslant 3 \min_{\mathbf{T}\in\mathcal{S}_{k,\tau}} \|(\mathbf{T}-\mathbf{T}^0)\mathbf{\Lambda}\|^2_{\mathcal{F},\Pi} + \mathfrak{c}_1 \left[k(d+\tau)\frac{\log(n)}{n} + \frac{1}{n}\log\left(\frac{4}{\alpha}\right)\right]$$

avec une probabilité supérieure à $1-\alpha$, où $\alpha \in (0,1)$ et $\mathfrak{c}_1 > 0$ une constante indépendante de n, k, d, τ et α. La structure de série temporelle permet d'améliorer le contrôle en $O(k(d+T)\log(n)/n)$ établi dans Koltchinskii et al., [1]. Nous avons également étendu le résultat sur la borne inférieure du théorème 5 de [1] au cas où $\tau \leqslant T$ et où les séries considérées sont τ-périodiques (voir théorème 3.5, [2]). Cette borne est donc optimale dans le cas où les séries sont τ-périodiques, c'est-à-dire lorsque $\mathbf{\Lambda} = (\mathbf{I}_\tau|\ldots|\mathbf{I}_\tau)$.

1.2 Séléction de modèle

Considérons l'estimateur adaptatif suivant: $\widehat{\mathbf{\Theta}} := \widehat{\mathbf{T}}\mathbf{\Lambda}$, où $\widehat{\mathbf{T}} := \widehat{\mathbf{T}}_{\widehat{k}}$. Nous introduisons dans cette section une procédure de séléction du paramètre k. Celui-ci est obtenu en minimisant le risque empirique de l'estimateur assorti d'une pénalité:

$$\widehat{k} \in \arg\min_{k\in\mathcal{K}}\{r_n(\widehat{\mathbf{T}}_k\mathbf{\Lambda}) + \mathrm{pen}(k)\} \text{ avec } \mathcal{K} = \{1,\ldots,k^*\} \subset \mathbb{N}^*,$$

et

$$\mathrm{pen}(k) := 16\mathfrak{c}_{\mathrm{pen}}\frac{\log(n)}{n}k(d+\tau),$$

où $\mathfrak{c}_{\mathrm{pen}}$ désigne une constante calibrée en pratique par heuristique de pente. Nous obtenons alors la borne de risque suivante pour l'estimateur adaptatif:

$$\begin{aligned}\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^0\|^2_{\mathcal{F},\Pi} \quad \leqslant \quad & 4\min_{k\in\mathcal{K}}\left\{3\min_{\mathbf{T}\in\mathcal{S}_k}\|(\mathbf{T}-\mathbf{T}^0)\mathbf{\Lambda}\|^2_{\mathcal{F},\Pi} + \mathfrak{c}_2 k(d+\tau)\frac{\log(n)}{n}\right\} \\ & + \frac{\mathfrak{c}_2}{n}\log\left(\frac{4k^*}{\alpha}\right) + \frac{\mathfrak{c}_3}{n}\end{aligned}$$

avec $\mathfrak{c}_2, \mathfrak{c}_3 > 0$ des constantes indépendantes de n, d, k, τ et α.

1.3 Expériences numériques

Des expériences numériques seront présentées durant l'exposé et viendront illustrer les précedents résultats. En particulier, nous avons réalisé une étude comparative entre l'estimateur $\widehat{\mathbf{\Theta}}_{k,\tau}$ obtenu pour le modèle avec structure de série temporelle et celui obtenu pour le modèle classique, c'est-à-dire sans prise en compte de structure de série temporelle. L'objectif majeur étant bien évidemment de montrer qu'on améliore la qualité de l'estimation lorsque l'on prend en compte la structure de série temporelle dans la matrice $\mathbf{M}$. Le résultat d'une de nos simulations est représenté à la Figure 1. Nous avons

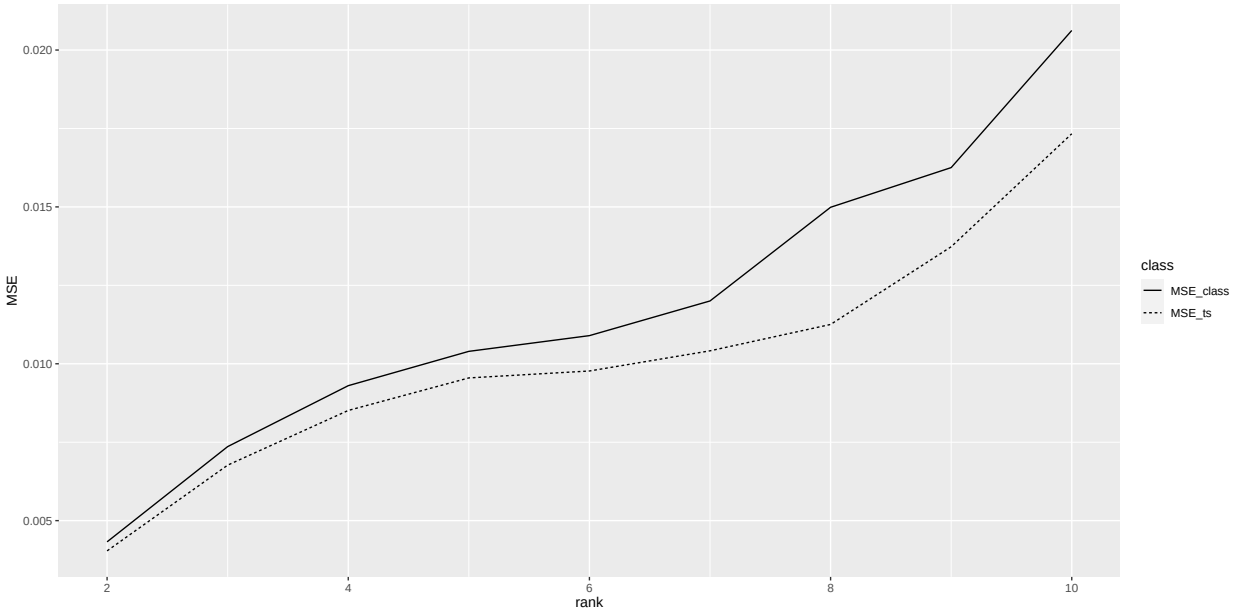


Figure 1: MSE des modèles (séries temporelles (pointillés) v.s. classique (trait plein)) par rapport au rang k.

également illustré nos résultats avec des données réelles provenant du package `FunFEM` (données *Vélib*). Ces données sont mises sous forme de matrice où chaque ligne représente l'occupation au cours du temps d'une station *Vélib* de la ville de Paris. Le terme $M_{i,t}$, $(i,t) \in \{1,\dots,d\} \times \{1,\dots,T\}$ de la matrice $\mathbf{M}$ correspond donc au taux d'occupation (nombre de vélos disponnibles/nombre total de vélos) de la station i au temps t. Les taux d'occupation ont été relevées toutes les heures pendant une semaine. Comme le montre la Figure 2, il s'agit clairement d'un jeu de séries temporelles périodiques, donc parfaitement adapté pour évaluer l'apport de notre méthode par rapport à la méthode classique. Une partie des résultats obtenus est présentée dans la Table 1, que nous commenterons lors de l'exposé.

	15%	30%
Modèle sans struct. périodique	0.0609	0.0315
Modèle avec struct. périodique	0.0436	0.0381

Table 1: MSE en fonction du nombre d'entrées observées (%).

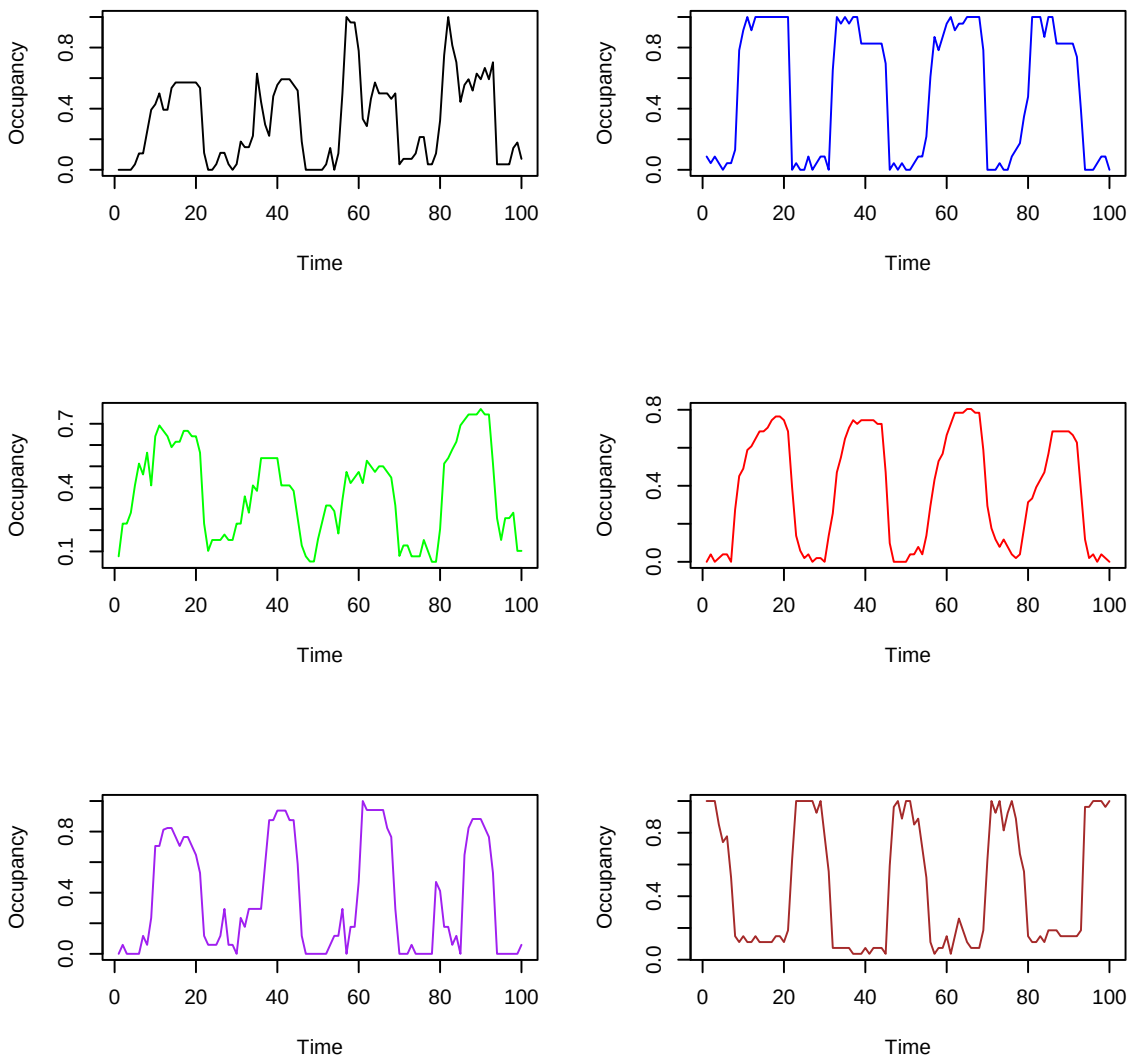


Figure 2: Taux d'occupation de six stations *Vélib* sur une semaine (week-end exclus).

References

[1] Koltchinskii, V., Lounici, K. and Tsybakov, A. B. *Nuclear-Norm Penalization and Optimal Rates for Noisy Low-Rank Matrix Completion.* The Annals of Statistics 39, 5, 2302-2329, 2011.

[2] Alquier, P., Marie, N. and Rosier, A. *Tight Risk Bound for High Dimensional Time Series Completion.* Electronic Journal of Statistics, 2022.

Statistique mathématique – grande dimension

Sparse PLS with group lasso: inside the `dual.spls` package

Louna Alsouki[1,3], François Wahl[1,2], Laurent Duval[2], Clément Marteau[1] & Rami El-Haddad[3]

[1] *Université Claude-Bernard Lyon 1, 43 boulevard du 11 Novembre 1918, 69100 Villeurbanne, France,*
[2] *IFP Energies nouvelles, 1-4 avenue de Bois-Préau, 92852 Rueil-Malmaison, France,*
[3] *Université Saint Joseph de Beyrouth, Mar Roukoz – Dekwaneh, B.P. 1514, Liban,*
louna.al-souki@univ-lyon1.fr, francois.wahl@math.univ-lyon1.fr, laurent.duval@ifpen.fr, marteau@univ-lyon1.fr, rami.haddad@usj.edu.lb

Résumé. Face aux problèmes de régression en grande dimension, deux stratégies de réduction de dimension sont souvent employées: les méthodes de projection, comme la PLS [6], ou les algorithmes de sélection de variables, comme le lasso [5]. Les méthodes de type SPLS (sparse PLS) combinent les deux. Elles visent la précision des prédictions et la localisations des variables importantes. Nous proposons une méthode généralisée nommée dual-SPLS (dual sparse PLS) qui met en œuvre différents types de pénalisations en s'appuiant sur l'algorithme PLS1. La parcimonie de ses résultats montre une bonne localisation tout en maintenant une prédiction adéquate. Grâce à sa flexibilité, elle fournit une stratégie qui permet de traiter différentes types données simultanément. La méthode est implémentée dans le package `dual.spls` disponible dans `R`. Il comprend des fonctions supplémentaires, notamment pour l'échantillonnage en calibration et la génération de données simulés. Nous présentons brièvement le principe de la méthode avec une démonstration sur l'utilisation du package. Nous illustrons la mise en œuvre avec le cas où deux ensembles de prédicteurs sont liés à la même réponse.

Mots-clés. Moindres carrés partiels, parcimonie, régression, norme duale, *group lasso*.

Abstract. The problem of high dimensionality is usually tackled with two dimension reduction techniques: projection methods, like the PLS [6], or variable selection algorithms, like the lasso [5]. Sparse PLS combines both. It aims at accuracy in predictions and in localization of important variables. We propose a generalized method: the dual sparse partial least squares (dual-SPLS) which relies on the PLS1 algorithm modified by different types of penalizations. The sparsity of its results shows good localizations while maintaining accurate prediction performance. Thanks to its flexibility, it provides a strategy that deals with different complementary data. The method is implemented in the package `dual.spls` available in `R`. It includes additional functions, notably, for calibration splits and data simulations. In this paper, we focus on a brief presentation of the method with a demo on the use of the package. We also illustrate the application on the case where two sets of predictors are related to the same response.

Keywords. Partial Least Squares, sparsity, regression, dual norm, group lasso.

1 Introduction

With the growing size of data in many fields, high dimensional situations are increasingly frequent. It has become crucial to resort to dimension reduction techniques. Thus, on the one hand, the latter may be a projection method, like the partial least squares (PLS) [6]. It aims at building a more compact latent space on which the explanatory variables $\mathbf{X}$ are projected, while taking into account the correlation between the p predictors and the n-dimensional reponse vector $\mathbf{y}$. Although this computational approach proved successful in many high dimensional settings thanks to its predictive performance, it often lacks appropriate localization of the important variables. On the other hand, variable selection algorithms also deal with high dimensional data aiming for a less complex model easier for interpretations. For instance, we recall the lasso algorithm [5] that adopts a shrinkage scheme inducing sparsity of the coefficients. Consequently, the procedure selects the most important features. However, the lasso is sensitive to the available data and does not always result in interpretable coefficients. Sparse partial least squares (SPLS) combines both approaches by adding to the PLS framework a selection step using a type of penalty. We propose a generalization of the classical PLS1 algorithm (dual-SPLS). It relies on the dual norm of a chosen penalty norm [1] where shrinkage is applied adaptively. It aims at 1) combining heterogeneous and high-dimensional data sources 2) providing accurate predictions 3) extracting pertinent knowledge for better localization.
The algorithms are implemented in `R` in a package called `dual.spls`. It includes mainly four dual-SPLS types of strategies inspired by the lasso, the group lasso, the least squares and the ridge methodologies. It also provides a calibration and validation splitting method. The latter is an improvement over the Kennard and Stone algorithm [3] and considers simultaneously both the predictors and the response.
In the following, we give an overview on the dual-SPLS procedure, citing breifly the four types of norms that we used. Then we describe the package functionalities with a demo on the group lasso dual-SPLS norm. Lastly, we show some results on simulated data.

2 Dual Sparse Partial Least Squares

Definition 2.1 *Let $\Omega(\cdot)$ be a norm on $\mathbb{R}^p$. The associated dual norm, denoted $\Omega^*(\cdot)$, is defined as, for any $\boldsymbol{z} \in \mathbb{R}^p$,*

$$\Omega^*(\boldsymbol{z}) = \max_{\boldsymbol{w}} (\boldsymbol{z}^T \boldsymbol{w}) \quad s.t. \quad \Omega(\boldsymbol{w}) = 1. \tag{1}$$

In particular, the dual of the classical ℓ_2-norm is the euclidien norm itself. Recalling the optimization problem of the PLS for the first component

$$\max_{\mathbf{w}} (\mathbf{y}^T \mathbf{X} \mathbf{w}) \quad \text{s.t.} \quad \|\mathbf{w}\|_2 = 1. \tag{2}$$

we notice that for $\mathbf{z} = \mathbf{X}^t\mathbf{y}$, the ℓ_2 dual norm amounts to finding the first weight vector of the PLS1 algorithm. Thereby, the dual-SPLS evaluates different types of norm, notably adding adaptive penalization. We study four types of norms, inspired by classical algorithms:

1. **'lasso' norm** [5] with `d.spls.lasso()`: it is an intuitive norm that combines the ℓ_1 and ℓ_2 norms:
$$\Omega(\mathbf{w}) = \lambda\|\mathbf{w}\|_1 + \|\mathbf{w}\|_2 \tag{3}$$

2. **'group lasso' norm** [4] with `d.spls.GL()`: it deals with combined complementary data situations. Three norms are sugessted:

   (a) **GLA** is the simplest GL norm, it applies the lasso norm to each group separately:
   $$\Omega(\mathbf{w}) = \|\mathbf{w}\|_2 + \sum_{g\in 1,\dots G} \lambda_g\|\mathbf{w}_g\|_1 \tag{4}$$

   (b) **GLB** decomposes the ℓ_2-norm and adds a constraint reflecting the importance assigned to each group:
   $$\Omega(\mathbf{w}) = \sum_{g\in 1,\dots G} \alpha_g\|\mathbf{w}_g\|_2 + \sum_{g\in 1,\dots G} \lambda_g\|\mathbf{w}_g\|_1 \tag{5}$$
   where $\sum_{g\in 1,\dots G}\alpha_g = 1$, and $\sum_{g\in 1,\dots G}\Omega(\mathbf{w}_g) = 1$.

   (c) **GLC** applies the lasso norm to each group separately but constraints the total set
   $$\Omega_g(\mathbf{w}) = \|\mathbf{w}_g||_2 + \lambda_g||\mathbf{w}_g\|_1 \tag{6}$$
   where $\Omega(\mathbf{w}) = \sum_{g\in 1,\dots G}\alpha_g\Omega_g(\mathbf{w}) = 1$ and $\sum_{g\in 1,\dots G}\alpha_g = 1$. Note that:
   - g represents a group of $p(g)$ index extracted from $\{1,\dots,p\}$;
   - G represents the number of groups;
   - $\mathbf{w}_g$ represents the values of index g in the loading vector $\mathbf{w}$.

3. **'least squares' norm** with `d.spls.LS()`: it examines a case where $\mathbf{X}$ is not singular and $\mathbf{N}_1$ is a p squared matrix:
$$\Omega(\mathbf{w}) = \lambda\|\mathbf{N_1}\mathbf{w}\|_1 + \|\mathbf{X}\mathbf{w}\|_2. \tag{7}$$
   The solurion for $\lambda = 0$ and $\mathbf{N}_1$ the identity matrix is the classical least squares solution.

4. **'ridge' norm** [2] with `d.spls.ridge()`: it deals with cases where $\mathbf{X}$ is singular and uses a ridge penalization scheme:
$$\Omega(\mathbf{w}) = \lambda_1\|\mathbf{w}\|_1 + \lambda_2\|\mathbf{X}\mathbf{w}\|_2 + \|\mathbf{w}\|_2. \tag{8}$$

In dual-SPLS, the goal is to find the weight vector $\mathbf{w}$ that corresponds to the dual norm of one of the four norms mentioned above. Each alternative has a closed form solution that can be expressed with the *soft thresholding* operator: $f_\nu(\mathbf{u}) = \text{sign}(\mathbf{u})(|\mathbf{u}| - \nu)_+$, where $\text{sign}(\mathbf{u})$ is the vector of signs of $\mathbf{u}$ and ν is the sparsity parameter. A common issue in shrinkage methods is the choice of the appropriate shrinking parameter. Thereby, we conceived an adaptive algorithm that computes ν according to the number of variables that we would like to keep in the active set at each iteration.

3 `dual.spls` package: an example of the `d.spls.GL` norm A

3.1 Data sets

The `dual.spls` package includes a function that simulates `G` sub matrices. Each is computed using a mixture of gaussian distributions of specified standard deviation repectively denoted as `nondes` and `sigmaondes`. It also simulates a response vector $\mathbf{y}$ as a linear combination of the predictors to which a `sigmay` uncertainty is added. In the following example, we propose a simulation of two mixtures of 50 and 100 predictors respectively and 100 observations. The first sub matrix is computed using 20 Gaussians of 0.05 standard deviation. The second uses 30 gaussian distributions of a standard deviation of 0.02. The uncertainty of $\mathbf{y}$ is set by default to 0.5.

```
R> n <- 100
R> p <- c(50,100)
R> data <- d.spls.simulate(n=n,p=p,nondes=c(50,100),
R>                         sigmaondes=c(0.05,0.02))
```

The returned values are a concatunated matrix $\mathbf{X}$ with 100 observations and 150 variables and a response vector $\mathbf{y}$.

```
R> X <- data$X
R> y <- data$y
```

3.2 Application of dual-SPLS group lasso norm A

Similarly to all the `dual.spls` functions, the number of component and the proportion of zeros that we would like to have in the coefficients must be specified. They are denoted `ncp` and `ppnu`. Here we apply dual-SPLS for 20 components. We also requested that 80% of the coefficients should be null in the first set of predictors and 85 % in the second set. Moreover, the `d.spls.GL()` requires a vector `indG` that indexes each observation by the group it belongs to.

```
R> indG <- c(rep(1,p[1]),rep(2,p[2]))
R> mod.splsA <-d.spls.GL(X=X,y=y,ncp=20,ppnu=c(0.8,0.85),
                  indG=indG,norm="A",verbose=FALSE)
```

The figure 1 represents the results after splitting the observations into two groups: 80% for calibration and the rest for validation using the implemented function `d.spls.calval()`. The results are compared with the standars PLS1 regression algorithm using the function `d.spls.pls()`. Visualization of the results can be done with the function `d.spls.plot()` that plots the coefficients. Moreover, RMSE and MAE errors and Rsquared are computed using the function `d.spls.metric()`.

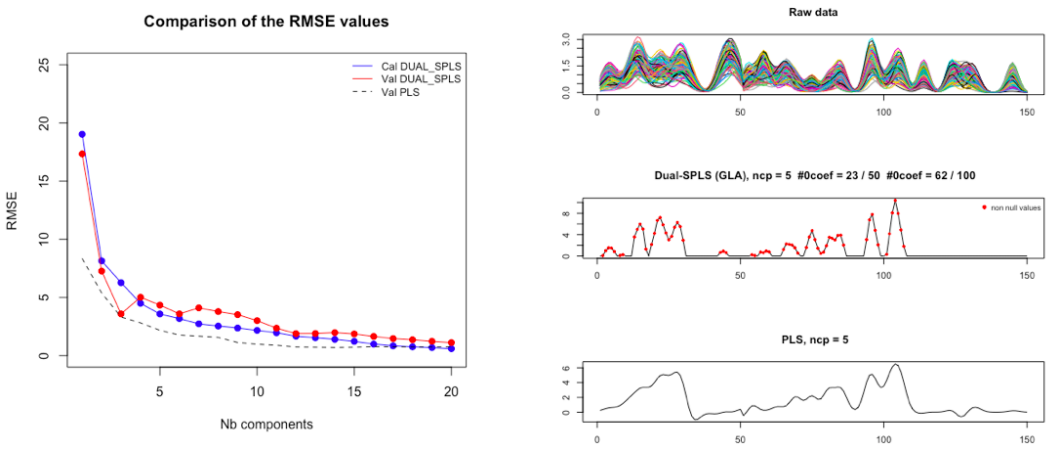


Figure 1: Dual-SPLS group lasso norm GLA results: prediction errors according to the number of components (left), raw data and coefficients localization for dual-SPLS and PLS (right).

According to Figure 1 (left), we notice that the predictions of dual-SPLS are slightly less accurate than with PLS regression. However, with an increasing number of components, the curves become more and more closer. We select 5 components to compare the coefficients localization as the RMSE curves exhibit a plateau above this value for the two regressions. On the right pane, the first figure shows the raw data. In the second figure, the red points represent only non null coefficients of dual-SPLS GLA regression. The line is here to help the visualization. In the third figure, we plot the coefficients obtained by the classical PLS1 algorithm; all of them are non zero. We see that dual-SPLS selects less than half of the total variables. We also notice that each variable is representative compared to the raw data plot. On the contrary, the PLS regression coefficients are hard to interpret and no pertinent information about the variables can be extracted.

4 Conclusions

The `R` package `dual.spls` regroups a set of functions used for the application of the dual-SPLS for four cases of norm. It also provides a function that simulates data sets and another that splits the data into calibration and validation. Additional functions are also present; for instance, `d.spls.plot()` and `d.spls.metric()` helps in evaluating results. The dual-SPLS algorithms (as in the example of the group lasso norm) tends to result in accurate predictions compared to the PLS regression. In addition, its performance is acheived with much less selected variables. Indeed, its results show sparsity with accurate localization.

References

[1] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Optimization with sparsity-inducing penalties. 4(1):1–106, 2012.

[2] A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[3] R. W. Kennard and L. A. Stone. Computer aided design of experiments. *Technometrics*, 11(1):137–148, 1969.

[4] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani. A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245, 2013.

[5] R. Tibshirani. Regression shrinkage and selection via the lasso. 58(1):267–288, 1996.

[6] H. Wold. Path models with latent variables: The NIPALS approach. In *Quantitative Sociology. International Perspectives on Mathematical and Statistical Modeling*, pages 307–357. Elsevier, 1975.

5 Acknowledgement

This work was performed within the framework of the LABEX MILYON (ANR-10-LABX-0070) of Université de Lyon, within the program "Investissements d'Avenir" (ANR-11-IDEX-0007) operated by the French National Research Agency (ANR).

Sélection de modèles linéaires emboîtés par règle d'arrêt prématuré

Samy Clementz [1] & Alain Celisse [1] & Sylvain Arlot [2]

[1] *Laboratoire SAMM EA 4543, Université Paris 1 Panthéon-Sorbonne, 90 rue de Tolbiac, 75 634 Paris Cedex 13, France, samy.clementz@univ-paris1.fr*
[2] *Laboratoire LMO Université Paris-Saclay, CNRS, Laboratoire de mathématiques d'Orsay, 91405, Orsay, France*

Résumé. Dans de nombreux problèmes de sélection de modèle, les estimateurs candidats sont construits de manière itérative. Il s'agit par exemple du cadre de l'Analyse en Composantes Principales (ACP), du débruitage par projection, ou encore de la sélection de variables ordonnées. Classiquement, l'étape de sélection de modèle ne débute que lorsque s'achève le processus complet de construction des estimateurs, coûteux en temps et en ressources de calcul. Pour contourner cette difficulté, nous analysons une méthode de sélection alternative, connue dans la littérature des problèmes inverses sous le nom de principe de discrépance. Cette méthode arrête prématurément le processus itératif de construction des estimateurs, et est donc adaptée aux situations où temps et ressources de calculs sont limités.

Mots-clés. Règles d'arrêt prématuré, principe de discrépance, sélection de modèle, analyse en composantes principales ...

Remerciements. Projet soutenu par l'attribution d'une allocation de recherche Région Ile-de-France

Abstract. In many model selection problems, candidate estimators are constructed iteratively. This holds for example with Principal Component Analysis (PCA), in the denoising by projection framework, or in the ordered-variable selection framework. Classically, model selection only starts after the construction of all estimators has been completed, which is costly in terms of time and computing resources. For addressing this issue, we analyze an alternative selection method, known in the inverse problem literature as the discrepancy principle. This method stops early the iterative construction of the successive estimators. It is therefore suited to situations where time and computational resources are limited.

Keywords. early stopping, discrepancy principle, model selection, principal component analysis ...

Acknowledgements This work was supported by grants from Région Ile-de-France.

1 Introduction

Supposons que l'on observe un signal bruité :

$$x = f^* + \varepsilon, \tag{1}$$

où $x \in \mathbb{R}^p$ est un signal, $f^* \in \mathbb{R}^p$ est le vecteur de moyenne du signal, déterministe ou aléatoire, et ε est un vecteur gaussien centré de matrice de covariance $\sigma^2 I_p$, indépendant de f^*. Le but est de reconstituer f^* à partir de x.

Si nous connaissons *a priori* un sous-espace vectoriel S de $\mathbb{R}^p$ auquel f^* appartient, nous pouvons considérer l'estimateur du maximum de vraisemblance contraint à S, qui est $\hat{f}_S = P_S x$, où $P_S : \mathbb{R}^p \to \mathbb{R}^p$ est la projection orthogonale sur S. Si nous ne savons pas *a priori* dans quel S se trouve le signal f^*, une méthode classique consiste à :

1. Se donner une collection de modèles $\{S_{\leq k}, k = 1, \dots, p\}$ emboîtés, c'est-à-dire :

$$\begin{cases} S_{\leq k} \subset S_{\leq k+1}, & k = 1, \dots, p-1 \\ \dim(S_{\leq k}) = k, & k = 1, \dots, p. \end{cases} \tag{2}$$

   Des exemples où les conditions d'emboîtement (2) sont vérifiées sont: (i) l'ACP (où les modèles sont préalablement appris sur des données d'entraînement indépendantes de x), (ii) le débruitage où l'on se donne une base orthonormée $(g_i)_{1 \leq i \leq p}$ de $\mathbb{R}^p$ (comme une base de Fourier discrète) puis on pose $S_{\leq k} = Vect(g_i, i \leq k)$. En régression linéaire où $f^* = X\beta^*$ et la matrice X est supposée de rang plein, la sélection de variables ordonnées constitue un autre exemple classique puisqu'elle donne lieu à $S_{\leq k} = Vect(X_i, i \leq k)$, où X_i est le $i$ème vecteur-colonne de X.

2. Construire itérativement pour chaque k les estimateurs $\hat{f}_{\leq k} = P_{\leq k} x$, où l'on note $P_{\leq k}$ la projection orthogonale sur $S_{\leq k}$;

3. Estimer f^* par le *meilleur* estimateur dans la collection $\{\hat{f}_{\leq k}, k = 1, \dots, p\}$ pour un certain critère à minimiser.

Tout comme l'étape 2, la recherche du *meilleur* estimateur à l'étape 3 est aussi coûteuse en ressources de calculs. En effet, elle peut consister à minimiser un risque empirique pénalisé de tous les estimateurs construits en étape 2. C'est typiquement le cas des critères AIC, BIC et plus généralement dans les procédures étudiées par Birgé et Massart (2001). La lourde charge de calcul se retrouve aussi en validation croisée (voir Arlot et Celisse (2010)).

Dans ce travail, nous introduisons une nouvelle règle de choix de modèle basé sur le principe de discrépance, permettant d'éviter le calcul exhaustif de tous les estimateurs à l'étape 2 en interrompant le calcul itératif de façon prématurée. En particulier, la méthode que nous proposons ne requiert que le calcul et le parcours d'un nombre limité

d'estimateurs en compétition. La performance statistique de cette règle d'arrêt est établie par le biais d'une inégalité oracle où la performance du meilleur estimateur choisi est comparée à celle de tous les estimateurs candidats, y compris ceux qui ne sont pas calculés car jugés trop mauvais. Enfin, nous illustrons la règle d'arrêt sur un exemple simple.

2 Critère d'arrêt prématuré

L'idée de l'approche proposée pour choisir le *meilleur* estimateur parmi $\hat{f}_{\leq 1}, \hat{f}_{\leq 2}, \ldots$ est d'interrompre le processus itératif de construction de ces estimateurs la première fois qu'un bon candidat est détecté. Plus précisément, ceci se traduit par l'arrêt de ce processus au premier instant pour lequel le risque empirique associé à la séquence d'estimateurs passe en dessous d'un certain seuil $\kappa > 0$:

$$k_{DP} = \min\{k \geq 1 : \|x - \hat{f}_{\leq k}\|^2 \leq \kappa\}. \tag{3}$$

Cette méthode de sélection a été introduite par Morozov (1966) dans le cadre des problèmes inverses pour le choix du paramètre de régularisation. Plus récemment, cette méthode a été analysée par Blanchard, Hoffman et Reiß(2018) pour les problèmes inverses en statistiques, et par Celisse et Wahl (2021) en régression non-paramétrique à noyau reproduisant pour le choix du nombre d'itérations d'un algorithme d'apprentissage de type descente de gradient.

Afin de simplifier la présentation du résultat, nous introduisons, pour tout $t \in [1, p]$, le pseudo-projecteur interpolant et l'estimateur associé :

$$P_{\leq t} := P_{\leq \lfloor t \rfloor} + \sqrt{\{t\}}(P_{\leq \lfloor t \rfloor + 1} - P_{\leq \lfloor t \rfloor}),$$

$$\hat{f}_{\leq t} = P_{\leq t} x.$$

Nous introduisons, également pour simplifier la présentation, la règle d'arrêt continue :

$$t_{DP} = \min\{t \geq 1 : \|x - \hat{f}_{\leq t}\|^2 \leq \kappa\}. \tag{4}$$

Nous montrons que, pour le choix $\kappa = \sigma^2 p$, **l'estimateur sélectionné par le principe de discrépance $\hat{f}_{\leq t_{DP}}$ vérifie une inégalité de type oracle impliquant *tous* les estimateurs de la collection, y compris sur ceux qui n'ont pas été construits en pratique car jugés trop mauvais**.

Théorème 1 *L'estimateur sélectionné par le principe de discrépance $\hat{f}_{\leq t_{DP}}$ vérifie, pour tout $u > 0$, avec probabilité $\geq 1 - 5\exp(-u)$:*

$$\|\hat{f}_{t_{DP}} - f^*\|^2 \leq C \left(\min_{1 \leq t \leq p} \mathbb{E}[\|\hat{f}_t - f^*\|^2] + \sigma^2(\sqrt{pu} + u) \right).$$

Le terme $\sigma^2\sqrt{p}$ est de taille raisonnable car le terme de variance contenu dans le risque est linéaire en σ^2, de maximum $\sigma^2 p$.

3 Exemple : choix du nombre de composantes de l'ACP

Dans le cas particulier où le vecteur f^* est aléatoire, supposons qu'on dispose préalablement d'un ensemble d'entraînement de n exemples de signaux bruités:

$$x_i = f_i^* + \varepsilon_i, \quad i = 1, \dots, n, \tag{5}$$

où seuls les vecteurs aléatoires $(x_i)_{1 \leq i \leq n}$ sont observés. Les vecteurs aléatoires ε_i sont des vecteurs gaussiens centrés de matrice de covariance $\sigma^2 I_p$, indépendants de f_i^*.

Supposons que la loi de f^* est portée par un sous-espace vectoriel de dimension k petite par rapport à p. Dans ce cas, il peut être intéressant de projeter l'observation x sur un sous-espace vectoriel (bien choisi) afin de conserver l'information disponible tout en réduisant la dimension. Le lemme suivant montre que les projecteurs de choix sont ceux obtenus par l'ACP de la matrice de covariance de x.

Lemme 1 *Le minimum de l'erreur quadratique moyenne*

$$L : P \mapsto \mathbb{E}[\|f^* - Px\|^2]$$

sur tous les projecteurs orthogonaux P de rang k est atteint en $P_{\leq k}$, qui est le projecteur orthogonal sur le sous-espace vectoriel engendré par les k premières directions propres de $\Sigma_{f^*} = \mathrm{Cov}(f^*)$.

En pratique, ces projecteurs peuvent être estimés par $\hat{P}_{\leq k}$, construits de manière itérative, en calculant les k directions propres de la matrice de covariance empirique $\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^n x_i x_i^T$ correspondant à ses k plus grandes valeurs propres. On définit alors l'excès de risque de l'ACP :

$$\mathcal{E}_{k,n} = L(\hat{P}_{\leq k}) - L(P_{\leq k})$$

Des bornes sur cet excès de risque, en grande probabilité et en espérance, sont données par Reiss et Wahl (2020) et Milbradt et Wahl (2020), en fonction de la décroissance des valeurs propres de la matrice de covariance de x. Le corollaire suivant montre en particulier que l'excès de risque $\mathcal{E}_{k,n}$ est exactement le terme d'erreur permettant de faire apparaître le biais alternatif $\mathbb{E}[\|f^* - P_{\leq k} f^*\|^2] = \sum_{j=k+1}^p \lambda_j(\Sigma_{f^*})$, qui est le biais obtenu via les projecteurs que l'on cherche à estimer.

Corollaire 1 *Avec probabilité $\geq 1 - 6\exp(-u)$:*

$$\|\hat{P}_{\leq t_{DP}} x - f^*\|^2 \leq C \left(\min_{1 \leq k \leq p} \left\{ \sum_{j=k+1}^p \lambda_j(\Sigma_{f^*}) + k\sigma^2 + \mathcal{E}_{k,n} \right\} + \sigma^2(\sqrt{pu} + u) \right.$$
$$\left. + \sqrt{\mathrm{tr}(\Sigma_{f^*})u} + \lambda_1(\Sigma_{f^*})u \right).$$

4 Structure de la preuve

La preuve se base sur la décomposition biais-variance-variance approchée de l'espérance du risque et du risque empirique. On pose pour tout $t \in [1, p]$:

$$B^2(t) = \|(I - P_{\leq t})f^*\|^2, \quad V(t) = t\sigma^2, \quad v(t) = (t + \sqrt{\{t\}})\sigma^2,$$

alors :

$$\mathbb{E}[\|f^* - \hat{f}_{\leq t}\|^2] = B^2(t) + V(t),$$

$$\mathbb{E}[\|x - \hat{f}_{\leq t}\|^2] = \sigma^2 + B^2(t) + V(t) - 2v(t).$$

L'idée est de comparer la performance de la règle d'arrêt empirique (3) avec celle de la règle d'arrêt de population :

$$\begin{aligned} t^* &= \min\{t \in [1, p] : \mathbb{E}[\|x - \hat{f}_{\leq t}\|^2] \leq \sigma^2\} \\ &= \min\{t \in [1, p] : B^2(t) \leq 2v(t) - V(t)\} \end{aligned}$$

qui est donc, si on fait l'approximation $v(t) = V(t)$, l'instant où le biais coïncide avec la variance. Cet instant t^* est optimal à constante près, puisque on peut montrer :

Lemme 2

$$\mathbb{E}[\|f^* - \hat{f}_{\leq t^*}\|^2] \leq 2 \min_{1 \leq t \leq p} \mathbb{E}[\|f^* - \hat{f}_{\leq t}\|^2].$$

Nous commençons par majorer le risque en t_{DP}. En posant $\hat{v}(t) = \langle \varepsilon, P_{\leq t}\varepsilon \rangle$, on a :

$$\|f^* - \hat{f}_{t_{DP}}\|^2 \leq 2\left(B^2(t_{DP}) + \hat{v}(t_{DP})\right).$$

Tout le travail consiste à comparer le biais en t_{DP} et biais en t^*, puis $\hat{v}$ en t_{DP} et la variance approchée v en t^*. C'est l'objet de la proposition suivante, dont la preuve utilise l'inégalité d'Hanson-Wright :

Proposition 1 *Pour tout $u > 0$:*

$$\mathbb{P}\left(B^2(t_{DP}) > 2B^2(t^*) + u\right) \leq 2\exp\left(-c\min\left(\frac{u^2}{\sigma^4 p}, \frac{u}{\sigma^2}\right)\right).$$

$$\mathbb{P}(\hat{v}(t_{DP}) > v(t^*) + 2u) \leq 3\exp\left(-c\min\left(\frac{u^2}{\sigma^4 p}, \frac{u}{\sigma^2}\right)\right)$$

Donc avec probabilité $\geq 1 - 5\exp(-u)$:

$$\|P_{\leq t_{DP}}x - f\|^2 \leq C(\min_{1 \leq t \leq p}\{B^2(t) + v(t)\} + \sigma^2(\sqrt{pu} + u)).$$

L'inégalité $v(t) \leq 2V(t)$, pour $t \geq 1$, puis le Lemme 2 permettent de montrer le Théorème 1.

5 Illustration

On illustre la procédure d'arrêt en considérant un signal f^* échantillonné selon $\mathcal{N}(0, \Sigma)$, avec $\Sigma = diag\left[(20e^{-0.02k})_{0\leq k<500}\right]$. On observe $x = f^* + \varepsilon$, avec $\varepsilon \sim \mathcal{N}(0, I_{500})$. On reconstitue le signal grâce aux projections $(\hat{f}_{\leq k})_{\leq k}$ sur l'espace vectoriel engendré par les k premiers vecteurs de la base canonique de $\mathbb{R}^{500}$.

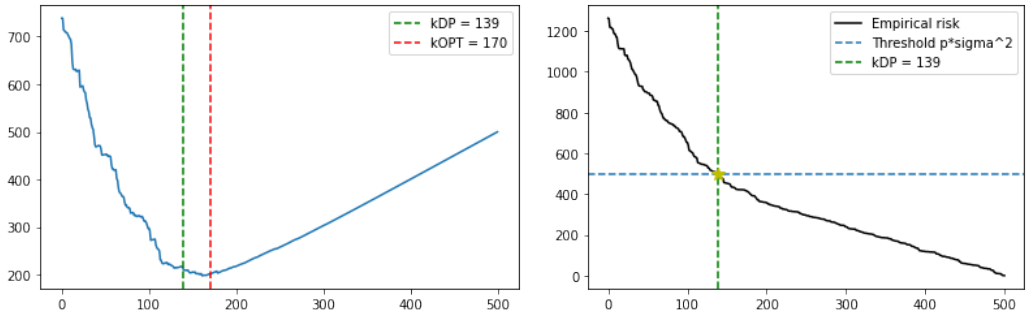


Figure 1: Illustration de la règle d'arrêt pour une réalisation de x. À gauche : le risque moyen, minimisé en l'instant idéal k_{opt}. À droite : le risque empirique. On retient k_{DP}, le premier instant pour lequel le risque empirique passe sous le seuil $500\sigma^2$.

Bibliographie

Morozov, V. A. (1966). On the solution of functional equations by the method of regularization. *Doklady Akademii Nauk* (Vol. 167, No. 3, pp. 510-512). Russian Academy of Sciences.

Celisse, A., et Wahl, M. (2021). Analyzing the discrepancy principle for kernelized spectral filter learning algorithms. *Journal of Machine Learning Research*, 22(76), 1-59.

Birgé, L., et Massart, P. (2001). Gaussian model selection. *Journal of the European Mathematical Society*, 3(3), 203-268.

Arlot, S., et Celisse, A. (2010). A survey of cross-validation procedures for model selection. Statistics surveys, 4, 40-79.

Régression non paramétrique par projection sur des bases à support non compact

Florian Dussap [1]

[1] *Université Paris Cité, CNRS, MAP5 UMR 8145, 75006 Paris*
florian.dussap@u-paris.fr

Résumé. On étudie le problème d'estimation non paramétrique d'une fonction de régression avec un design aléatoire sur $\mathbb{R}^p$ pour $p \geqslant 2$. On utilise pour cela un estimateur par projection calculé avec un critère de moindres carrés. Notre contribution est de considérer des domaines d'estimation non compacts et d'étudier théoriquement le risque de l'estimateur pondéré par la loi du design. On propose une procédure de sélection de modèle dans laquelle la collection de modèles est aléatoire et prend en compte l'écart entre la norme empirique et la norme associée à la loi du design. On démontre que l'estimateur résultant optimise automatiquement le compromis biais-variance pour les deux normes.

Mots-clés. estimation non paramétrique, régression non paramétrique, base orthonormée, sélection de modèles .

Abstract. We study the nonparametric regression estimation problem with a random design in $\mathbb{R}^p$ with $p \geqslant 2$. We do so by using a projection estimator obtained by least squares minimization. Our contribution is to consider non-compact estimation domains in $\mathbb{R}^p$ on which we recover the function, and to provide a theoretical study of the risk of the estimator relative to a norm weighted by the distribution of the design. We propose a model selection procedure in which the model collection is random and takes into account the discrepancy between the empirical norm and the norm associated with the distribution of design. We prove that the resulting estimator automatically optimizes the bias-variance trade-off in both norms.

Keywords. nonparametric estimation, nonparametric regression, orthonormal basis, model selection.

1 Modèle statistique

On considère le modèle de régression non paramétrique :

$$Y_i = b(\boldsymbol{X}_i) + \varepsilon_i, \quad i = 1, \dots, n$$

où les variables $\boldsymbol{X}_i \in \mathbb{R}^p$ sont indépendantes mais pas nécessairement identiquement distribuées, les variables de bruit $\varepsilon_i \in \mathbb{R}$ sont i.i.d. centrées avec une variance finie σ^2 et

indépendantes des $\boldsymbol{X}_i$, et $b\colon \mathbb{R}^p \to \mathbb{R}$ est la fonction de régression. Notre objectif est de reconstruire la fonction b sur un domaine $A \subset \mathbb{R}^p$ à partir des observations $(\boldsymbol{X}_i, Y_i)$.

On travaille dans le cadre suivant. On suppose que la variance de bruit est connue. On note μ_i la loi de $\boldsymbol{X}_i$ qu'on ne suppose pas connue. Cependant, on fixe ν une mesure de référence sur A et on suppose que la mesure $\mu \coloneqq \frac{1}{n}\sum_{i=1}^n \mu_i$ admet une densité bornée par rapport à ν, de sorte que $\mathrm{L}^2(A,\mu) \subset \mathrm{L}^2(A,\nu)$. En particulier, cette hypothèse implique que $\mathrm{supp}(\mu) \subset A$. Enfin, on suppose que A s'écrit comme un produit cartésien $A_1 \times \cdots \times A_p$ et que ν s'écrit comme une mesure produit $\nu_1 \otimes \cdots \otimes \nu_p$ avec ν_i une mesure sur A_i.

Notre but est d'estimer la fonction b sur le domaine A et de contrôler l'erreur moyenne pour la norme de $\mathrm{L}^2(A,\mu)$, c'est-à-dire l'erreur quadratique intégrée pondérée par la loi du design. L'erreur pour cette norme peut s'interpréter comme une erreur de prédiction :

$$\forall\, \hat{b} \text{ estimateur}, \quad \|b-\hat{b}\|_\mu^2 = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\boldsymbol{X}_i' \sim \mu_i}\Big[\Big(b(\boldsymbol{X}_i') - \hat{b}(\boldsymbol{X}_i')\Big)^2 \,\Big|\, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\Big],$$

à savoir l'erreur quadratique pour une nouvelle observation tirée uniformément parmi les lois μ_i.

2 Construction de l'estimateur

Pour chaque $i \in \{1,\ldots,p\}$, on choisit $(\varphi_k^i)_{k\in\mathbb{N}}$ une base orthonormée de $\mathrm{L}^2(A_i,\nu_i)$ constituée de fonctions bornées et on construit une base orthonormée de $\mathrm{L}^2(A,\nu)$ par tensorisation :

$$\forall \boldsymbol{x} \in A,\ \varphi_{\boldsymbol{k}}(\boldsymbol{x}) \coloneqq \left(\varphi_{k_1}^1 \otimes \cdots \otimes \varphi_{k_p}^p\right)(\boldsymbol{x}) \coloneqq \varphi_{k_1}^1(x_1) \times \cdots \times \varphi_{k_p}^p(x_p).$$

Le principe de l'estimation par projection est de décomposer la fonction b sur la base $(\varphi_{\boldsymbol{k}})_{\boldsymbol{k}\in\mathbb{N}^p}$ et d'estimer ses coefficients. Plus précisément, pour $\boldsymbol{m} \in \mathbb{N}_+^p$ notons $S_{\boldsymbol{m}}$ l'espace vectoriel engendré par les fonctions $\varphi_{\boldsymbol{k}}$ avec $k_i < m_i$. On estime b par l'estimateur des moindres carrés sur $S_{\boldsymbol{m}}$:

$$\hat{b}_{\boldsymbol{m}} \coloneqq \operatorname*{arg\,min}_{t\in S_{\boldsymbol{m}}} \frac{1}{n}\sum_{i=1}^n \Big(Y_i - t(\boldsymbol{X}_i)\Big)^2.$$

Cet estimateur s'exprime à l'aide d'hypermatrices (tableaux de nombres multidimensionnels) :

$$\hat{b}_{\boldsymbol{m}} = \sum_{\forall i,\, k_i < m_i} \hat{a}_{\boldsymbol{k}}^{(\boldsymbol{m})} \varphi_{\boldsymbol{k}}, \quad \hat{\mathbf{a}}_{\boldsymbol{k}}^{(\boldsymbol{m})} = \frac{1}{n}\widehat{\mathbf{G}}_{\boldsymbol{m}}^{-1} \times_d \widehat{\boldsymbol{\Phi}}_{\boldsymbol{m}}^* \times_1 \mathbf{Y},$$

où $\times_i$ désigne le produit contracté d'ordre i d'hypermatrices[1] et où $\mathbf{Y} \coloneqq (Y_1,\ldots,Y_n)$, $\widehat{\boldsymbol{\Phi}}_{\boldsymbol{m}}^* \coloneqq [\varphi_{\boldsymbol{k}}(\boldsymbol{X}_i)]_{\boldsymbol{k},i} \in \mathbb{R}^{\boldsymbol{m}\times n}$ et où $\widehat{\mathbf{G}}_{\boldsymbol{m}} \in \mathbb{R}^{\boldsymbol{m}\times\boldsymbol{m}}$ est l'hypermatrice de Gram pour le produit

1. $\times_i$ est l'opération : $[\mathbf{A} \times_i \mathbf{B}]_{\boldsymbol{j},\boldsymbol{\ell}} \coloneqq \sum_{\boldsymbol{k}=(k_1,\ldots,k_i)} A_{\boldsymbol{j},\boldsymbol{k}} \times B_{\boldsymbol{k},\boldsymbol{\ell}}$.

scalaire empirique :

$$\widehat{\mathbf{G}}_{\boldsymbol{m}} \coloneqq \Big[\langle \varphi_{\boldsymbol{j}}, \varphi_{\boldsymbol{k}} \rangle_n\Big]_{\boldsymbol{j},\boldsymbol{k}}, \quad \langle t, s \rangle_n \coloneqq \frac{1}{n}\sum_{i=1}^{n} t(\boldsymbol{X}_i)s(\boldsymbol{X}_i).$$

Dans la suite, on note $\mathbf{G}_{\boldsymbol{m}}$ l'espérance de $\widehat{\mathbf{G}}_{\boldsymbol{m}}$, c'est-à-dire l'hypermatrice de Gram pour le produit scalaire pondéré par la loi du design :

$$\mathbf{G}_{\boldsymbol{m}} \coloneqq \Big[\langle \varphi_{\boldsymbol{j}}, \varphi_{\boldsymbol{k}} \rangle_\mu\Big]_{\boldsymbol{j},\boldsymbol{k}}, \quad \langle t, s \rangle_\mu \coloneqq \int_A t(\boldsymbol{x})s(\boldsymbol{x})\,\mathrm{d}\mu(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\Big[t(\boldsymbol{X}_i)s(\boldsymbol{X}_i)\Big].$$

3 Étude du risque pour un modèle fixé

Commençons par rappeler la décomposition biais-variance en norme empirique de l'estimateur dans le cas d'un design fixe (i.e. conditionnellement au design) :

$$\mathbb{E}\Big[\|b - \hat{b}_{\boldsymbol{m}}\|_n^2 \,\Big|\, \boldsymbol{X}_1, \dots, \boldsymbol{X}_n\Big] = \inf_{t \in S_{\boldsymbol{m}}} \|b - t\|_n^2 + \sigma^2 \frac{D_{\boldsymbol{m}}}{n},$$

où $D_{\boldsymbol{m}} \coloneqq m_1 \cdots m_p$ est la dimension de $S_{\boldsymbol{m}}$ et $\|\cdot\|_n$ est la norme empirique. En intégrant ce résultat contre la loi du design, on obtient :

$$\mathbb{E}\|b - \hat{b}_{\boldsymbol{m}}\|_n^2 = \mathbb{E}\Big[\inf_{t \in S_{\boldsymbol{m}}} \|b - t\|_n^2\Big] + \sigma\frac{D_{\boldsymbol{m}}}{n} \leqslant \inf_{t \in S_{\boldsymbol{m}}} \|b - t\|_\mu^2 + \sigma^2\frac{D_{\boldsymbol{m}}}{n}.$$

Pour obtenir un résultat en norme $\|\cdot\|_\mu$, il faut étudier les déviations de la norme empirique par rapport à son espérance, sur le sous-espace $S_{\boldsymbol{m}}$. En s'inspirant des travaux de Cohen *et al.* (2013), on définit l'évènement :

$$\forall \delta \in \,]0,1[,\ \Omega_{\boldsymbol{m}}(\delta) \coloneqq \Big\{\forall t \in S_{\boldsymbol{m}},\ \|t\|_\mu^2 \leqslant \frac{1}{1-\delta}\|t\|_n^2\Big\} = \Big\{\lambda_{\min}(\mathbf{H}_{\boldsymbol{m}}) \geqslant 1 - \delta\Big\},$$

où $\mathbf{H}_{\boldsymbol{m}}$ est la matrice de Gram relativement au produit scalaire empirique d'une b.o.n. de $S_{\boldsymbol{m}}$ pour le produit scalaire $\langle \cdot, \cdot \rangle_\mu$. En utilisant une inégalité de concentration de Tropp (2012) sur les valeurs propres de matrices aléatoires, on montre que :

$$\mathbb{P}\Big[\Omega_{\boldsymbol{m}}(\delta)^{\mathrm{c}}\Big] \leqslant D_{\boldsymbol{m}} \exp\left(-h(\delta)\frac{n}{L(\boldsymbol{m})\|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}}\right),$$

où $h(\delta) \coloneqq \delta + (1-\delta)\log(1-\delta)$ et où $L(\boldsymbol{m}) \coloneqq \|\sum_{\forall i, k_i < m_i} \varphi_{\boldsymbol{k}}^2\|_\infty$. Ainsi, si on se restreint à la collection de modèles :

$$\mathcal{M}_{n,\alpha}^{(1)} \coloneqq \left\{\boldsymbol{m} \in \mathbb{N}_+^p \,\middle|\, L(\boldsymbol{m})\Big(\|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}} \vee 1\Big) \leqslant \alpha\frac{n}{\log n}\right\}, \tag{1}$$

alors on a :

$$\forall \boldsymbol{m} \in \mathcal{M}_{n,\alpha},\ \mathbb{P}\Big[\Omega_{\boldsymbol{m}}(\delta)^{\mathrm{c}}\Big] \leqslant D_{\boldsymbol{m}}\ n^{-\frac{h(\delta)}{\alpha}} \leqslant n^{-\frac{h(\delta)}{\alpha}+1}.$$

Grâce à ce résultat, on établit le théorème suivant sur la décomposition biais-variance du risque en norme $\|\cdot\|_\mu$.

Théorème 3.1. *Supposons que $b \in \mathrm{L}^{2r}(\mu)$ avec $r \in]1,+\infty]$ et notons $r' \in [1,+\infty[$ son exposant conjugué. Alors pour tous $\alpha \in (0, \frac{1}{2r'+1})$ et $\boldsymbol{m} \in \mathcal{M}_{n,\alpha}^{(1)}$, on a :*

$$\mathbb{E}\|b-\hat{b}_{\boldsymbol{m}}\|_\mu^2 \leqslant C_n(\alpha,r') \inf_{t\in S_{\boldsymbol{m}}} \|b-t\|_\mu^2 + C'(\alpha,r')\,\sigma^2 \frac{D_{\boldsymbol{m}}}{n} + \frac{C''\Big(\|b\|_{\mathrm{L}^{2r}(\mu)},\sigma^2,\alpha\Big)}{n\log n},$$

où les constantes $C_n(\alpha,r')$ et $C'(\alpha,r')$ sont données par :

$$C_n(\alpha,r') \coloneqq 1 + \frac{2}{1-\delta(\alpha,r')}\left(\frac{\alpha}{\Big(1-\delta(\alpha,r')\Big)\log n} \wedge 1\right), \quad C'(\alpha,r') \coloneqq \frac{2}{1-\delta(\alpha,r')},$$

et où $\delta(\alpha,r') \in (0,1)$ tends vers 1 quand α tends vers $\frac{1}{2r'+1}$.

4 Adaptation

Pour choisir le modèle qui optimise le compromis entre le biais et la variance automatiquement, on utilise une procédure de sélection de modèle par pénalisation. Comme on l'a vu précédemment, si on veut que les résultats pour la norme empirique se transfèrent à la norme $\|\cdot\|_\mu$, il faut se restreindre à la collection de modèles $\mathcal{M}_{n,\alpha}^{(1)}$ définie par (1). En pratique cette collection n'est pas accessible car elle dépend de la matrice inconnue $\mathbf{G}_{\boldsymbol{m}}$. On considère donc la version empirique de la collection de modèles $\mathcal{M}_{n,\alpha}^{(1)}$:

$$\widehat{\mathcal{M}}_{n,\beta}^{(1)} \coloneqq \left\{ \boldsymbol{m} \in \mathbb{N}_+^p \;\middle|\; L(\boldsymbol{m})\Big(\|\widehat{\mathbf{G}}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}} \vee 1\Big) \leqslant \beta \frac{n}{\log n} \right\},$$

Suivant l'approche développée par Barron *et al.* (1999) et améliorée par Baraud (2000), on sélectionne le modèle à l'aide d'un critère des moindres carrés pénalisé par la dimension :

$$\hat{\boldsymbol{m}}_1 \coloneqq \operatorname*{arg\,min}_{\boldsymbol{m}\in\widehat{\mathcal{M}}_{n,\beta}^{(1)}} \left(-\|\hat{b}_{\boldsymbol{m}}\|_n^2 + (1+\theta)\sigma^2\frac{D_{\boldsymbol{m}}}{n}\right), \quad \theta > 0. \tag{2}$$

En s'appuyant sur le résultat de Baraud (2000) pour le cas d'un design fixe, on montre que l'estimateur $\hat{b}_{\hat{\boldsymbol{m}}_1}$ est adaptatif pour la norme empirique.

Théorème 4.1. *Si $b \in \mathrm{L}^{2r}(\mu)$ pour $r \in]1,+\infty]$ et si $\mathbb{E}|\varepsilon_1|^q$ est fini pour un certain $q > 6$, alors il existe une constante $\alpha_{\beta,r'} > 0$ qui dépend de β et r' (l'exposant conjugué de r)*

telle que pour tout $\alpha \in]0, \alpha_{\beta,r'}[$*, on ait la majoration suivante sur le risque de* $\hat{b}_{\hat{\boldsymbol{m}}_1}$ *avec* $\hat{\boldsymbol{m}}_1$ *défini par* (2) *:*

$$\mathbb{E}\|b - \hat{b}_{\hat{\boldsymbol{m}}_1}\|_n^2 \leqslant C(\theta) \inf_{\boldsymbol{m} \in \mathcal{M}_{n,\alpha}^{(1)}} \left(\inf_{t \in S_{\boldsymbol{m}}} \|b - t\|_\mu^2 + \sigma^2 \frac{D_{\boldsymbol{m}}}{n} \right) + \sigma^2 \frac{\Sigma(\theta, q)}{n} + R_n,$$

où :

$$C(\theta) \coloneqq (2 + 8\theta^{-1})(1 + \theta), \qquad \Sigma(\theta, q) \coloneqq C''(\theta, q) \frac{\mathbb{E}|\varepsilon_1|^q}{\sigma^q} \sum_{\boldsymbol{m} \in \mathbb{N}_+^p} D_{\boldsymbol{m}}^{-(\frac{q}{2}-2)},$$

et où le terme de reste est donné par :

$$R_n \coloneqq C'(\|b\|_{\mathrm{L}^{2r}(\mu)}, \sigma^2) \frac{(\log n)^{(p-1)/r'}}{n^{\kappa(\alpha,\beta)/r'}},$$

avec $\kappa(\alpha, \beta)$ *une constante positive vérifiant* $\frac{\kappa(\alpha,\beta)}{r'} > 1$ *et* $\frac{\kappa(\alpha,\beta)}{r'} \to 1$ *quand* $\alpha \to \alpha_{\beta,r'}$*.*

Pour transférer le résultat précédent de la norme empirique à la norme $\|\cdot\|_\mu$, on utilise encore une fois des inégalités de concentration matricielles sur $\widehat{\mathbf{G}}_{\boldsymbol{m}}$. Cependant, on fait une distinction entre le cas A compact et le cas A non compact. Lorsque A est compact, on peut faire l'hypothèse usuelle que la densité $\frac{\mathrm{d}\mu}{\mathrm{d}\nu}$ est minorée sur A et utiliser l'inégalité de Chernov matricielle de Gittens et Tropp (2011). Ce lemme dépend de façon essentielle de l'hypothèse de minoration de la densité, on ne peut donc pas l'utiliser dans le cas non compact.

Dans le cas non compact, on utilise à la place l'inégalité de Bernstein matricielle (cf. Tropp (2012)). Cependant, cette inégalité ne fait pas intervenir les mêmes quantités dans la majoration que l'inégalité de Chernov matricielle et il faut réduire la collection de modèles pour pouvoir l'utiliser.

1. **Cas compact**. On suppose qu'il existe $f_0 > 0$ telle que pour tout $x \in A$, $\frac{\mathrm{d}\mu}{\mathrm{d}\nu}(x) > f_0$. Dans ce cas, on a :

$$\|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}} = \sup_{t \in S_{\boldsymbol{m}} \setminus \{0\}} \frac{\|t\|_\nu^2}{\|t\|_\mu^2} \leqslant \frac{1}{f_0}.$$

2. **Cas général**. On considère des collections de modèles plus petites :

$$\mathcal{M}_{n,\alpha}^{(2)} \coloneqq \left\{ \boldsymbol{m} \in \mathbb{N}_+^p \,\middle|\, L(\boldsymbol{m}) \left(\|\mathbf{G}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}^2 \vee 1 \right) \leqslant \alpha \frac{n}{\log n} \right\},$$
$$\widehat{\mathcal{M}}_{n,\beta}^{(2)} \coloneqq \left\{ \boldsymbol{m} \in \mathbb{N}_+^p \,\middle|\, L(\boldsymbol{m}) \left(\|\widehat{\mathbf{G}}_{\boldsymbol{m}}^{-1}\|_{\mathrm{op}}^2 \vee 1 \right) \leqslant \beta \frac{n}{\log n} \right\},$$

   et on sélectionne $\hat{\boldsymbol{m}}_2 \in \widehat{\mathcal{M}}_{n,\beta}^{(2)}$ comme minimisant :

$$\hat{\boldsymbol{m}}_2 \coloneqq \operatorname*{arg\,min}_{\boldsymbol{m} \in \widehat{\mathcal{M}}_{n,\beta}^{(2)}} \left(-\|\hat{b}_{\boldsymbol{m}}\|_n^2 + (1 + \theta)\sigma^2 \frac{D_{\boldsymbol{m}}}{n} \right), \quad \theta > 0. \tag{3}$$

Théorème 4.2. *Soit $r \in]1, +\infty]$ et supposons que $b \in \mathrm{L}^{2r}(\mu)$ et que $\mathbb{E}|\varepsilon_1|^q$ est fini pour un certain $q > 6$.*

1. ***Cas compact**. Soit $f_0 > 0$ tel que $\frac{\mathrm{d}\mu}{\mathrm{d}\nu}(x) \geqslant f_0$ pour tout $x \in A$, alors il existe $\beta_{f_0,r'} > 0$ tel que pour tout $\beta \in]0, \beta_{f_0,r'}[$, il existe $\alpha_{\beta,r'} > 0$ tel que pour tout $\alpha \in]0, \alpha_{\beta,r'}[$, l'estimateur $\hat{b}_{\hat{\boldsymbol{m}}_1}$ avec $\hat{\boldsymbol{m}}_1$ défini par (2) vérifie l'inégalité oracle :*

$$\mathbb{E}\|b - \hat{b}_{\hat{\boldsymbol{m}}_1}\|_\mu^2 \leqslant C(\theta, \beta, r) \inf_{\boldsymbol{m} \in \mathcal{M}_{n,\alpha}^{(1)}} \left(\inf_{t \in S_{\boldsymbol{m}}} \|b - t\|_\mu^2 + \sigma^2 \frac{D_{\boldsymbol{m}}}{n} \right) + C'(\beta, r)\sigma^2 \frac{\Sigma(\theta, q)}{n} + R_n,$$

où le terme de reste est donné par :

$$R_n = C''\Big(\|b\|_{\mathrm{L}^{2r}(\mu)}, \sigma^2, \beta, r\Big) \left(n^{-\frac{\kappa(\alpha,\beta)}{r'}} (\log n)^{\frac{p-1}{r'}} + n^{-\lambda(\beta,r,f_0)} (\log n)^{\frac{p-1}{r'}-1} \right),$$

avec $\lambda(\beta, r, f_0) > 1$ et $\frac{\kappa(\alpha,\beta)}{r'} > 1$.

2. ***Cas général**. Soit $B \coloneqq (\|\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\|_\infty + \frac{2}{3})^{-1}$, alors il existe $\beta_{B,r'} > 0$ tel que pour tout $\beta \in]0, \beta_{B,r'}[$, il existe $\tilde{\alpha}_{\beta,r'} > 0$ tel que pour tout $\alpha \in]0, \tilde{\alpha}_{\beta,r'}[$, l'estimateur $\hat{b}_{\hat{\boldsymbol{m}}_2}$ avec $\hat{\boldsymbol{m}}_2$ défini par (3) vérifie l'inégalité oracle :*

$$\mathbb{E}\|b - \hat{b}_{\hat{\boldsymbol{m}}_2}\|_\mu^2 \leqslant C(\theta, \beta, r) \inf_{\boldsymbol{m} \in \mathcal{M}_{n,\alpha}^{(2)}} \left(\inf_{t \in S_{\boldsymbol{m}}} \|b - t\|_\mu^2 + \sigma^2 \frac{D_{\boldsymbol{m}}}{n} \right) + C'(\beta, r)\sigma^2 \frac{\Sigma(\theta, q)}{n} + R_n,$$

où le terme de reste est donné par :

$$R_n = C''\Big(\|b\|_{\mathrm{L}^{2r}(\mu)}, \sigma^2, \beta, r\Big) \left(n^{-\frac{\tilde{\kappa}(\alpha,\beta)}{r'}} (\log n)^{\frac{p-1}{r'}} + n^{-\lambda(\beta,r,B)} (\log n)^{\frac{p-1}{r'}-1} \right),$$

avec $\lambda(\beta, r, B) > 1$ et $\frac{\tilde{\kappa}(\alpha,\beta)}{r'} > 1$.

Bibliographie

Baraud, Y. (2000). Model selection for regression on a fixed design. *Probability Theory and Related Fields*, 117(4), pp. 467–493.

Baraud, Y. (2002). Model selection for regression on a random design. *ESAIM: Probability and Statistics*, 6, pp. 127–146.

Barron, A., Birgé, L. et Massart, P. (1999). Risk bounds for model selection via penalization. *Probability and Related Fields*, 113(3), pp. 301–413.

Cohen, A., Devenport, A. et Leviatan, D. (2013). On the Stability and Accuracy of Least Squares Approximations. *Foundations of Computational Mathematics*, 13(5), pp. 819–834.

Comte, F. et Genon-Catalot, V. (2020). Regression function estimation as a partly inverse problem. *Annals of the Institute of Statistical Mathematics*, 72(4), pp. 1023–1054.

Gittens, A. et Tropp, J. A. (2011). Tail bounds for all eigenvalues of a sum of randim matrices. *arXiv 1104.4513 [math.PR]*.

Tropp, J. A. (2011). User-Friendly Tail Bounds for Sums of Random Matrices. *Foundations of Computational Mathematics*, 12(4), pp. 389–434.

Phénomène de Stein pour l'estimation multiple de moyennes en grande dimension.

Jean-Baptiste Fermanian [1] & Hannah Marienwald [2] & Gilles Blanchard [3]

[1] *Université Paris-Saclay, Laboratoire de Mathématiques d'Orsay, 91405, Orsay, France; jean-baptiste.fermanian@universite-paris-saclay.fr.*

[2] *Universität Potsdam, Potsdam, Germany, et Technische Universität Berlin, Berlin, Germany; marienwald@uni-potsdam.de.*

[3] *Université Paris-Saclay, CNRS, Inria, Laboratoire de Mathématiques d'Orsay, 91405, Orsay, France; gilles.blanchard@universite-paris-saclay.fr.*

Résumé. Nous proposons une amélioration de l'estimation multiple de moyennes, dont le but est l'estimation conjointe des moyennes de plusieurs distributions à partir d'échantillons séparés et indépendants. L'approche naïve consiste à estimer la moyenne empirique de chaque échantillon séparément, alors que la méthode que nous proposons exploite les similarités possibles entre les moyennes, sans qu'aucune information a priori ne soit connue. Tout d'abord, pour chaque échantillon, on détecte les moyennes similaires ou voisines à l'aide de tests multiples. Ensuite, chaque estimateur naïf est décalé vers la moyenne locale de ses voisins. Cette transformation fait écho à l'estimateur de James Stein où la moyenne empirique est décalée vers zéro. Ainsi, ici notre point de référence n'est pas zéro mais la moyenne des voisins de l'échantillon considéré. Bien que du biais soit ajouté, l'estimation est améliorée par la réduction de la variance. Cette amélioration peut être significative lorsque la dimension de l'espace est importante, démontrant un phénomène de "bénédiction de la dimension". Une application de cette approche est l'estimation multiple de kernel mean embeddings, qui joue un rôle important dans de nombreuses applications modernes. Les résultats théoriques sont vérifiés sur des données artificielles et réelles.

Mots-clés. Estimation multiple, test multiple, phénomène de Stein, grande dimension, kernel mean embedding.

Abstract. We propose an improved estimator for the multi-task averaging problem, whose goal is the joint estimation of the means of multiple distributions using separate, independent data sets. The naive approach is to take the empirical mean of each data set individually, whereas the proposed method exploits similarities between tasks, without any related information being known in advance. First, for each data set, similar or neighboring means are determined from the data by multiple testing. Then each naive estimator is shrunk towards the local average of its neighbors. This shrinkage echoes James Stein's estimator where the empirical mean is shrunk to zero. Here the reference point is not zero but the means of the neighbours. Although bias is added, the estimate is improved by reducing the variance. This improvement can be significant when the dimension of

the input space is large, demonstrating a "blessing of dimensionality" phenomenon. An application of this approach is the estimation of multiple kernel mean embeddings, which plays an important role in many modern applications. The theoretical results are verified on artificial and real world data.

Keywords. Multiple estimations, multiple tests, Stein's phenomenon, high dimension, kernel mean embedding.

1 Introduction

Considérant B lois de probabilités $\mathbb{P}_1, \dots, \mathbb{P}_B$ sur un espace de Hilbert $\mathcal{X}$, notre objectif est d'estimer leurs espérances $\mu_1, \dots, \mu_B$ en minimisant le risque quadratique ("Mean Squared Error") individuel ou moyen :

$$\mathrm{MSE}(b, \widehat{\mu}_b) := \mathbb{E}\left[\|\mu_b - \widehat{\mu}_b\|^2\right], \quad \overline{\mathrm{MSE}}(\widehat{\mu}_\bullet) := \frac{1}{B}\sum_{b=1}^{M} \mathrm{MSE}(b, \widehat{\mu}_b)\,.$$

Ce problème très classique d'estimation multiple est parfois appelé "multi task averaging" (MTA) (Feldman et al. (2014)). Les tâches sont les distributions $\mathbb{P}_b$ et pour chacune d'entre elles nous avons un échantillon appelé *bag* $X_\bullet^{(b)}$. Nous supposons :

$$\begin{cases} X_\bullet^{(b)} = (X_i^{(b)})_{1 \le i \le N_b}, \text{ où les } X_i^{(b)} \text{ sont i.i.d. de loi } \mathbb{P}_b, b \in [\![B]\!] = \{1, \dots, B\}; \\ X_i^{(b)} \in \mathcal{X}; \\ \mu_b = \mathbb{E}_{X \sim \mathbb{P}_b}[X]; \\ X_\bullet^{(1)}, \dots, X_\bullet^{(B)} \text{ sont indépendants.} \end{cases}$$

Le cas qui nous intéresse est celui où l'espace $\mathcal{X}$ est de grande dimension, finie, par exemple lorsque $\mathcal{X} = \mathbb{R}^d$ avec d grand par rapport au nombre de bags B et aux tailles d'échantillons N_b, ou infinie avec des espaces fonctionnels comme par exemple des espaces de Hilbert à noyaux reproduisants (RKHS) où nous chercherons à estimer des Kernel Mean Embeddings (KME).

Nous cherchons à construire des estimateurs plus performants que les estimateurs naïfs qui consistent à prendre la moyenne empirique de chacun des échantillons :

$$\widehat{\mu}_b^{\mathtt{NE}} := \frac{1}{N_b}\sum_{i=1}^{N_b} X_i^{(b)} \quad \text{avec} \quad \mathrm{MSE}(b, \widehat{\mu}_b^{\mathtt{NE}}) = \frac{\operatorname{Tr}\Sigma_b}{N_b}\,, \tag{1}$$

où $\Sigma_b := \mathrm{Cov}\left[X_1^{(b)}\right]$ est la covariance de la distribution $\mathbb{P}_b$.

2 Phénomène de Stein

Pour l'estimation d'une unique moyenne, $B = 1$, de loi gaussienne isotrope de variance connue, $\mathbb{P}_1 = \mathcal{N}(\mu_1, \sigma^2 I_d)$, Stein (1956) montre que l'estimateur empirique (1) n'est pas admissible lorsque que la dimension d de l'espace est supérieure à 3. C'est-à-dire qu'il existe un estimateur $\widehat{\mu}$ de risque strictement inférieur à l'estimateur empirique pour toute moyenne μ_1. Cet estimateur est construit en décalant la moyenne empirique vers un certain point de référence $\mu_\circ \in \mathbb{R}^d$:

$$\widehat{\mu}_1^{\gamma} = \gamma \mu_\circ + (1-\gamma)\widehat{\mu}_1^{\texttt{NE}}\,. \tag{2}$$

On retrouve l'estimateur de James-Stein (James, Stein (1961)) en prenant $\gamma_{\texttt{JS}} = \frac{(d-2)\sigma_N^2}{\|\widehat{\mu}_1^{\texttt{NE}} - \mu_\circ\|^2}$, où $\sigma_N^2 := \sigma^2/N_1$. En réalité on parle d'estimateur de James-Stein lorsque $\mu_\circ = 0$, mais on peut prendre $\mu_\circ$ quelconque par un changement d'origine. On peut majorer le risque de cet estimateur :

$$\mathrm{MSE}(1, \widehat{\mu}_1^{\gamma_{\texttt{JS}}}) \leq d\sigma_N^2 - \frac{(d-2)^2\sigma_N^4}{(d-2)\sigma_N^2 + \|\mu_1 - \mu_\circ\|^2} < d\sigma_N^2 = \mathrm{MSE}(1, \widehat{\mu}_1^{\texttt{NE}})\,.$$

On observe que le gain obtenu par rapport à l'estimation naïve augmente :

- lorsque la dimension de l'espace augmente : dans notre méthode les données ne seront plus forcément isotropes ou seront de dimension infinie, le rôle de d sera joué par ce qu'on appelle la *dimension effective* de la loi de probabilités ;
- lorsque la distance $\|\mu_1 - \mu_\circ\|^2$ diminue : l'estimateur de James-Stein choisit arbitrairement le point de référence $\mu_\circ$. L'idée de notre estimateur est d'utiliser les autres bags pour trouver à l'aide de tests un $\mu_\circ$ moins arbitraire. Dans certaines situations, même si on sait que les moyennes sont a priori différentes, celles-ci pourraient cependant être proches : notre méthode cherche à exploiter ces similarités.

L'estimateur de James-Stein atteint la borne minimax de Pinsker (1980). Si l'on considère l'ensemble $\mathcal{G}_d(\tau, \mu_\circ)$ des lois de moyenne dans une boule autour de $\mu_\circ$: $\mathcal{G}_d(\tau, \mu_\circ) = \{\mathcal{N}(\mu, \sigma^2 I_d), \|\mu - \mu_\circ\|^2 \leq \tau\sigma^2 d\}$, alors :

$$\lim_{d\to\infty} \inf_{\widehat{\mu}} \sup_{\mathbb{P}\in\mathcal{G}_d(\tau,\mu_\circ)} \frac{\mathrm{MSE}(\widehat{\mu})}{d\sigma^2} = \frac{\tau}{1+\tau}\,. \tag{3}$$

3 Distance minimale de séparation de moyennes

Pour détecter ces similarités entre moyennes, il est important de connaître jusqu'à quelle distance on est capable de distinguer deux moyennes. Cette distance minimale de séparation va dépendre des dimensions effectives des distributions.

3.1 Dimension effective

La dimension effective d'une loi de probabilité $\mathbb{P}$ est une notion de mesure de ses degrés de libertés. C'est une quantité qui apparaît naturellement à la place de la dimension lorsque les données ne sont pas isotropes. Nous considérons ici pour une certaine loi $\mathbb{P}$ deux dimensions différentes définies toutes deux à partir de l'opérateur de covariance Σ de la loi :

$$d^e(\mathbb{P}) := \frac{\|\Sigma\|_1}{\|\Sigma\|_\infty}\,, \quad d^*(\mathbb{P}) := \frac{\|\Sigma\|_2^2}{\|\Sigma\|_\infty}\,, \quad \text{où } \Sigma = \mathrm{Cov}[\mathbb{P}].$$

Ces dimensions sont égales à la dimension de l'espace lorsque les données sont isotropes : si $\Sigma = \sigma^2 I_d$, alors on a $d^e = \mathrm{Tr}\,\Sigma/\|\Sigma\|_{op} = d$ et $d^* = \mathrm{Tr}\,\Sigma^2/\|\Sigma\|_\infty^2 = d$. Pour les problèmes d'estimation c'est d^e qui interviendra alors que pour les problèmes de tests ce sera plutôt d^*. On peut remarquer que $d^* \leq d^e$.

3.2 Distance minimale de séparation de moyennes

Considérons le problème de test d'égalité entre les moyennes μ_1 et μ_2 :

$$(H_0) : \|\mu_1 - \mu_2\|^2 = 0\,, \quad (H_1(\delta)) : \|\mu_1 - \mu_2\|^2 > \delta^2\,. \tag{4}$$

On appelle *distance minimale de séparation* $\delta^*(\alpha)$ pour $\alpha \in (0,1)$, la distance minimale entre μ_1 et μ_2 pour laquelle on est capable de détecter si celles-ci sont différentes. Plus formellement, on définit δ^* par:

$$\delta^*(\alpha) = \inf\Big\{\delta \geq 0 \Big| \exists \,\text{test } T \ : \mathbb{P}_{(H_0)}(T=1) + \mathbb{P}_{(H_1(\delta))}(T=0) \leq \alpha\Big\}\,. \tag{5}$$

Dans le cas où $\mathrm{Cov}[\mathbb{P}_1] = \mathrm{Cov}[\mathbb{P}_2] = \Sigma$ et $N_1 = N_2 = N$, alors on peut contrôler simplement la distance de séparation (5) à l'aide de la dimension effective d^* (voir Blanchard et al. (2018) pour le cas isotrope et Blanchard et Fermanian (2021) pour le cas non isotrope) :

$$\delta^*(\alpha) \simeq (d^*)^{1/4}\sqrt{\|\Sigma\|_\infty/N}\,. \tag{6}$$

4 Estimation multiple de moyennes

4.1 Echantillons gaussiens homogènes

Intéressons nous au cas homogène où chaque distribution $\mathbb{P}_b$ a la même covariance et les tailles d'échantillons sont égales : $N_b = N$ et $\Sigma_b = \Sigma$ pour tout $b \in [\![B]\!]$. L'idée maintenant est de construire un estimateur de μ_1 de la forme (2) mais en utilisant les échantillons $X_\bullet^{(2)}, \ldots, X_\bullet^{(B)}$ pour choisir $\mu_\circ$. Supposons tout d'abord que nous connaissons déjà les

moyennes proches de μ_1, i.e. nous avons à disposition un ensemble $V(\tau) \subset [\![B]\!]$, $\tau > 0$ fixé tel que pour tout $b \in V := V(\tau)$:

$$\|\mu_1 - \mu_b\|^2 \leq \tau \sigma_N^2 d^e = \tau \mathrm{MSE}(1, \widehat{\mu}_1^{\mathtt{NE}}), \quad \text{où } \sigma_N^2 = \frac{\|\Sigma\|_\infty}{N} \ .$$

On considère l'estimateur :

$$\tilde{\mu}_1 = \gamma \widehat{\mu}_1^{\mathtt{NE}} + (1-\gamma) \frac{1}{|V|} \sum_{b \in V} \widehat{\mu}_b^{\mathtt{NE}} .$$

Choisissons $\gamma = \frac{\tau(|V|-1)}{\tau(|V|-1)+|V|}$, on obtient ainsi une amélioration par rapport à l'estimateur naïf. On a en effet :

$$\frac{\mathrm{MSE}(1, \tilde{\mu}_1)}{\sigma_N^2 d^e} \leq \frac{\tau + |V|^{-1}}{1+\tau} .$$

Avec la connaissance des voisins nous approchons la borne minimax de Pinsker (3) où la dimension effective d^e remplace la dimension de l'espace. L'idée ici sera de se passer de cette information en utilisant des tests multiples pour détecter les voisins. Pour cela, nous construisons une famille de tests $(T_{1b})_{b \in [\![B]\!]}$ pour les hypothèses :

$$(H_0) : \|\mu_1 - \mu_b\|^2 > \tau \sigma_N^2 d^e, \quad \text{contre} \quad (H_1) : \|\mu_1 - \mu_b\|^2 \leq \tau' \sigma_N^2 d^e .$$

On prend alors comme ensemble V l'ensemble les échantillons correspondant à un rejet de (H_0) : $V := \{b | T_{1b} = 1\}$. Notons que l'usage de tests implique une valeur minimale pour τ, en effet d'après (6), τ doit être supérieur à $(d^*)^{1/2}/d^e$ pour être capable de contrôler les erreurs des tests. Comme pour l'estimateur de James-Stein, le gain par rapport à l'estimateur empirique augmente avec la dimension de l'espace. L'avantage est que nous ne choisissons plus $\mu_\circ$ arbitrairement.

On peut de même contrôler le risque moyen sur les tâches. On construit de la même manière une famille de tests $(T_{ab})_{a,b \in [\![B]\!]}$ pour détecter si les moyennes μ_a et μ_b sont proches en contrôlant les erreurs de première et seconde espèce. Avec grande probabilité on peut ainsi majorer le risque moyen :

$$\frac{\overline{\mathrm{MSE}}(\tilde{\mu}_\bullet)}{\sigma_N^2 d^e} \leq \frac{\tau + \frac{\mathcal{N}}{B}}{1+\tau} ,$$

où $\mathcal{N}$ est le nombre de couverture des moyennes μ_b par des boules de rayons $\sqrt{\tau' \sigma_N^2 d^e}/2$. On peut généraliser ces résultats à des données non homogènes, c'est-à-dire avec des covariances et des tailles d'échantillons différentes ainsi qu'à des données bornées. Le but étant d'appliquer cette méthode à l'estimation de Kernel Mean Embedding (KME).

4.2 Application à l'estimation de kernel mean embeddings

Etant donné une carte $\phi : \mathcal{X} \to \mathcal{H}$ où $\mathcal{H}$ est un espace de Hilbert, on appelle KME d'une distribution $\mathbb{P}$ sur $\mathcal{X}$ l'élément $\mu_{\mathbb{P}}$ de $\mathcal{H}$ défini par :

$$\mu_{\mathbb{P}} = \phi(\mathbb{P}) := \mathbb{E}_{X\sim\mathbb{P}}[\phi(X)]\,.$$

Sous certaines conditions peu coûteuses sur ϕ, l'application $\mathbb{P} \mapsto \mu_{\mathbb{P}}$ peut être injective ce qui donne un grand intérêt à cette quantité. L'utilisation de KME trouve de nombreuses applications, en régression par exemple, et a l'avantage d'avoir un faible coût algorithmique. Estimer les KME $\mu_{\mathbb{P}_b}$ de distributions $\mathbb{P}_1, \dots, \mathbb{P}_B$ se ramène au problème d'estimations multiples pour les échantillons $(\phi(X_i^{(b)}))_i$ avec $b \in [\![B]\!]$.
L'estimation de la matrice de Gram des KMEs :

$$G = \left(\langle \mu_{\mathbb{P}_a}, \mu_{\mathbb{P}_b}\rangle_{\mathcal{H}}\right)_{a,b\in[\![B]\!]}\,,$$

est souvent nécessaire. On peut relier l'erreur d'estimation de la matrice de Gram G par $\widehat{G} = \left(\langle \widehat{\mu}_{\mathbb{P}_a}, \widehat{\mu}_{\mathbb{P}_b}\rangle_{\mathcal{H}}\right)_{a,b\in[\![B]\!]}$ à l'erreur moyenne d'estimation de chacune des moyennes :

$$\mathbb{E}\left[\left\|\frac{1}{B}(G-\widehat{G})\right\|\right] \leq 4\|\phi\|_\infty^2 \overline{\mathrm{MSE}}_{\mathcal{H}}(\widehat{\mu}_\bullet)\,.$$

Ainsi une amélioration de l'estimation individuelle des KME améliorera l'estimation de la matrice de Gram.

Bibliographie

Blanchard, G., Carpentier, A., and Gutzeit, M. (2018). Minimax Euclidean separation rates for testing convex hypotheses in $\mathbb{R}^d$. *Electronic Journal of Statistics* 12 (2), 3713–3735.
Blanchard G., Fermanian J-B. (2021). Nonasymptotic one-and two-sample tests in high dimension with unknown covariance structure. (ArXiv/ To appear in: Festschrift in the honor of V. Spokoiny).
Feldman, S., Gupta, M. R., and Frigyik, B. A. (2014). Revisiting Stein's paradox: multi-task averaging. *Journal of Machine Learning Research*, 15(106):3621–3662.
James, W. and Stein, C. (1961). Estimation with quadratic loss. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, pages 361–379. University of California Press.
Pinsker M. S. (1980). Optimal Filtering of Square-Integrable Signals in Gaussian Noise, *Problems Inform. Transmission*, 16:2, 120–133.
Stein, C. (1956), Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proceedings of the Third Berkeley symposium on mathematical statistics and probability*. Vol. 1. No. 1.

Estimation of high dimensional gamma convolutions through random projections

Oskar Laverny [1]

[1] *Université Claude Bernard Lyon 1, Insitut Camille Jordan UMR 5208, France & SCOR SE, oskar.laverny@univ-lyon1.fr*

Résumé. La classe des convolutions généralisées de lois gammas est définie par une structure convolutionelle semi-paramétrique. La flexibilité de ses structures de dépendance, les possibilités marginales et la structure convolutionelle pratique de ces distributions en fait une classe intéressante pour le praticien. Cependant, l'estimation de ces distributions quand la dimension augmente est complexe et représente un défi. Nous proposons une procédure d'estimation stochastique basée sur une approximation de l'erreur quadratique en base de Laguerre via des cumulants (shiftés), évaluée sur des projections aléatoires du jeu de donnée. À travers l'analyse de la perte via des outils venant des cubatures de Grassmanniens, de l'optimisation de mesure sparse et de flots de gradients en espace Wasserstein, nous démontrons la convergence de la descente de gradient stochastique vers un estimateur consistent de la distribution de grande dimension. Nous proposons plusieurs exemples en dimension base et en grande dimension.

Mots-clés. Convolutions de gamma généralisées, estimation, grande dimension, mesure sparse, flots de gradients ...

Abstract. The multivariate generalized gamma convolutions is a class of distributions defined by a convolutional semi-parametric structure. Their flexible dependence structures, the marginal possibilities and their useful convolutional expression makes them appealing to the practitioner. However, fitting such distributions when the dimension gets high is a challenge. We propose stochastic estimations procedures based on (shifted) cumulants approximation of a Laguerre integrated square error, evaluated on random projections of the dataset. Through the analysis of our loss via tools from Grassmannian cubatures, sparse optimization on measures and Wasserstein gradient flows, we show the convergence of the stochastic gradient descent to a proper estimator of the high dimensional distribution. We propose several examples on both low and high-dimensional settings.

Keywords. Generalized gamma convolutions, estimation, high-dimension, sparse measure, gradient flows, ...

Une convolution de lois gamma en dimension d est une distribution de probabilité correspondant au vecteur aléatoire $\boldsymbol{X}$ tel que :

$$\begin{pmatrix} X_1 \\ ... \\ X_d \end{pmatrix} = \begin{pmatrix} s_{1,1} & ... & ... & s_{1,n} \\ ... & ... & ... & ... \\ s_{d,1} & ... & ... & s_{d,n} \end{pmatrix} \cdot \begin{pmatrix} G_1 \\ ... \\ G_n \end{pmatrix},$$

ou la matrice $\boldsymbol{s}$ est à valeurs positive ou nulle, les paramètres d'échelle, et les variables aléatoires $G_1, ...G_n$ suivent toutes des lois gammas indépendantes, de paramètres de forme respectifs $\alpha_1, ..., \alpha_n$ et d'échelle 1. Typiquement dans ces modèles, la matrice $\boldsymbol{s}$ est sparse; Le modèle peut être sous-déterminé ($n \gg d$), ou au contraire sous-déterminé ($d \gg n$).

La généralisation consiste à faire tendre n vers l'infini, et l'on définit alors la distribution du vecteur aléatoire $\boldsymbol{X}$ par sa fonction génératrice des cumulants:

$$K(\boldsymbol{t}) = \ln \mathbb{E}\left(e^{\langle \boldsymbol{t}, \boldsymbol{X} \rangle}\right) = -\int \ln\left(1 - \langle \boldsymbol{s}, \boldsymbol{t}\right) \nu(d\boldsymbol{s}),$$

ou ν est une mesure positive sur $\mathbb{R}_+^d$, que l'on appelle la mesure de Thorin. Le cas précédent se retrouve facilement en prenant $\nu = \sum_{i=1}^n \alpha_i \delta_{\boldsymbol{s}_i}$, et des conditions sur ν dans le cas continu qui permette d'assurer que K est bien une fonction génératrice des cumulants sont disponibles dans Pérez-Abreu & Stelzer (2012).

Ces modèles ont été étudier notamment par Bondesson (1992, 2009, 2015), en tant que généralisation du travail de Thorin (1977, 1977a) dans le cas univarié. Si beaucoup de résultats théoriques existent sur les propriétés de ces distributions (la classe univarié contient notamment les log-normales, les Pareto, certaines distributions α-stables...), la littérature ne proposait que peu de possibilité d'estimation de ces distributions, surtout quand la dimension d du problème est plus grande que 1 (et d'autant plus pour d grand).

On peut approcher la densité de ces distributions via la base de Laguerre, que l'on peut décrire en dimension d par l'ensemble des fonctions $\left\{\phi_{\boldsymbol{k}}, \boldsymbol{k} \in \mathbb{N}^d\right\}$ telles que :

$$\phi_k(x) = \sqrt{2} \sum_{j=0}^{k} \binom{k}{j} \frac{(-2x)^j}{j!} e^{-x} \text{ pour } k \in \mathbb{N} \text{ et } \phi_{\boldsymbol{k}}(\boldsymbol{x}) = \prod_{i=1}^{d} \phi_{k_i}(x_i) \text{ for } \boldsymbol{k} \in \mathbb{N}^d.$$

Les premiers coefficients de la densité f de $\boldsymbol{X}$ dans la base sont noté $\boldsymbol{a} = \{a_{\boldsymbol{k}}, \boldsymbol{k} \leq \boldsymbol{m}\}$, et s'exprime de manière bijective en fonction de certains moments généralisés de ν, que l'on appelle les moments de Thorin $\boldsymbol{\tau} = \{\tau_{\boldsymbol{k}}, \boldsymbol{k} \leq \boldsymbol{m}\}$, et qui sont définis par :

1. $\tau_{\boldsymbol{0}}(\delta_{\boldsymbol{s}}) = -\ln(1 + |\boldsymbol{s}|)$

2. $\tau_{\boldsymbol{k}}(\delta_{\boldsymbol{s}}) = \boldsymbol{s}^{\boldsymbol{k}} \left(1 + |\boldsymbol{s}|\right)^{-|\boldsymbol{k}|}$ for $\boldsymbol{k} \neq 0$, $\boldsymbol{k} \in \mathbb{N}^d$.
3. $\tau_{\boldsymbol{k}}(\nu) = \int \tau_{\boldsymbol{k}}(\delta_{\boldsymbol{s}})\nu(d\boldsymbol{s})$

On obtient alors $\boldsymbol{a} = \boldsymbol{B}(\boldsymbol{\tau})$, ou la fonction $\boldsymbol{B}$ est numériquement complexe, mais bijective et infiniment dérivable (elle contient, entre autre, des polynômes de Bell). Le paramètre $\boldsymbol{m} \in \mathbb{N}^d$ gère la précision de l'approximation.

À partir d'un échantillon aléatoire $\boldsymbol{x} \in \mathbb{R}_+^{N \times d}$ de N observations en dimension d, i.i.d, du vecteur aléatoire $\boldsymbol{X}$, on peut estimer les coefficients de Laguerre par Monte-Carlo:

$$\hat{a}_{\boldsymbol{k}} = \sum_{i=1}^{N} \phi_{\boldsymbol{k}}(\boldsymbol{x}_i).$$

Une perte est alors toute trouvée pour estimer ces distributions:

$$\mathcal{L}(\boldsymbol{x}, \nu) = \|\hat{\boldsymbol{a}}(\boldsymbol{x}) - \boldsymbol{B}(\boldsymbol{\tau}(\nu))|_2^2.$$

Perte que l'on approxime dans un premier temps a l'ordre un via une approximation de Taylor sur la fonction $\boldsymbol{B}$, autour de $\hat{\boldsymbol{\tau}}(\boldsymbol{x}) = \boldsymbol{B}^{-1}(\hat{\boldsymbol{a}}(\boldsymbol{x}))$:

$$\widehat{\mathcal{L}}(\boldsymbol{x}, \nu) = \|\hat{\boldsymbol{\tau}}(\boldsymbol{x}) - \boldsymbol{\tau}(\nu)|_{\boldsymbol{\nabla}(\boldsymbol{x})}^2 \approx \mathcal{L}(\boldsymbol{x}, \nu),$$

ou la norme $\|\cdot\|_{\boldsymbol{\nabla}(\boldsymbol{x})}$ correspond à cette approximation ($\boldsymbol{\nabla}(\boldsymbol{x})$ est la jacobienne de $\boldsymbol{B}$ prise en $\hat{\boldsymbol{\tau}}(\boldsymbol{x})$).

Cependant, si cette approximation de premier ordre permet de simplifier la problématique numériquement (la fonction $\boldsymbol{B}$ ne devant être calculé qu'une seule fois au début de l'optimisation, et non plus à chaque gradient), elle ne règle pas le problème majeur de l'estimation de densité via des bases orthonormales : le problème de la dimension.

En effet, lorsque d augmente, la nécessité de conserver dans chaque direction une précision suffisante nous force à faire augmenter (exponentiellement en la dimension) le nombre de coefficients, rendant le calcul même des $\hat{\boldsymbol{\tau}}(\boldsymbol{x})$ tout simplement impossible et l'optimisation encore moins.

Nous proposons alors d'utiliser une technique de projection aléatoire. En effet, on remarque que pour tout $\boldsymbol{c} \in \mathbb{R}_+^d$, $\langle \boldsymbol{c}, \boldsymbol{X} \rangle$ suit une convolution de lois gammas univarié, de mesure de Thorin $\nu_{\langle \boldsymbol{c} \rangle}$ définie par :

$$\nu_{\langle \boldsymbol{c} \rangle}(s) = \nu\left(\left\{\boldsymbol{y} \in \mathbb{R}_+^d : \langle \boldsymbol{c}, \boldsymbol{y} \rangle = s\right\}\right) \ \forall s \in \mathbb{R}_+^d.$$

Ainsi, nous proposons d'utiliser la perte $\widetilde{\mathcal{L}}$ définie par :

$$\widetilde{\mathcal{L}}(\boldsymbol{x}, \nu) := \int_{[0,1]^d} \widehat{\mathcal{L}}\left(\langle \boldsymbol{c}, \boldsymbol{x} \rangle, \nu_{\langle \boldsymbol{c} \rangle}\right) d\boldsymbol{c}$$
$$= \int_{[0,1]^d} \|\hat{\boldsymbol{\tau}}(\langle \boldsymbol{c}, \boldsymbol{x} \rangle) - \boldsymbol{\tau}(\nu_{\langle \boldsymbol{c} \rangle})\|^2_{\nabla(\langle \boldsymbol{c}, \boldsymbol{x} \rangle)} \, d\boldsymbol{c},$$

Par des arguments venant de la théorie des projections aléatoires, de la théorie de Grassmanniens, et de travaux précédents sur $\mathcal{L}$, nous arrivons à montrer une consistance de cet estimateur : une mesure de Thorin $\tilde{\nu}$ qui minimise $\widetilde{\mathcal{L}}(\boldsymbol{x}, \nu)$ satisfait alors (sous certaines conditions techniques) :

$$\|f - f_{\tilde{\nu}}\|_2^2 \xrightarrow[\substack{N \to \infty \\ \boldsymbol{m} \to \infty}]{a.s} 0.$$

De plus, par des arguments venant des flots de gradient dans des espaces métriques, voir Chizat (2021), nous montrons que le flot de gradient associé à la perte (non convexe) $\widetilde{\mathcal{L}}$ converge vers un minimum global. L'ajout d'une pénalisation lasso sur la mesure de Thorin est également possible dans le cadre de ce résultat.

Nous concluons l'étude par des exemples simulés dans des dimensions variées, jusqu'à un modèle multiplicatif en dimension 2000.

Bibliographie

Thorin, O. (1977) On the Infinite Divisibility of the Pareto Distribution. *Scandinavian Actuarial Journal*.

Thorin, O. (1977a) On the Infinite Divisibility of the Lognormal Distribution. *Scandinavian Actuarial Journal*.

Bondesson, L. (1992). *Generalized Gamma Convolutions and Related Classes of Distributions and Densities*. Springer New York.

Bondesson, L. (2009). On Univariate and Bivariate Generalized Gamma Convolutions. *Journal of Statistical Planning and Inference*.

Bondesson, L. (2015). A Class of Probability Distributions That Is Closed with Respect to Addition as Well as Multiplication of Independent Random Variables. *Journal of Theoretical Probability*.

Pérez-Abreu, V. and Stelzer, R. (2012). Infinitely Divisible Multivariate and Matrix Gamma Distributions, *Journal of Multivariate Analysis*.

Chizat, L. (2021). Sparse optimization on measures with over-parameterized gradient descent. *Mathematical Programming*.

Laverny, O., Masiello, E., Maume-Deschamps, V. and Rullière, D. Estimation of multivariate generalized gamma convolutions through Laguerre expansions. *Electronic Journal of Statistics*.

Applications : biologie et santé

Méthodes pour l'inférence post-clustering appliquées à l'expression génique

Benjamin Hivert[1,2,3], Denis Agniel[4,5], Rodolphe Thiébaut[1,2,3,6] & Boris Hejblum[1,2,3]

[1] *Univ. Bordeaux, Inserm Bordeaux Population Health Research Center, SISTM team, UMR 1219, Bordeaux F33076, France*
[2] *INRIA Bordeaux Sud Ouest, SISTM team Talence F-33400, France*
[3] *Vaccine Research Institute, VRI, Hôpital Henri Mondor, Créteil F-94000, France*
[4] *Rand Corporation, Santa Monica, CA 90401, USA*
[5] *Harvard Medical School, Boston, MA 02115, USA*
[6] *CHU Pellegrin, Groupe Hospitalier Pellegrin, Bordeaux F-33076, France*

benjamin.hivert@u-bordeaux.fr, denis.agniel@gmail.com, rodolphe.thiebaut@u-bordeaux.fr, boris.hejblum@u-bordeaux.fr

Résumé. L'analyse des données d'expression génique est souvent organisée autour de deux étapes successives : i) une classification non supervisée utilisant l'ensemble des gènes pour regrouper les unités d'observations (patients, échantillons ou cellules) en sous-groupes distincts et homogènes ; puis ii) l'analyse différentielle se faisant à l'aide de tests d'hypothèse visant à identifier quels gènes, c'est-à-dire quelles variables, sont différentiellement exprimés entre ces sous-groupes. Cependant, cette approche utilisant les même données lors des deux étapes ne permet pas de garantir un bon contrôle de l'erreur de type I à l'étape ii). Nous proposons deux méthodes d'inférence pour tenir compte de l'étape initiale de classification non supervisée lors de l'analyse différentielle et ainsi garantir un contrôle effectif de l'erreur de type I. La première méthode se base sur le concept d'inférence sélective tandis que la seconde repose sur une définition de la séparation de classes faisant uniquement intervenir les concepts d'unimodalité et de multimodalité. Nous avons évalué les performances des deux méthodes grâces à différentes simulations numériques, ainsi que dans une application sur un jeu de données réelles de faible dimension. Les méthodes proposées conduisent à des p-valeurs valides sous l'hypothèse nulle d'absence de différence entre les sous-groupes dans l'expression d'un gène sélectionné, indépendamment de la classification, tout en conservant une bonne puissance statistique. En grande dimension, cette inflation de l'erreur de type I peut-être contre-balancée par la dilution du signal utilisé pour la classification, à condition que les variables soient indépendantes. En revanche, en présence de corrélation (comme c'est le cas en pratique pour l'expression génique), des classes artificielles apparaissent alors que celles-ci ne sont pas séparables. Une adaptation des méthodes à ce contexte de grande dimension est donc nécessaire.

Mots-clés. Classification non supervisée, inférence sélective, analyse circulaire, génomique statistique, données de grande dimension ...

Abstract. The analysis of RNA-seq gene expression data is often organised around two successive steps : i) clustering using all of the genes to group the observation units (patients, cells, etc.) into separate and homogeneous subgroups ; then ii) differential analysis of individual genes using hypothesis testing to identify which genes, i.e. which variables, are differentially expressed between the subgroups. However, several subgroups constructed in i) can actually contain only units coming from the same homogeneous population : clustering will then artificially create differences between those spurious subgroups, leading to false positives in ii). We propose two inference methods to take into account the initial clustering step for differential analysis and thus guarantee an effective control of the type I error. This first method is based on the concept of selective inference while the second one use unimodality and multimodality to describe the separation between clusters. We evaluate the performance of both approaches in extensive numerical simulations as well as in an application to a real, low dimensional dataset. Both proposed methods lead to valid p-values under their null hypothesis of no difference between subgroups in expression at a selected gene independently of the clustering, while maintaining good statistical power. In high dimension, this type I error inflation can be overcome by the dilution of the clustering information, provided that the variables are independent. Yet, in the presence of correlation (as for gene expression), spurious clusters appear, even though they are not separable. An adaptation of the above methods to this high dimensional context is thus necessary.

Keywords. Clustering, selective inference, double-dipping, statistical genomics, high-dimensional data ...

1 Introduction

Le séquençage de l'ARN est une technologie mesurant l'expression génique à différentes échelles qui permet une meilleure compréhension des mécanismes biologiques. L'analyse de ces données d'expression génique repose souvent sur deux grandes étapes : i) une étape de classification non supervisée et ii) une étape d'analyse différentielle. La classification non supervisée a pour but de regrouper les observations afin de former des sous-groupes (des classes) homogènes et séparés. Il est néanmoins toujours possible de construire des classes à l'aide d'une méthode de classification même si toutes les observations proviennent en réalité d'une même population homogène. Dans ce cas les différences entre les classes obtenues ne s'expliquent pas par la présence d'un processus biologique séparant ces groupes d'observations, mais simplement par un mauvais partitionnement des observations. Ce problème est particulièrement dérangeant dans le cadre de l'analyse des données d'expression génique puisque la construction des classes va permettre de formuler les hypothèses de test durant l'étape d'analyse différentielle. Durant cette étape, on va chercher à identifier les gènes séparant les classes (*i.e* les gènes s'exprimant de manière différentielle entre les classes). Cependant, si les deux classes considérées contiennent uni-

quement des observations provenant en réalité d'une même population, alors les gènes qui seront identifiés comme différentiellement exprimés ne le seront pas pour des raisons biologiques, mais simplement comme une conséquence de la classification non supervisée qui aura artificiellement forcée des différences entre les observations qu'elles contiennent. Ainsi, les méthodes d'inférence classiques ne contrôlent plus l'erreur de type I en particulier puisque que les hypothèses de tests ne sont pas construites a-priori, mais à l'aide des mêmes données que celles utilisées pour le test d'hypothèse. Ce problème d'inférence post-classification pour les données RNA-seq est devenu l'un des problèmes majeurs liés à leur analyse [Lähnemann et al., 2020] et fait l'objet de développements méthodologiques récents [Zhang et al., 2019, Gao et al., 2020].

Nous avons donc pour objectif de proposer de nouvelles méthodes d'inférence post-classification permettant d'identifier les gènes différentiellement exprimés entre des classes construites à partir des données à l'aide d'une méthode de classification non supervisée. Pour cela, nous proposons deux méthodes d'inférence permettant de tenir compte de l'étape de classification non supervisée dans l'analyse différentielle : a) nous étendons le travail de Gao et al. [2020] au cas uni-variable qui s'appuie sur le concept d'inférence sélective [Tibshirani et al., 2016] où l'on conditionne sur la classification dans le test statistique ; b) avec une définition plus restrictive d'un sous-groupe (utilisant les notions d'unimodalité et de multimodalité pour caractériser respectivement l'homogénéité au sein d'un sous-groupe et la séparation entre deux sous-groupes) nous étudions la séparabilité de deux classes selon une variable d'intérêt par un test de multimodalité.

2 Méthodes

Soit $\boldsymbol{X}$ une matrice $n \times p$. Sur cette matrice $\boldsymbol{X}$, on applique une méthode de classification non supervisée C afin de construire $C(\boldsymbol{X})$, une partition des n observations en K classes disjointes $C_1, \ldots, C_K$. On souhaite alors étudier la capacité d'une variable $\boldsymbol{X}_g \in \mathbb{R}^n$ de $\boldsymbol{X}$ à séparer deux classes estimées C_k and $C_{k'}$ grâce à la méthode de classification C.

2.1 Inférence sélective post-classification

Une première solution permettant d'étudier la séparation de C_k et $C_{k'}$ sur $\boldsymbol{X}_g$ est de tester une différence de moyenne entre les individus de C_k et ceux de $C_{k'}$ sur $\boldsymbol{X}_g$. En supposant que la variable $\boldsymbol{X}_g$ suit une distribution gaussienne : $\boldsymbol{X}_g \sim \mathcal{N}_n(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I}_n)$, il est alors possible d'étudier la séparation de C_k et $C_{k'}$ sur $\boldsymbol{X}_g$ en considérant les hypothèses :

$$\mathcal{H}_0 : \overline{\mu}_g^{C_k} = \overline{\mu}_g^{C_{k'}} \qquad \text{vs} \qquad \mathcal{H}_1 : \overline{\mu}_g^{C_k} \neq \overline{\mu}_g^{C_{k'}} \tag{1}$$

où $\overline{\mu}_g^{C_l} = \frac{1}{|C_l|} \sum_{i \in C_l} \mu_{gi}$ pour $l \in \{k, k'\}$.

L'hypothèse du test (*i.e* la séparation des deux classes) étant établie à partir des données elles mêmes et non a priori, un moyen d'obtenir des tests statistiques valides est

d'utiliser les concepts d'inférence sélective [Fithian et al., 2014, Tibshirani et al., 2016]. Selon ces concepts, puisque les hypothèses ont été construites en utilisant les données, il est nécessaire de conditionner sur cette utilisation des données dans la définition de la p-valeur. Ainsi, nous définissions notre p-valeur par :

$$p_g^{C_k,C_{k'}} \equiv \mathbb{P}_{H_0}\left(\left|\overline{X}_g^{C_k} - \overline{X}_g^{C_{k'}}\right| > \left|\overline{x}_g^{C_k} - \overline{x}_g^{C_{k'}}\right| \mid C_k, C_{k'} \in C(\boldsymbol{X})\right) \tag{2}$$

En particulier, nous conditionnons sur le fait que les deux classes d'intérêt ont été estimées (à l'aide de la méthode de classification non supervisée C) via les mêmes données que celles servant à faire le test. En utilisant les travaux de Gao et al. [2020], il est possible de ré-écrire cette p-valeur (2) comme :

$$p_g^{C_k,C_{k'}} = \mathbb{P}_{H_0}\left(|\phi_g| > \left|\overline{x}_g^{C_k} - \overline{x}_g^{C_{k'}}\right| \mid \phi_g \in S_g\right) \tag{3}$$

où $S_g = \{\phi_g : C_k, C_{k'} \in C\left(\boldsymbol{X}\left(\phi_g\right)\right)\}$ et $\phi_g = \left|\overline{X}_g^{C_k} - \overline{X}_g^{C_{k'}}\right| \overset{H_0}{\sim} \mathcal{N}\left(0, \tau_g^2\right)$. $\boldsymbol{X}(\phi_g)$ est une version perturbée des données selon la statistique de test ϕ_g où seulement la $g^{\text{ème}}$ variable est impactée par cette perturbation. Cette p-valeur définie en (3) est ensuite calculée par une approche de Monte-Carlo. La variance τ_g^2 de la statistique de test ϕ_g est le seul paramètre à estimer et cette estimation se fait de manière à respecter l'hypothèse nulle d'absence de séparation entre les deux classes.

Ce conditionnement sur l'évènement de clustering peut aboutir à une perte de puissance statistique dans certains cas. Pour pallier ce problème, nous proposons une extension de ce test d'inférence sélective basée sur l'agrégation des p-valeurs issues du test sur l'ensemble des clusters adjacents entre C_k et $C_{k'}$ sur $\boldsymbol{X}_g$.

2.2 Test de multimodalité

Pour s'affranchir des limites du test d'inférence sélective (temps de calculs importants dus à l'estimation de Monte-Carlo, perte de puissance statistique dans certains cas), nous proposons également un test de multimodalité. Les notions d'unimodalité et de multimodalité ont déjà été utilisées pour caractériser respectivement la distribution de données avec et sans une structure en classes [Kalogeratos and Likas, 2012, Siffer et al., 2018]. Nous proposons d'utiliser ces deux notions afin d'évaluer la présence d'un continuum dans la distribution de $\boldsymbol{X}_g$ entre C_k et $C_{k'}$. En effet, si un tel continuum n'existe pas pour aller de C_k à $C_{k'}$ (en passant par toutes les adjacentes se trouvant entre ces deux), alors cela signifie que C_k et $C_{k'}$ sont séparées sur $\boldsymbol{X}_g$.

Tester la présence de ce continuum entre C_k et $C_{k'}$ sur $\boldsymbol{X}_g$ revient alors à tester l'unimodalité de cette variable en ne considérant que les individus se trouvant dans l'ensemble des classes adjacentes entre C_k et $C_{k'}$. Le test d'unimodalité que nous utilisons est le DipTest [Hartigan et al., 1985], un test non-paramétrique se basant sur la fonction de répartition des données. En pratique, le DipTest compare la fonction de répartition

empirique des données à celle de la distribution uniforme, considérée comme la pire des distributions unimodales d'après la statistique Dip.

Cette seconde approche est en fait très proche de celle basée sur l'inférence sélective. En effet, les perturbations des données sur lesquelles repose notre première approche sont un moyen d'étudier la possibilité à créer un continuum entre C_k et $C_{k'}$ sur $\boldsymbol{X}_g$. Avec cette seconde approche, nous testons directement la présence de ce continuum dans les données à l'aide d'un test de multimodalité.

3 Résultats

3.1 Simulations numériques

Nous avons étudié le comportement des trois tests que nous proposons (le test d'inférence sélective, son extension par combinaison de p-valeurs ainsi que le test de multimodalité) en comparaison avec un test classique, le t-test, dans le cadre de la recherche des variables séparant des pairs de classes. Pour cela, nous avons généré des données selon deux scénario : i) un scénario sous $\mathcal{H}_0$ dans lequel les données sont simulées comme provenant toutes d'une même distribution gaussienne et ii) un scénario sous $\mathcal{H}_1$ dans lequel les données sont simulées telles qu'il y ait des vrais sous-groupes (Figure 1A). Dans les deux cas, nous avons estimé un nombre de classes correspondant au vrai nombre de sous-groupes présents sous $\mathcal{H}_1$. Alors que sous $\mathcal{H}_1$, l'algorithme de classification non supervisée ne fait que découvrir la structure des données en identifiant les vrais sous-groupes d'observations présents dans les données, sous $\mathcal{H}_0$, il force des différences entre des sous-groupes d'observations afin de construire des classes. Cela se traduit par des p-valeurs pour le t-test anormalement significatives, témoignant d'un mauvais contrôle de l'erreur de type I. Nos tests parviennent bien à corriger ce problème garantissant un contrôle effectif de l'erreur de type I en cas de mauvais partitionnement des données. Sous $\mathcal{H}_1$, nos tests restent suffisamment puissants pour détecter la séparation entre deux classes (Figure 1B).

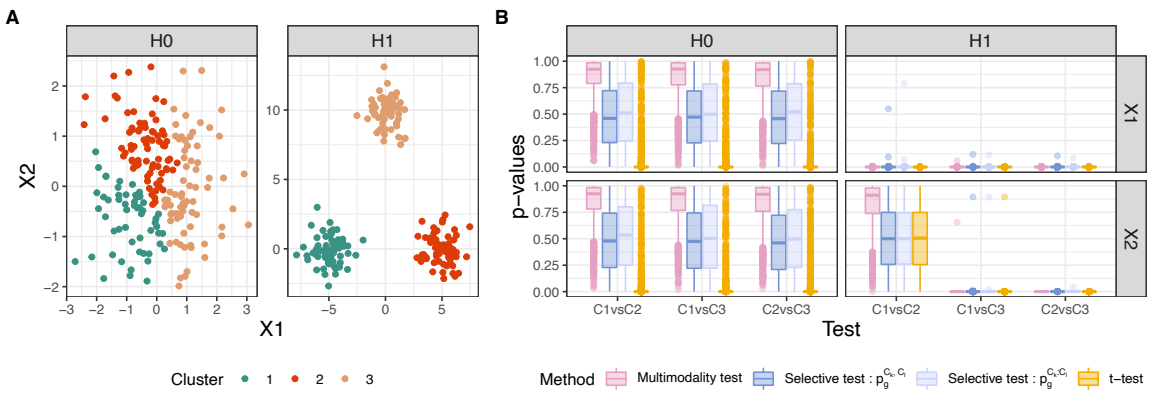


FIGURE 1 – Distributions des p-valeurs des tests proposés sous deux scénario différents

3.2 Grande dimension

Le problème en grande dimension s'avère différent. En effet, l'impact de la classification non supervisée sur la distribution des variables n'est pas le même qu'en petite dimension. Pour commencer, la grande quantité de bruit par rapport à un vrai signal relatif à une structure en classes rend la tâche de classification elle-même plus compliquée. En l'absence de réels sous-groupes d'observations dans les données, seulement un faible nombre de variables seront impactées par l'algorithme de classification, et au vue du grand nombre de variables observées. La correction pour la multiplicité des tests suffit alors à garantir un contrôle effectif de l'erreur de type I des méthodes d'inférence classiques. Cependant, une corrélation entre les variables peut amplifier le nombre de variables touchées par l'algorithme de clustering : le problème d'inférence post-classification demeure alors toujours présent.

Références

David Lähnemann, Johannes Köster, Ewa Szczurek, Davis J McCarthy, Stephanie C Hicks, Mark D Robinson, Catalina A Vallejos, Kieran R Campbell, Niko Beerenwinkel, Ahmed Mahfouz, et al. Eleven grand challenges in single-cell data science. *Genome biology*, 21(1) :1–35, 2020.

Jesse M Zhang, Govinda M Kamath, and N Tse David. Valid post-clustering differential analysis for single-cell rna-seq. *Cell systems*, 9(4) :383–392, 2019.

Lucy L Gao, Jacob Bien, and Daniela Witten. Selective inference for hierarchical clustering. *arXiv preprint arXiv :2012.02936*, 2020.

Ryan J Tibshirani, Jonathan Taylor, Richard Lockhart, and Robert Tibshirani. Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association*, 111(514) :600–620, 2016.

William Fithian, Dennis Sun, and Jonathan Taylor. Optimal inference after model selection. *arXiv preprint arXiv :1410.2597*, 2014.

Argyris Kalogeratos and Aristidis Likas. Dip-means : an incremental clustering method for estimating the number of clusters. *Advances in neural information processing systems*, 25 :2393–2401, 2012.

Alban Siffer, Pierre-Alain Fouque, Alexandre Termier, and Christine Largouët. Are your data gathered ? In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2210–2218, 2018.

John A Hartigan, Pamela M Hartigan, et al. The dip test of unimodality. *Annals of statistics*, 13(1) :70–84, 1985.

Prédiction génomique en grande dimension : densité de marqueurs nécessaire pour une prédiction fiable en sélection génomique

Charles-Elie Rabier [1] & Simona Grusea [2]

[1] *IMAG, Université de Montpellier, CNRS, France*
charles-elie.rabier@umontpellier.fr
[2] *Institut de Mathématiques de Toulouse, Université de Toulouse, INSA de Toulouse, France*
grusea@insa-toulouse.fr

Résumé. La sélection génomique (GS) consiste à sélectionner des individus sur la base de prédictions génomiques effectuées à l'aide d'une grande densité de marqueurs. Une question d'importance en GS est de déterminer le nombre de marqueurs nécessaires pour une prédiction fiable. Pour ce faire, nous introduisons de nouveaux proxies pour la précision de la prédiction. Ces proxies sont appropriés dans le cadre d'une carte génétique discrète, où il est fréquent d'observer du déséquibre de liaison incomplet, i.e. la situation où les allèles à l'emplacement d'un gène et à l'emplacement d'un marqueur à proximité, diffèrent. De plus, nos proxies suggérés sont utiles pour concevoir des puces SNPs basées sur une densité modérée de marqueurs. Nous analysons des données de riz des Philippines, et nous nous focalisons sur la date de floraison recueillie durant la saison sèche 2012. En utilisant différentes densités de marqueurs, nous montrons qu'au moins 1553 marqueurs sont nécessaires afin d'obtenir une prédiction fiable et de mettre en place la GS. Déterminer le nombre optimal de marqueurs est crucial afin d'optimiser le programme de sélection.

Mots-clés. Statistique en grande dimension, Prédiction, Régression Ridge, Sélection génomique, Déséquilibre de liaison incomplet, Données de riz

Abstract. Genomic selection (GS) consists in predicting breeding values of selection candidates, using a large number of genetic markers. An important question in GS is to determine the number of markers required for a good prediction. For this purpose, we introduce new proxies for the accuracy of the prediction. These proxies are suitable under sparse genetic map where it is likely to observe some imperfect linkage disequilibrium, i.e. the situation where the alleles at a gene location and at a marker located nearby vary. Moreover, our suggested proxies are helpful for designing cost-effective SNP chips based on a moderate density of markers. We analyze rice data from Los Banos, Philippines and focus on the flowering time collected during the dry season 2012. Using different densities of markers, we show that at least 1553 markers are required to implement GS. Finding the optimal number of markers is crucial in order to optimize the breeding program.

Keywords. High Dimensional Data Analysis, Prediction, Ridge Regression, Genomic Selection, Imperfect Linkage Disequilibrium, Rice Data Analysis

Cet acte de congrès se veut un résumé de la publication Rabier et Grusea (2021).

1 Introduction

La sélection génomique (GS) est une méthode extrêmement populaire en génétique (Meuwissen et al., 2001), consistant à sélectionner des individus sur la base de prédictions génomiques. Ces prédictions, effectuées à l'aide d'une grande densité de marqueurs génétiques couvrant le génome, se doivent d'être précises afin de pouvoir sélectionner les meilleurs candidats pour le programme de sélection. La GS a été tout d'abord appliquée aux animaux et est désormais largement employée chez les plantes. On peut citer par exemple, les études sur la pomme (Muranty et al., 2015), l'eucalyptus (Tan et al., 2017), les poires japonaises (Minamikawa et al., 2018), les fraises (Gezan et al., 2017), la banane (Nyine et al., 2018) et le café (Ferrao et al., 2018).

Rappelons que l'on nomme QTLs, les loci sur le génome codant pour un caractère (i.e. trait ou phénotype) quantitatif. D'un point de vue méthodologique, la GS repose sur l'espoir que chaque QTL sera étroitement corrélé avec au moins un des marqueurs, en raison de la grande densité de marqueurs. En génétique, cette corrélation entre un QTL et un marqueur est nommée Déséquilibre de Liaison (DL) : cela correspond à la non indépendance des allèles à 2 loci différents (cf. Durett, 2008). Contrairement aux études d'associations à l'échelle du génome (GWAS) qui recherchent des QTLs, l'objectif de la GS est d'effectuer des prédictions à l'aide d'un grand nombre de marqueurs, sans avoir à détecter les QTLs. L'avantage des prédictions génomiques en GS, par rapport aux prédictions basées sur les loci détectés par GWAS, réside dans le fait que les QTLs à petits effets sont très difficiles à détecter pour la plupart des traits, qualifiés de traits complexes car gouvernés par beaucoup de QTLs à petits effets.

De nombreux facteurs sont connus pour influencer la qualité des prédictions en GS : la taille du jeu d'entrainement sur lequel le modèle est appris, l'héritabilité du caractère liée au rapport signal/bruit, la densité de marqueurs, le DL, le lien entre jeux d'entrainement et de validations ... Dans Rabier et al (2018), nous nous étions interessés uniquement au DL complet : les QTLs étaient localisés uniquement sur quelques marqueurs. Cependant, lorsque les QTLs ne coïncident pas avec les positions des marqueurs, les allèles au QTL et au marqueur à proximité, diffèrent. On parle dès lors de DL incomplet. Ainsi, l'objectif de ce travail est d'étudier la prédiction génomique dans le cadre d'un DL incomplet.

Une thématique de recherche connexe est la détermination du nombre de marqueurs requis pour mettre en place la GS. Dans leur étude sur le maïs, Zhang et al (2015) ont montré que la prédiction d'un trait complexe nécessitait un grand nombre de marqueurs génétiques (environ 58000 marqueurs par GBS), alors que 200 marqueurs étaient suffisants pour prédire un trait simple. D'autre part, dans une étude sur le café, Ferrao et al (2019) ont montré que les prédictions basées sur 4000 marqueurs donnaient les mêmes résultats que celles basées sur 35000 marqueurs. Pour finir, en aquaculture, Kriaridou et al. (2020)

ont obtenu des résultats similaires en comparant 1000 marqueurs et une haute densité de marqueurs. Un sujet également proche s'avère la conception de puces LD (e.g. Bolormaa et al. 2015, Corbin et al., 2014). Notre travail pourrait s'avérer utile pour concevoir des puces LD qui réduisent les coûts dus au génotypage en considérant une densité modérée de marqueurs, contrairement aux puces haute densité (Wu et al, 2016).

1.1 Modèle linéaire causal en présence de DL incomplet

Introduisons tout d'abord le modèle statistique causal. Nous nous intéressons à un trait quantitatif (i.e. un phénotype) qui est observé sur un échantillon d'entraînement de taille n. $Y_1, \ldots, Y_n$ désignent les variables aléatoires relatives à ce trait quantitatif. On considère qu'il existe m QTLs sur le génome, ayant une influence sur le trait quantitatif.

Pour $1 \leq j \leq m$, on note $\beta_j^\star$ le $j$ème effet QTL. $X^\star$ désigne la matrice de taille $n \times m$ contenant les allèles aux QTLs des n individus. La $i$ème ligne de $X^\star$, qui s'écrit $\boldsymbol{x_i}' = \left(X_{i,1}^\star, ..., X_{i,p}^\star\right)$, correspond au $i$ème individu, en utilisant $'$ pour symboliser la transposée.

On considère le modèle linéaire causal suivant pour le trait quantitatif :

$$\boldsymbol{Y} = X^\star \boldsymbol{\beta^\star} + \boldsymbol{\varepsilon}, \tag{1}$$

où $\boldsymbol{Y} = (Y_1, ..., Y_n)'$, $\boldsymbol{\beta^\star} = (\beta_1^\star, ..., \beta_m^\star)'$, $\boldsymbol{\varepsilon} \sim N(0, \sigma_e^2 I_n)$, I_n est la matrice identité de taille n et σ_e^2 représente la variance environnementale.

On suppose que $X^\star$ et $\boldsymbol{\varepsilon}$ sont indépendants. De plus, l'information génomique est disponible à p marqueurs, avec $p > n$ (contexte de grande dimension). A noter que les QTLs **ne coïncident pas forcément** avec les marqueurs (cadre du DL incomplet).

Soit X la matrice de taille $n \times p$ contenant l'information génomique aux p marqueurs pour les n individus. Rappelons que dans notre étude précédente, nous avions considéré le DL complet, où les QTLs se trouvent parmi les marqueurs. Dans ce cas, le modèle considéré s'écrivait :

$$\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2}$$

avec $1 \leq j \leq p$, $\beta_j = 0$ si le marqueur n'est pas un QTL, et $\beta_j \neq 0$ si le jème marqueur est un QTL.

Remark 1 *A noter que le DL complet peut être vu comme un cas particulier du DL incomplet. C'est le cas lorsque les m QTLs sont localisés sur quelques marqueurs et dans ce cas, $\boldsymbol{\beta}$ représente le vecteur sparse de taille p, contenant les composantes de $\boldsymbol{\beta^\star}$.*

1.2 Individu Test (TST)

On suppose que pour un individu aléatoire supplémentaire, dit individu test (TST), noté *new*, on dispose de son information génomique (i.e. ses allèles aux différents marqueurs)

mais pas de son phénotype. Soit $\boldsymbol{x}^{\star}_{\boldsymbol{new}}$ le vecteur colonne contenant les allèles aux m QTLs de l'individu new. Le trait quantitatif Y_{new} vérifie :

$$Y_{new} = \boldsymbol{x}_{\boldsymbol{new}}{}^{\star\prime}\, \boldsymbol{\beta}^{\star} + \varepsilon_{new},$$

où $\varepsilon_{new} \sim N(0, \sigma_e^2)$, et $\boldsymbol{x}^{\star}_{\boldsymbol{new}}$, ε_{new} et $\boldsymbol{\varepsilon}$ sont tous indépendants.

De plus, $\boldsymbol{x}_{\boldsymbol{new}}$ fait référence au génome aléatoire aux marqueurs. A noter que $\boldsymbol{x}_{\boldsymbol{new}}$ et $\boldsymbol{x}^{\star}_{\boldsymbol{new}}$ sont corrélés en raison de la liaison génétique due à la taille fixe du génome.

Dans ce qui suit, nous assumerons que $\boldsymbol{Y}$, Y_{new}, $\boldsymbol{x}_{\boldsymbol{new}}$, $\boldsymbol{x}^{\star}_{\boldsymbol{new}}$, les colonnes de X et les colonnes de $X^{\star}$ sont centrés.

1.3 Modèle de prédiction et précision

Construisons un estimateur $\hat{Y}_{new}$ pour la valeur phénotypique de l'individu new, à l'aide de l'information génomique disponible pour les n individus d'entrainement.

Prédicteur basé sur la régression Ridge
Nous utiliserons comme modèle instrumental, la régression Ridge qui peut être vue comme une régression Bayésienne. L'estimateur de la régression Ridge pour les effets marqueurs $\boldsymbol{\beta}$ (Tikhonov 1963, Hoerl et Kennard 1970) s'écrit :

$$\hat{\boldsymbol{\beta}} = \left(X'X + \lambda I_p\right)^{-1} X'\boldsymbol{Y} = X'V^{-1}\boldsymbol{Y}, \tag{3}$$

où $V = XX' + \lambda I_n$, et où λ et I_p désignent respectivement le paramètre de régularisation, et la matrice identité de taille $p \times p$.

A noter que l'estimateur Ridge est basé sur l'information génomique aux marqueurs et est approprié dans le cadre de la grande dimension (i.e. $p > n$, cf. e.g. Shao and Deng 2012, et Bühlmann 2013). La prédicteur s'écrit $\hat{Y}_{new} = \boldsymbol{x}_{\boldsymbol{new}}{}'\hat{\boldsymbol{\beta}}$.

Critère de précision
En génétique, la capacité prédictive est quantifiée selon la *précision phénotypique*, ρ_{ph} (e.g. Visscher et al., 2010) ou la *précision génotypique*, ρ_g (e.g. Daetwyler et al., 2008 et 2010):

$$\rho_{ph} := \frac{\mathrm{Cov}\left(\hat{Y}_{new},\, Y_{new}\right)}{\sqrt{\mathrm{Var}\left(\hat{Y}_{new}\right)\mathrm{Var}\left(Y_{new}\right)}}\,,\ \rho_g := \frac{\mathrm{Cov}\left(\hat{Y}_{new},\, \boldsymbol{x}_{\boldsymbol{new}}{}^{\star\prime}\boldsymbol{\beta}^{\star}\right)}{\sqrt{\mathrm{Var}\left(\hat{Y}_{new}\right)\mathrm{Var}\left(\boldsymbol{x}_{\boldsymbol{new}}{}^{\star\prime}\boldsymbol{\beta}^{\star}\right)}}\,. \tag{4}$$

En particulier, cette corrélation est une composante essentielle dans l'équation du sélectionneur (cf. Lynch and Walsh, 1998). A noter que dans Rabier et Grusea (2021), nous nous intéressons également à l'erreur quadratique, plus commune dans la communauté statistique.

Héritabilité
Comme $\boldsymbol{x}^{\star}_{\boldsymbol{new}}$, ε_{new} et $\boldsymbol{\varepsilon}$ sont tous supposés indépendants, nous avons la relation $\rho_{ph}/\rho_g = h$, où h est la racine de carré de l'héritabilité du trait :

$$h^2 := \frac{\mathrm{Var}\left(\boldsymbol{x_{new}}^{\star\prime}\,\boldsymbol{\beta}^{\star}\right)}{\mathrm{Var}\left(Y_{new}\right)}\,. \tag{5}$$

Avec la notation $\sigma^2_G = \mathrm{Var}\left(\boldsymbol{x_{new}}^{\star\prime}\,\boldsymbol{\beta}^{\star}\right)$, nous avons la relation $h^2 = \sigma^2_G/\left(\sigma^2_G + \sigma^2_e\right)$.

1.4 Contributions

Dans Rabier et Grusea (2021), nous montrons que la projection de la fonction de régression $X^{\star}\boldsymbol{\beta}^{\star}$ sur le sous-espace linéaire engendré par les colonnes de X, noté $\mathrm{CS}(X)$, est un élément clé pour la précision génotypique. A partir de ces résultats théoriques, il est possible de retouver les résultats obtenus sous le DL complet : le facteur clé devient la projection du signal $\boldsymbol{\beta}$ sur le sous-espace linéaire engendré par les lignes de X, noté $\mathrm{RS}(X)$. Assumant que le signal $\boldsymbol{\beta}^{\star}$ est réparti uniformément sur chaque sous-espace, nous montrons également au travers d'un lemme que la précision optimale (oracle) est atteinte dès que la limite d'un facteur de perte, dû au DL incomplet, est égal à 0.

Par la suite, nous introduisons un prédicteur modifié, inspiré de la régression Ridge, et qui améliore la qualité de la prédiction. Il repose sur la projection de $\boldsymbol{Y}$ sur un sous-espace bien choisi de $\mathrm{CS}(X)$. Nous étudions la précision de ce prédicteur : comme attendu, en présence de DL incomplet, la qualité de la prédiction dépend de la projection de la fonction de régression $X^{\star}\boldsymbol{\beta}^{\star}$ sur le sous-espace choisi. Nous présentons des résultats qui permettent de comparer la précision liée à l'estimateur Ridge et celle liée au prédicteur modifié, dans un contexte de DL incomplet.

D'autre part, nous proposons de nouveaux proxies reposant sur nos résultats théoriques. Les performances de ces proxies seront illustrés sur des données simulées et réelles. Nos proxies dépendent uniquement des phénotypes et des marqueurs des individus d'entrainement, mais prennent en compte la sparsité de la carte génétique des individus TST. Ainsi, afin de construire ces proxies, nous nous attaquons à un nouveau sujet en GS : la précision de la prédiction des individus TST lorsque la carte génétique des individus d'entrainement diffère de celle des TST. En particulier, nous suggérons de considérer une carte plus dense pour les individus d'entrainement que pour les TST : la carte dense nous aidera à estimer les paramètres de nuisance $X^{\star}$ et $\boldsymbol{\beta}^{\star}$ requis pour calculer nos proxies. Ce concept repose sur l'espoir que les QTLs se trouveront en DL complet sous la carte dense d'entrainement, ce qui n'est pas le cas pour la carte TST (DL incomplet). Contrairement à notre étude portant sur le DL complet (Rabier et al, 2018) où l'Adaptive LASSO (Zou, 2006) était le meilleur substitut pour $\boldsymbol{\beta}$, nous avons trouvé ici que le LASSO (Tibshirani, 1996) s'avérait le meilleur substitut pour $\boldsymbol{\beta}^{\star}$ lorsque la carte discrète TST était considérée.

Pour finir, nous analysons des données de riz des Philippines (Spindel et al., 2015), et nous nous focalisons sur la date de floraison recueillie durant la saison sèche 2012.

En considérant différentes densités de marqueurs et nos nouveaux proxies appropriés en présence de DL incomplet, nous montrons que les généticiens peuvent évaluer la précision des futures prédictions. Si la précision n'est satisfaisante, nous leur conseillons de re-densifier leur carte génétique à l'aide de marqueurs supplémentaires, afin d'améliorer la fiabilité des prédictions.

Bibliographie

[1] Bühlmann, P. (2013). Statistical significance in high-dimensional linear models. *Bernoulli*, 19(4), 1212-1242.

[2] Ferrao, L.F.V., Ferrao, R.G., Ferrao, M.A.G., Fonseca, A., Carbonetto, P., Stephens, M., and Garcia, A.A.F. (2019). Accurate genomic prediction of Coffea canephora in multiple environments using whole-genome statistical models. *Heredity*, 122(3), 261-275.

[3] Gezan, S.A., Osorio, L.F., Verma, S., and Whitaker, V.M. (2017). An experimental validation of genomic selection in octoploid strawberry. *Horticulture research*, (4), 16070.

[4] Hoerl AE, and Kennard RW (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1), 55-67.

[5] Minamikawa, M.F, Takada, N., Terakami, S., Saito, T., Onogi, A., Kajiya-Kanegae, H. ..., and Iwata, H. Genome-wide association study and genomic prediction using parental and breeding populations of Japanese pear (Pyrus pyrifolia Nakai). *Scientific reports*, 8(1), 11994.

[6] Rabier, C.E., Grusea, S. (2021), Prediction in high-dimensional linear models and application to genomic selection under imperfect linkage disequilibrium. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 70(4), 1001-1026.

[7] Rabier, C. E., Mangin, B., Grusea, S. (2019). On the accuracy in high-dimensional linear models and its application to genomic selection. *Scandinavian journal of statistics*, 46(1), 289-313.

[8] Shao, J., Deng, X. (2012). Estimation in high-dimensional linear models with deterministic design matrices. *Annals of statistics*, 40(2), 812-831.

[9] Spindel, J., Begum, H., Akdemir, D., Virk, P., Collard, B., and Redoña, E., et al (2015). Genomic Selection and Association Mapping in rice (Oryza sativa): Effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of rice genomic selection in elite, tropical rice breeding lines. *PLoS Genetics*, 11(2), e1004982.

[10] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, 267-288.

[11] Tikhonov, A.N. (1963). On the solution of ill-posed problems and the method of regularization. *Dokl. Akad. Nauk.*, (151), 501-504..

[12] Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476), 1418-1429.

Identification of prognostic and predictive biomarkers in high-dimensional data

Wencan Zhu [1,3] & Céline Lévy-Leduc [2] & Nils Ternès [3]

[1] *UMR MIA-Paris, AgroParisTech, INRAE - Université Paris-Saclay, 75005 Paris, France, wencan.zhu@agroparistech.fr*

[2] *UMR MIA-Paris, AgroParisTech, INRAE - Université Paris-Saclay, 75005 Paris, France, celine.levy-leduc@agroparistech.fr*

[3] *Biostatistics and Programming department, Sanofi R&D, 91380 Chilly Mazarin, France, nils.ternes@sanofi.com*

Abstract. In clinical trials, identification of prognostic and predictive biomarkers is essential to precision medicine. Prognostic biomarkers can be useful for the prevention of the occurrence of the disease, and predictive biomarkers can be used to identify patients with potential benefit from the treatment. Previous researches were mainly focused on clinical characteristics, and the use of genomic data in such an area is hardly studied. We propose a new approach called PPLasso (Prognostic Predictive Lasso) integrating prognostic and predictive effects into one statistical model as a variable selection problem in an ANCOVA (Analysis of Covariance) type model. PPLasso also takes into account the correlations between biomarkers that can alter the biomarker selection accuracy. Through numerical experiments, we show that PPLasso outperforms the traditional Lasso approach on both prognostic and predictive biomarker identification in various scenarios.

Keywords. Variable selection, Precision medicine, high-dimensional data, predictive biomarker

1 Method

Let $\mathbf{y}$ be a continuous response and t_1, t_2 two treatments. Let also $\mathbf{X}_1$ (resp. $\mathbf{X}_2$) denote the design matrix for the n_1 (resp. n_2) patients with treatment t_1 (resp. t_2), each containing measurements on p candidate biomarkers:

$$\mathbf{X}_1 = \begin{bmatrix} X_{11}^1 & X_{11}^2 & \dots & X_{11}^p \\ X_{12}^1 & X_{12}^2 & \dots & X_{12}^p \\ \dots & & & \\ X_{1n_1}^1 & X_{1n_1}^2 & \dots & X_{1n_1}^p \end{bmatrix}, \mathbf{X}_2 = \begin{bmatrix} X_{21}^1 & X_{21}^2 & \dots & X_{21}^p \\ X_{22}^1 & X_{22}^2 & \dots & X_{22}^p \\ \dots & & & \\ X_{2n_2}^1 & X_{2n_2}^2 & \dots & X_{2n_2}^p \end{bmatrix}. \quad (1)$$

To take into account the potential correlation that may exist between the biomarkers in the different treatments, we shall assume that the rows of $\mathbf{X}_1$ (resp. $\mathbf{X}_2$) are independent centered Gaussian random vectors with a covariance matrice equal to $\mathbf{\Sigma_1}$ (resp. $\mathbf{\Sigma_2}$).

To model the link that exists between $\mathbf{y}$ and the different types of biomarkers we propose using the following model:

$$\mathbf{y} = \begin{pmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ y_{22} \\ \vdots \\ y_{2n_2} \end{pmatrix} = \mathbf{X} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \beta_{11} \\ \beta_{12} \\ \vdots \\ \beta_{1p} \\ \beta_{21} \\ \beta_{22} \\ \vdots \\ \beta_{2p} \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \vdots \\ \epsilon_{1n_1} \\ \epsilon_{21} \\ \epsilon_{22} \\ \vdots \\ \epsilon_{2n_2} \end{pmatrix}, \tag{2}$$

where $(y_{i1}, \ldots, y_{in_i})$ corresponds to the response of patients with treatment t_i, i being equal to 1 or 2,

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & X_{11}^1 & X_{11}^2 & \ldots & X_{11}^p & 0 & 0 & \ldots & 0 \\ 1 & 0 & X_{12}^1 & X_{12}^2 & \ldots & X_{12}^p & 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & & & & \\ 1 & 0 & X_{1n_1}^1 & X_{1n_1}^2 & \ldots & X_{1n_1}^p & 0 & 0 & \ldots & 0 \\ 0 & 1 & 0 & 0 & \ldots & 0 & X_{21}^1 & X_{21}^2 & \ldots & X_{21}^p \\ 0 & 1 & 0 & 0 & \ldots & 0 & X_{22}^1 & X_{22}^2 & \ldots & X_{22}^p \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & 1 & 0 & 0 & \ldots & 0 & X_{2n_2}^1 & X_{2n_2}^2 & \ldots & X_{2n_2}^p \end{bmatrix},$$

with α_1 (resp. α_2) corresponding to the effects of treatment t_1 (resp. t_2). Moreover, $\boldsymbol{\beta}_1 = (\beta_{11}, \beta_{12}, \ldots, \beta_{1p})'$ (resp. $\boldsymbol{\beta}_2 = (\beta_{21}, \beta_{22}, \ldots, \beta_{2p})'$) are the coefficients associated to each of the p biomarkers in treatment t_1 (resp. t_2) group, $'$ denoting the matrix transposition and $\epsilon_{11}, \ldots, \epsilon_{2n_2}$ are standard independent Gaussian random variables independent of $\mathbf{X}_1$ and $\mathbf{X}_2$. When t_1 stands for the standard treatment or placebo, prognostic (resp. predictive) biomarkers are defined as those having non-zero coefficients in $\boldsymbol{\beta}_1$ (resp. in $\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1$) and non prognostic (resp. non predictive) biomarkers correspond to the indices having null coefficients in $\boldsymbol{\beta}_1$ (resp. in $\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1$). Model (2) can be written as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\gamma} + \boldsymbol{\epsilon}, \tag{3}$$

with $\boldsymbol{\gamma} = (\alpha_1, \alpha_2, \boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$.

To estimate $\boldsymbol{\gamma}$ using a sparsity enforcing constraint, we consider a first estimator of $\boldsymbol{\gamma}$ obtained by minimizing the following criterion with respect to $\boldsymbol{\gamma}$:

$$\frac{1}{2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\gamma}\|_2^2 + \lambda_1 \left\| \begin{bmatrix} \mathbf{0}_{p,1} & \mathbf{0}_{p,1} & D_1 \\ \mathbf{0}_{p,1} & \mathbf{0}_{p,1} & \frac{\lambda_2}{\lambda_1} D_2 \end{bmatrix} \boldsymbol{\gamma} \right\|_1, \tag{4}$$

where $D_1 = [\mathrm{Id}_p, \mathbf{0}_{p,p}]$ and $D_2 = [-\mathrm{Id}_p, \mathrm{Id}_p]$, Id_p denoting the identity matrix of size p and $\mathbf{0}_{i,j}$ denoting a matrix having i rows and j columns and containing only zeros.

However, since the inconsistency of Lasso biomarker selection is originated from the correlations between the biomarkers, we propose to remove the correlation by "whitening" the matrix $\mathbf{X}$. More precisely, we consider $\widetilde{\mathbf{X}} = \mathbf{X}\boldsymbol{\Sigma}^{-1/2}$, where

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \boldsymbol{\Sigma}_1 & 0 \\ 0 & 0 & 0 & \boldsymbol{\Sigma}_2 \end{bmatrix} \tag{5}$$

and define $\boldsymbol{\Sigma}^{-1/2}$ by replacing in (5) $\boldsymbol{\Sigma}_i$ by $\boldsymbol{\Sigma}_i^{-1/2}$, where $\boldsymbol{\Sigma}_i^{-1/2} = \mathbf{U}_i\mathbf{D}_i^{-1/2}\mathbf{U}_i^T$, $\mathbf{U}_i$ and $\mathbf{D}_i$ being the matrices involved in the spectral decomposition of $\boldsymbol{\Sigma}_i$ for $i = 1$ or 2. With such a transformation the columns of $\widetilde{\mathbf{X}}$ are decorrelated and Model (3) can be rewritten as follows:

$$\mathbf{y} = \widetilde{\mathbf{X}}\widetilde{\boldsymbol{\gamma}} + \boldsymbol{\epsilon} \tag{6}$$

where $\widetilde{\boldsymbol{\gamma}} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\gamma}$. The objective function (4) thus becomes:

$$L^{\mathrm{PPLasso}}_{\lambda_1,\lambda_2}(\widetilde{\boldsymbol{\gamma}}) = \frac{1}{2}\left\|\mathbf{y} - \widetilde{\mathbf{X}}\widetilde{\boldsymbol{\gamma}}\right\|_2^2 + \lambda_1 \left\| \begin{bmatrix} \mathbf{0}_{p,1} & \mathbf{0}_{p,1} & D_1 \\ \mathbf{0}_{p,1} & \mathbf{0}_{p,1} & \frac{\lambda_2}{\lambda_1}D_2 \end{bmatrix} \boldsymbol{\Sigma}^{-1/2}\widetilde{\boldsymbol{\gamma}} \right\|_1 . \tag{7}$$

In order to abtain the estimation of $\widetilde{\boldsymbol{\gamma}}$ and evenutally of $\boldsymbol{\gamma}$, we define a first estimator of $\widetilde{\boldsymbol{\gamma}} = (\widetilde{\alpha}_1, \widetilde{\alpha}_2, \widetilde{\boldsymbol{\beta}}_1', \widetilde{\boldsymbol{\beta}}_2')$ as follows:

$$\widehat{\widetilde{\boldsymbol{\gamma}}}_0(\lambda_1,\lambda_2) = (\widehat{\widetilde{\alpha}}_1, \widehat{\widetilde{\alpha}}_2, \widehat{\widetilde{\boldsymbol{\beta}}}_{10}', \widehat{\widetilde{\boldsymbol{\beta}}}_{20}') = \arg\min_{\widetilde{\boldsymbol{\gamma}}} L^{\mathrm{PPLasso}}_{\lambda_1,\lambda_2}(\widetilde{\boldsymbol{\gamma}}), \tag{8}$$

for each fixed λ_1 and λ_2. To better estimate $\widetilde{\boldsymbol{\beta}}_1$ and $\widetilde{\boldsymbol{\beta}}_2$, a thresholding was applied to $\widehat{\widetilde{\boldsymbol{\beta}}}_0(\lambda_1,\lambda_2) = (\widehat{\widetilde{\boldsymbol{\beta}}}_{10}(\lambda_1,\lambda_2)', \widehat{\widetilde{\boldsymbol{\beta}}}_{20}(\lambda_1,\lambda_2)')'$. For K_1 (resp. K_2) in $\{1,\ldots,p\}$, let Top_{K_1} (resp. Top_{K_2}) be the set of indices corresponding to the K_1 (resp. K_2) largest values of the components of $|\widehat{\widetilde{\boldsymbol{\beta}}}_{10}(\lambda_1,\lambda_2)|$ (resp. $|\widehat{\widetilde{\boldsymbol{\beta}}}_{20}(\lambda_1,\lambda_2)|$), then the estimator of $\widetilde{\boldsymbol{\beta}} = (\widetilde{\boldsymbol{\beta}}_1', \widetilde{\boldsymbol{\beta}}_2')$ after the correction is denoted by $\widehat{\widetilde{\boldsymbol{\beta}}}(\lambda_1,\lambda_2) = (\widehat{\widetilde{\boldsymbol{\beta}}}_1^{(\widehat{K}_1)}(\lambda_1,\lambda_2), \widehat{\widetilde{\boldsymbol{\beta}}}_2^{(\widehat{K}_2)}(\lambda_1,\lambda_2))$ where the jth component of $\widehat{\widetilde{\boldsymbol{\beta}}}_i^{(K_i)}(\lambda_1,\lambda_2)$, for $i = 1$ or 2, is defined by:

$$\widehat{\widetilde{\boldsymbol{\beta}}}_{ij}^{(K_i)}(\lambda_1,\lambda_2) = \begin{cases} \widehat{\widetilde{\boldsymbol{\beta}}}_{i0j}(\lambda_1,\lambda_2), & j \in \mathrm{Top}_{K_i} \\ K_1\text{th largest value of } |\widehat{\widetilde{\boldsymbol{\beta}}}_{i0j}(\lambda_1,\lambda_2)|, & j \notin \mathrm{Top}_{K_i}. \end{cases} \tag{9}$$

With $\widehat{\widetilde{\boldsymbol{\beta}}} = (\widehat{\widetilde{\boldsymbol{\beta}}}_1', \widehat{\widetilde{\boldsymbol{\beta}}}_2')$, the estimators of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1$ can be obtained by $\widehat{\boldsymbol{\beta}}_{10} = \boldsymbol{\Sigma}_1^{-1/2}\widehat{\widetilde{\boldsymbol{\beta}}}_1$ and $(\widehat{\boldsymbol{\beta}}_{20} - \widehat{\boldsymbol{\beta}}_{10}) = \boldsymbol{\Sigma}_2^{-1/2}\widehat{\widetilde{\boldsymbol{\beta}}}_2 - \boldsymbol{\Sigma}_1^{-1/2}\widehat{\widetilde{\boldsymbol{\beta}}}_1$. As previously, another thresholding was applied to $\widehat{\boldsymbol{\beta}}_{10}$ and $\widehat{\boldsymbol{\beta}}_{20}$: for $i = 1$ or 2,

$$\widehat{\boldsymbol{\beta}}_{ij}^{(M_i)}(\lambda_1,\lambda_2)=\begin{cases}\widehat{\boldsymbol{\beta}}_{i0j}(\lambda_1,\lambda_2), & j\in \mathrm{Top}_{M_i}\\ 0, & j\notin \mathrm{Top}_{M_i},\end{cases} \tag{10}$$

for each fixed λ_1 and λ_2. The biomarkers with non-zero coefficients in $\widehat{\boldsymbol{\beta}}_1=\widehat{\boldsymbol{\beta}}_1^{(M_1)}$ (resp. $\widehat{\boldsymbol{\beta}}_2^{(M_2)}-\widehat{\boldsymbol{\beta}}_1^{(M_1)}$) are considered as prognostic (resp. predictive) biomarkers.

In practice, $\boldsymbol{\Sigma}$ is replaced by its estimator obtained by using the approach proposed by Boileau et al. (2021).

2 Numerical experiments

We propose hereafter to compare the performance of our method with the standard Lasso criterion in terms of prognostic and predictive biomarker selection. For the Lasso criterion, in order to directly estimate prognostic and predictive effects, $\mathbf{X}$ and $\boldsymbol{\gamma}$ in Model (3) were replaced by

$$\mathbf{X}^*=\begin{bmatrix}\mathbf{1}_{n_1,1} & \mathbf{0}_{n_1,1} & \mathbf{X}_1 & \mathbf{0}_{n_1,p}\\ \mathbf{0}_{n_2,1} & \mathbf{1}_{n_2,1} & \mathbf{X}_2 & \mathbf{X}_2\end{bmatrix},$$

and $\boldsymbol{\gamma}^*=(\alpha_1,\alpha_2,\boldsymbol{\beta}_1^*,\boldsymbol{\beta}_2^*)$, respectively, where $\mathbf{X}_1$ and $\mathbf{X}_2$ are defined in (1), $\mathbf{0}_{i,j}$ (resp. $\mathbf{1}_{i,j}$) denotes a matrix having i rows and j columns and containing only zeros (resp. ones). Note that this is the modeling proposed by Lipkovich et al. (2017). The sparsity enforcing constraint was put on the coefficients $\boldsymbol{\beta}_1^*$ and $\boldsymbol{\beta}_2^*$ which boils down to putting a sparsity enforcing constraint on $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2-\boldsymbol{\beta}_1$.

2.1 Simulation setting

All simulated datasets were generated from Model (3) where the n_1 (n_2) rows of $\mathbf{X}_1$ ($\mathbf{X}_2$) are assumed to be independent Gaussian random vectors with a covariance matrix $\boldsymbol{\Sigma}_1=\boldsymbol{\Sigma}_2=\boldsymbol{\Sigma}_{bm}$, and $\boldsymbol{\epsilon}$ is a standard Gaussian random vector independent of $\mathbf{X}_1$ and $\mathbf{X}_2$. We defined $\boldsymbol{\Sigma}_{bm}$ as:

$$\boldsymbol{\Sigma}_{bm}=\begin{bmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}\\ \boldsymbol{\Sigma}_{12}^T & \boldsymbol{\Sigma}_{22}\end{bmatrix} \tag{11}$$

where $\boldsymbol{\Sigma}_{11}$ (resp. $\boldsymbol{\Sigma}_{22}$) are the correlation matrix of prognostic (resp. non-prognostic) biomarkers with off-diagonal entries equal to a_1 (resp. a_3). Morever, $\boldsymbol{\Sigma}_{12}$ is the correlation matrix between prognostic and non-prognostic variables with entries equal to a_2. In our simulations $(a_1,a_2,a_3)=(0.3,0.5,0.7)$, which is a framework proposed by Xue and Qu (2017). We checked that the Irrepresentable Condition (IC) of Zhao and Yu (2006) is violated and thus the standard Lasso cannot recover the positions of the null and non null variables. For each dataset we assumed $n_1=n_2=50$. We further assume a relative treatment effect of 1 ($\alpha_1=0$ and $\alpha_2=1$). The number of biomarkers p varies from 200 to 2000. The number of active biomarkers was set to 10 (*i.e.* 5 purely prognostic biomarkers

with $\boldsymbol{\beta}_{1j} = \boldsymbol{\beta}_{2j} = b_1 = 1$ $(j = 1, ..., 5)$ and 5 biomarkers both prognostic and predictive with $\boldsymbol{\beta}_{1j} = b_1$ and $\boldsymbol{\beta}_{2j} = b_2 = 2$ $(j = 6, ..., 10)$).

To assess the performance of the methods in selecting the prognostic and predictive biomarkers: the TPR_{prog} as the true positive rate (i.e. rate of active biomarkers selected) and FPR_{prog} the false positive rate (i.e. rate of inactive biomarkers selected) of the selection of prognostic biomarkers, and similarly for predictive biomarkers with TPR_{pred} and FPR_{pred}. We further note TPR_{all} and FPR_{all} the criterion of overall selection among all candidate biomarkers regardless their prognostic or predictive effect. The objective of the selection is to maximize the TPR_{all} and minimize the FPR_{all}. All metrics were calculated by averaging the results of 100 replications for each scenario.

2.2 Biomarker selection results

$\text{PPLasso}_{\text{oracle}}$ (resp. $\text{PPLasso}_{\text{est}}$) corresponds to the results of the method by considering the true (resp. estimated) matrix $\boldsymbol{\Sigma}_{bm}$. The parameter λ was chosen in order to maximize $(\text{TPR}_{\text{all}} - \text{FPR}_{\text{all}})$ for the different approaches.

Figure 1 shows that PPLasso achieved to select all prognostic factors (TPR_{prog} almost 1) even for large p, with limited false positive prognostic biomarkers selected. The results obtained from the oracle and estimated $\boldsymbol{\Sigma}_{bm}$ are comparable. In this scenario where the IC is violated, PPLasso globally outperforms Lasso. The Lasso approach failed in selecting all truly prognostic and predictive biomarkers, and the number of missed active biomarkers increased with the dimension p ($\text{TPR}_{\text{prog}} = 0.825$ and 0.448, $\text{TPR}_{\text{pred}} = 0.810$ and 0.454 for $p = 200$ and 2000, respectively).

For further experiments, we refer the reader to Zhu et al. (2022).

References

Zhu, W., Lévy-Leduc C., and Ternès, N. (2022). Identification of prognostic and predictive biomarkers in high-dimensional data with PPLasso. arXiv:2202.01970.

Xue, F. and Qu, A. (2017) Variable selection for highly correlated predictors. arXiv:1709.04840.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research*, 7(90), pp. 2541–2563.

Lipkovich, I., Dmitrienko, A., and B. DAgostino Sr., R. (2017). Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials. *Statistics in Medicine*, 36(1), pp. 136–196.

Boileau, P., Hejazi, N. S., van der Laan, M. J., and Dudoit, S. (2021). Cross-validated loss-based covariance matrix estimator selection in high dimensions. arXiv:2102.09715.

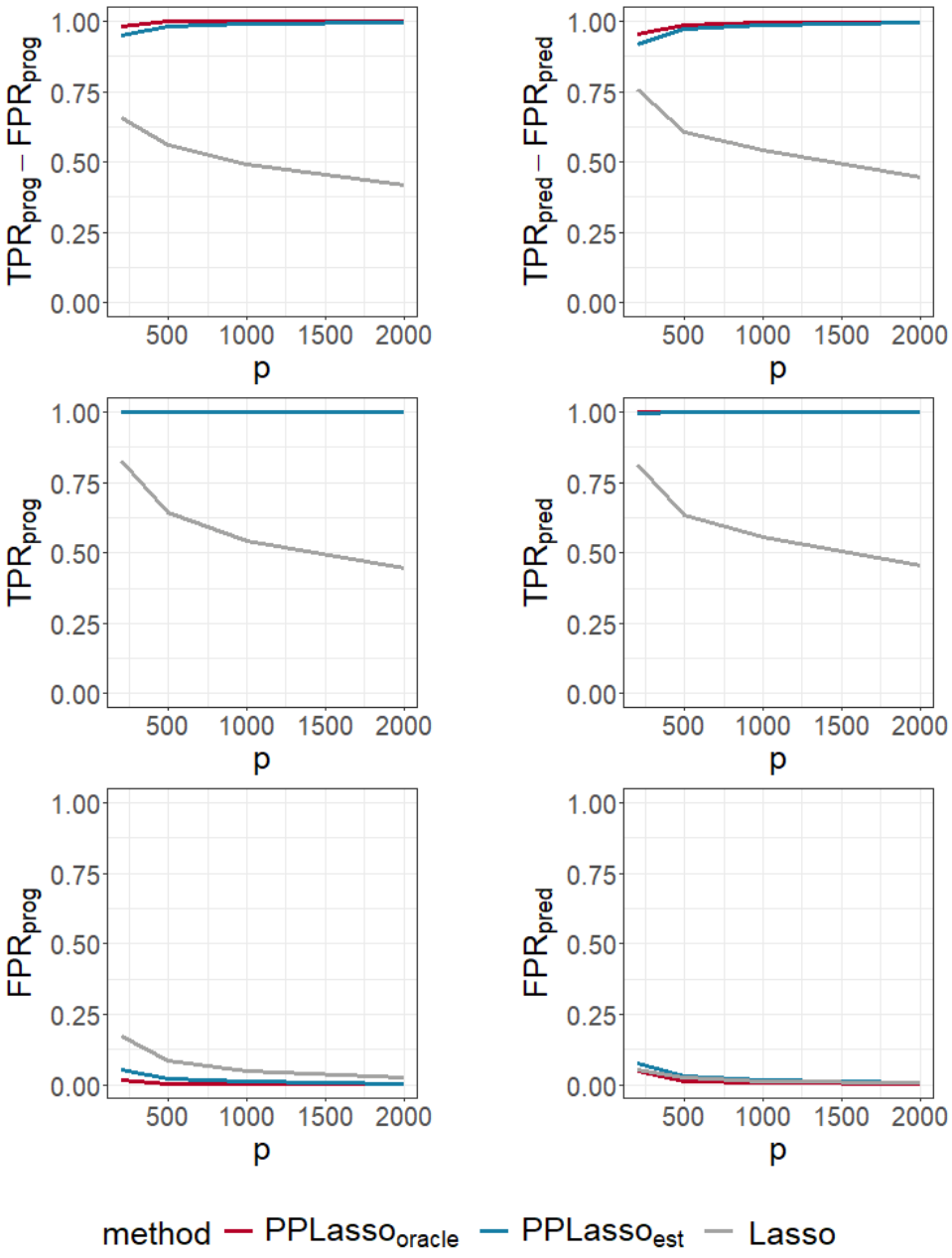


Figure 1: (TPR-FPR) and the corresponding True Positive Rate (TPR) and False Positive Rate (FPR) for prognostic (left) and predictive (right) biomarkers.

Impact d'une copule non-Gaussienne dans l'estimation REML du modèle génétique animal bivarié pour des populations sous sélection

Tom Rohmer [1], Anne Ricard [2,3] & Ingrid David [1]

[1] *GenPhySE, Université de Toulouse, INRAE, ENVT, F-31326, Castanet Tolosan.*
[2] *Université Paris-Saclay, INRAE, AgroParisTech, GABI, Jouy-en-Josas*
[3] *Institut Français du Cheval et de l'Equitation, Pôle Développement, Innovation et Recherche, Exmes*
{*tom.rohmer, anne.ricard, ingrid.david*}*@inrae.fr*

Résumé. Dans les modèles multi-caractères utilisés en génétique animal, les composantes de variances sont très souvent estimées par des méthodes de maximum de vraisemblance restreinte (REML) sous l'hypothèse de normalité jointe des caractères, bien qu'en pratique cette hypothèse n'est pas toujours réaliste. Nous avons simulé des populations mimant un schéma de sélection classique rencontré dans les élevages porcins et mesuré l'impact d'une distribution multivariée non Gaussienne sur les résidus (de part une copule non-Gaussienne) sur les estimations réalisées en pratique. Les résultats ont montré que lorsque les reproducteurs sont sélectionnés au hasard, nous n'observons aucun impact significatif sur les estimés, malgré l'hypothèse Gaussienne sous-jacente. Néanmoins, lorsque les reproducteurs sont sélectionnés de façon à améliorer les deux caractères d'intérêt par un processus de troncation basé sur les prédictions BLUP des valeurs génétiques, on observe des différences significatives avec les paramètres théoriques, en particulier avec des distributions bivariées asymétriques sur la partie résiduelle.

Mots-clés. Modèle génétique animal, multi-caractères, copule.

Abstract. In the classical multiple-trait animal models used in a genetic context, variance components are frequently estimated using Restricted Maximum Likelihood method (REML) under the Gaussian assumption for the bivariate traits. Nevertheless, in practice, this hypothesis is not always realistic. We simulated populations mimicking a selection schema encountered in pig farming and assessed the impact of a non-Gaussian multivariate distribution on the residual part generated by a non-Gaussian copula. When the breeders were selected at random, we did not observe any significant impact on the estimated parameters despite the underlying Gaussian assumption. Nevertheless, when the breeders were selected in order to improve the two traits, with a truncation selection process based on BLUP estimation of the breeding values, we observed significant differences with the true parameters, particularly with asymmetric bivariate distributions on the residual part.

Keywords. genetic animal model, multi-trait, copula

1 Introduction

En génétique animal, les modèles mixtes sont fréquemment utilisés afin de dissocier la part de variabilité due à la génétique et la part de variabilité due à l'environnement sur les phénotypes étudiés. Nous citons ici à titre d'exemple Mrode (2014). Ces derniers sont étendus au cas bivarié pour tenir compte des covariances entre caractères lorsque deux phénotypes $y_{i,1}, y_{i,2}$ sont observés pour l'animal $i = 1, \ldots, n$ voir par ex. Meyer (1991). Dans le modèle mixte, la matrice de variance-covariance entre les vecteurs $\boldsymbol{a}_j = (a_{1,j}, \ldots, a_{N,j})$, $N \geq n$, supposés Gaussiens, constitués des valeurs génétiques des animaux du pédigrée pour le caractère $j = 1, 2$, est la suivante :

$$\text{var} \begin{bmatrix} \boldsymbol{a}_1 \\ \boldsymbol{a}_2 \end{bmatrix} = G \otimes A, \quad \text{avec } G = \begin{pmatrix} \sigma^2_{\boldsymbol{a}_1} & \sigma_{\boldsymbol{a}_{12}} \\ \sigma_{\boldsymbol{a}_{12}} & \sigma^2_{\boldsymbol{a}_2} \end{pmatrix},$$

avec A la matrice des coefficients de parenté associée au pédigrée des animaux et '$\otimes$' désigne le produit de Kronecker.

Les composantes de variances sont généralement estimées par des méthodes de maximum de vraisemblance restreinte (REML) sous l'hypothèse de normalité jointe des phénotypes bien qu'en pratique cette hypothèse n'est pas toujours réaliste, en particulier de part des résidus non nécessairement Gaussiens. Et même si chacune des marges a une distribution Gaussienne, cela ne garantit pas la normalité des vecteurs bivariés.

Plus précisément, considérons un vecteur aléatoire $\boldsymbol{\mathcal{X}} = (\mathcal{X}_1, \mathcal{X}_2)$ de fonction de répartition (f.d.r.) $\boldsymbol{F}$, et de fonctions de répartitions marginales F_1, F_2 supposées continues. Le théorème de Sklar (1959) permet d'affirmer qu'il existe une unique fonction C : $[0, 1]^d \to [0, 1]$ telle que :

$$\boldsymbol{F}(\boldsymbol{x}) = C\{F_1(x_1), F_2(x_2)\}, \quad \boldsymbol{x} = (x_1, x_2) \in \mathbb{R}^2.$$

Cette fonction C dite copule caractérise la structure de dépendance du vecteur aléatoire $\boldsymbol{\mathcal{X}}$. En particulier une structure de dépendance non-Gaussienne entraîne une distribution bivariée non-gaussienne. Les contour plots des distributions considérées dans cette étude pour la partie résiduelle du modèle sont représentés dans la Figure 1.

Le but de ce papier est de mesurer l'impact d'une copule non-Gaussienne dans la partie résiduelle du modèle mixte, sur les estimations REML des composantes de variance (supposant le modèle Gaussien), lorsque les marges sont Gaussiennes.

2 Simulations

Les populations d'animaux apparentés ont été générées en suivant Gonzales et al. (2020), mimant un schéma de sélection pouvant être rencontré dans un élevage porcin. Les fondateurs (G_0) consistaient en 204 femelles et 12 mâles non apparentés. Dans les générations

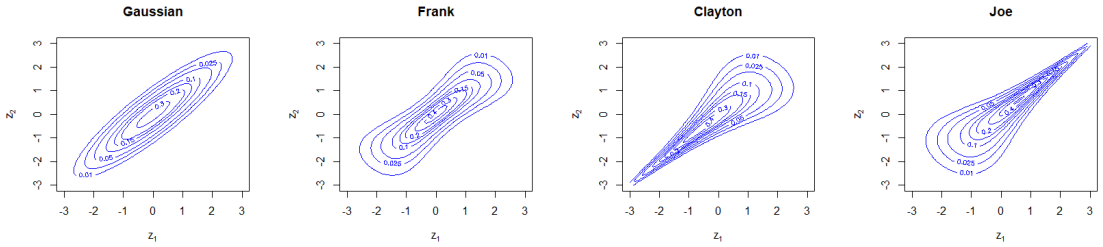


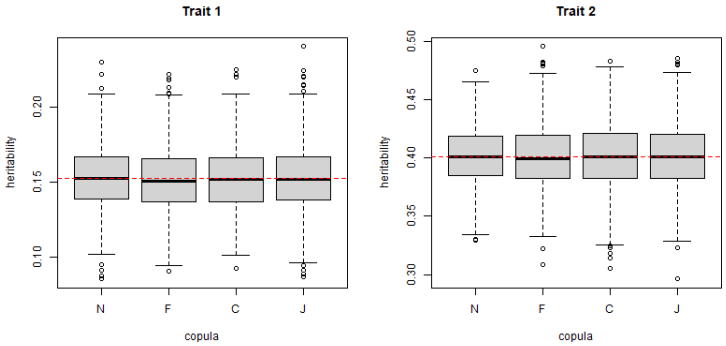
FIGURE 1 – Contours plot des distributions bivariées avec marge Gaussienne standard et copule normale, Frank, Clayton et Joe avec une corrélation de Kendall $\tau = 0.7$

suivantes (G_1–G_8), chaque mâle reproducteur a été accouplé à 17 femelles pour produire 10 femelles et 2 mâles phénotypés par femelle, soit $n = 19800$ individus ($N = n$) à l'issu de 8 générations. Dans les 3 premières générations, les mâles reproducteurs ont été choisi au hasard. Dans les 5 générations suivantes, pour chaque animal, les valeurs génétiques pour les deux caractères ont été prédites par BLUP (voir Henderson, 1975). Au sein de chaque famille de même père, les individus dont la somme de leurs BLUP étaient les plus fortes ont été sélectionné (sélection par troncation). Dans chaque génération, un ratio de sélection pour la reproduction de 2.9% de mâles et 10% des femelles a été considéré. Les accouplements ont été réalisé au hasard de telle sorte que les pleins frères et demi-frères ne soient jamais accouplés afin de réduire la consanguinité.

Les valeurs génétiques des fondateurs ont été simulées indépendamment selon une Gaussienne bivariée de matrice de covariance G. Pour les générations G_1–G_8, les valeurs génétiques $(a_{i,1}, a_{i,2})$, $i = 217, \ldots, n$ ont été simulé selon une loi normale bivariée respectant un schéma Mendélien : pour $j = 1, 2$, $a_{i,j} = 0.5(a_{i_p,j} + a_{i_m,j}) + M_{ij}$, où i_p et i_m sont les indices du père et de la mère de l'animal i et les vecteurs (M_{i1}, M_{i2}) sont simulés selon une Gaussienne multivariée de matrice de variance-covariance $G/2$.

Les vecteurs des résidus du modèle mixte ont été échantillonnés indépendamment selon une distribution bivariée avec des marges Gaussiennes de variance σ_j^2 fixée à 1 et de copule normal(F), Frank(F), Clayton(C) et Joe(J) avec corrélation de Kendall fixée à 0.7, voir Nelson 2007. La corrélation génétique théorique a été fixée à 0.59 (Kendall à 0.4). Les variances génétiques théoriques étaient de 0.18 ou 0.67 amenant des héritabilité théoriques $h_j^2 = \sigma_{a_j}^2/(\sigma_{a_j}^2 + \sigma_{e_j}^2)$, $j = 1, 2$ de 0.153 ou 0.401 pour les caractères simulés. De la génération G_3 à la génération G_8, nous avons estimé l'héritabilité des caractères ainsi que les corrélations génétiques et résiduelles à partir des estimations REML des composantes de variance sous l'hypothèse de normalité jointe des phénotypes.

a. 3ème génération (sélection aléatoire)



b. 8ème génération (sélection par troncation)

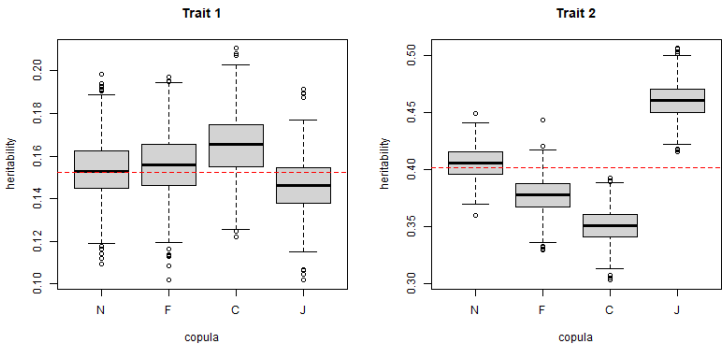


FIGURE 2 – Boxplot des héritabilités estimées des deux caractères (a.) sans sélection (b.) avec selection, pour les copules Normale (N), Frank (F), Clayton (C) et Joe(J) avec une dépendance résiduelle forte (corrélation de Kendall à 0.7) et une dépendance génétique modérée (corrélation de Kendall à 0.4). Les héritabilités théoriques (lignes rouges) des deux caractères sont respectivement 0.153 et 0.401.

3 Résultats

Les résultats ont montré qu'avec une sélection aléatoire des reproducteurs, quel que soit la copule sur les résidus, les biais pour les héritabilités estimées (Figure 2.a) ainsi que les biais pour les corrélations estimées (Table 1.a) étaient extrêmement faibles, en valeur absolue entre 6.77×10^{-6} et 1.43×10^{-3} pour les héritabilités (SE entre 0.020 et 0.030) et entre 2.4×10^{-7} et 0.010 pour les corrélations (SE entre 0.004 et 0.093). Aucunes différences entre les valeurs théoriques et estimées n'apparaissent significatives pour un t-test au niveau $\alpha = 5\%$.

À contrario, dans le cas d'une population soumise à une stratégie de sélection non aléatoire (Figure 2.b et Table 1.b), nous avons observé des différences significatives entre les paramètres estimés et ceux simulés lorsque les résidus n'étaient pas gaussiens et les

corrélations résiduelles suffisamment importantes. Notamment les différences avec les valeurs théoriques pour les héritabilités apparaissaient significatives pour le caractère ayant l'héritabilité la plus élevé, lorsque les héritabilités théoriques étaient différentes pour les deux caractères et pour les copules asymétriques C et J. Le biais estimé était de -0.051 pour la copule C (SE de 0.014) marquant une sous-estimation et de 0.059 pour la copule J (SE de 0.015) marquant une surestimation. Le biais estimé pour la copule F apparaît plus modéré, -0.024 (SE de 0.015).

Les différences avec les valeurs théoriques pour les corrélations étaient les plus fortes lorsque les héritabilités des deux caractères étaient identiques, et modérées ($h_j^2 = 0.401$, $j = 1, 2$). Nous remarquons dans ce cas une sous-estimation de la corrélation génétique pour J (biais estimé de -0.128, SE à 0.027) et une surestimation de la corrélation génétique pour C et G (bais resp. de 0.106 et 0.049, SE resp. de 0.018 et 0.020).

Ces résultats indiquent que le processus de sélection effectué dans les queues droites des distributions, améliorant simultanément les deux caractères, va entraîner des surestimations des corrélations résiduelles, lorsque la distribution jointe, portée par les résidus, a des corrélations dans la queue droite de la distribution plus fortes que le cas Gaussien (c'est le cas par exemple de la copule J) et des sous-estimation dans le cas contraire (par exemple C et F). De là, une sur(resp. sous)-estimation de la corrélation résiduelle va entraîner une sous(resp. sur)-estimation de la partie génétique. Vraisemblablement, dans le cas de différentes héritabilités pour les deux caractères, le processus de sélection mis en œuvre en présence d'une copule non-Gaussienne (en particulier asymétrique et à queue lourde) va altérer les valeurs génétiques du caractère avec la plus grande héritabilité et ainsi causer des sur/sous estimations des héritabilités.

Bibliographie

González-Diéguez D., Tusell L., Bouquet A., Legarra A., and Vitezica Z. G. (2020), Purebred and crossbred genomic evaluation and mate allocation strategies to exploit dominance in pig crossbreeding schemes, *G3 : Genes, Genomes, Genetics*, vol. 10, no. 8, pp. 2829–2841.

Henderson C. R. (1975), Best linear unbiased estimation and prediction under a selection model, *Biometrics*, pp. 423–447.

Meyer K. (1991) Estimating variances and covariances for multivariate animal models by restricted maximum likelihood, *Genetics Selection Evolution*, vol. 23, no. 1, pp. 1–17.

Mrode R. A. (2014), Linear models for the prediction of animal breeding values, *Cabi.*

Nelson RB. (2007), An introduction to copulas, *Springer Science & Business Media.*

Sklar M. (1959) Fonctions de répartition an dimensions et leurs marges, *Publ. inst. statist. univ. Paris.*

TABLE 1 – Biais et standard error (SE) des corrélations génétiques et résiduelles estimées à la génération G_3 et G_8.

a. 3ème génération (sélection aléatoire)

	hérit. théorique			Corrélations estimées							
				Corrélation génétique				Corrélation résiduelle			
	h_1^2	h_2^2		N	F	C	J	N	F	C	J
	0.153	0.153	biais	-0.004	-0.007	-0.004	-0.003	-0.000	0.001	0.000	0.001
			SE	0.062	0.063	0.067	0.063	0.004	0.006	0.006	0.006
G_3	0.153	0.401	biais	-0.001	-0.003	-0.001	-0.001	0.000	0.001	0.000	0.001
			SE	0.049	0.051	0.054	0.051	0.007	0.008	0.009	0.009
	0.401	0.401	biais	-0.001	-0.003	-0.001	-0.002	0.000	0.001	0.000	0.001
			SE	0.036	0.037	0.040	0.039	0.008	0.010	0.011	0.011

b. 8ème génération (sélection par troncation)

	hérit. théorique			Corrélations estimés							
	h_1^2	h_2^2		N	F	C	J	N	F	C	J
	0.153	0.153	biais	-0.003	0.034	0.089*	-0.108*	-0.000	-0.004	-0.010*	0.013*
			SE	0.035	0.034	0.031	0.044	0.003	0.003	0.003	0.005
G_8	0.153	0.401	biais	-0.001	0.003	0.030	0.039	-0.000	-0.009*	-0.023*	0.011*
			SE	0.027	0.028	0.026	0.026	0.004	0.005	0.005	0.005
	0.401	0.401	biais	-0.001	0.049*	0.106*	-0.128*	-0.001	-0.018*	-0.040*	0.052*
			SE	0.021	0.020	0.018	0.027	0.005	0.006	0.006	0.008

Les biais et les SE ont été obtenus à partir de 1000 simulations. Les copules des résidus du modèle mixte étaient normale(N), Frank(F), Clayton(Cl) and Joe(J). La corrélation génétique théorique était de 0.588, correspondant à un tau de Kendall de 0.4. Les corrélations résiduelles pour N, F, C and J étaient respectivement de 0.891, 0.846, 0.852 et 0.850 correspondant à un tau de Kendall de 0.7. '*' : Différence significative entre corrélation estimée et théorique pour le t-test au level $\alpha = 0.05$.

ÉVALUATION DE L'IMPACT DE SCENARIOS DE CONSOMMATION DE TABAC SUR LA CHARGE FUTURE DE L'INFARCTUS DU MYOCARDE EN FRANCE JUSQU'EN 2035 : UNE APPROCHE BASEE SUR UN MODELE ILLNESS-DEATH

Johann Kuhn[1] & Yann Le Strat[2] & Christophe Bonaldi[3] & Clémence Grave[4] & Valérie Olié[5] & Pierre Joly[6]

[1] *Santé publique France, Direction Appui, Traitements et Analyse de Données (DATA), Saint-Maurice 94415, France – johann.kuhn@santepubliquefrance.fr*
[1] *Université Paris-Est Créteil, Créteil 94010, France*
[2] *Santé publique France, Direction Appui, Traitements et Analyse de Données (DATA), Saint-Maurice 94415, France – yann.lestrat@santepubliquefrance.fr*
[3] *Santé publique France, Direction Appui, Traitements et Analyse de Données (DATA), Saint-Maurice 94415, France – christophe.bonaldi@santepubliquefrance.fr*
[4] *Santé publique France, Direction des Maladies non Transmissibles et Traumatismes (DMNTT), Saint-Maurice 94415, France – clemence.grave@santepubliquefrance.fr*
[5] *Santé publique France, Direction des Maladies non Transmissibles et Traumatismes (DMNTT), Saint-Maurice 94415, France – valerie.olie@santepubliquefrance.fr*
[6] *Université de Bordeaux, ISPED, Centre Inserm U1219 – Bordeaux Population Health, Bordeaux 33076, France – pierre.joly@u-bordeaux.fr*

Résumé. En France, l'infarctus du myocarde est une cause importante de morbidité, de recours aux soins, d'altération de la qualité de vie et de mortalité. Dans un premier travail, nous avons projeté plusieurs indicateurs de l'infarctus du myocarde (nombre de cas prévalents, prévalence, âge moyen des cas incidents) jusqu'en 2035, montrant une augmentation de la prévalence avec un quasi-doublement des cas chez les hommes et les femmes en 2035 comparé à 2015 et une augmentation du nombre de cas chez les femmes de moins de 65 ans. Dans ce second travail, nous avons quantifié l'impact de la diminution du tabagisme en France sur les mêmes indicateurs épidémiologiques que ceux du premier travail. Pour cela, nous avons créé différents scénarios de prévention et modélisé, pour chacun, le statut tabagique des individus et l'arrêt du tabac. Nous avons ensuite estimé l'impact de ces scénarios sur plusieurs indicateurs épidémiologiques en 2035 et comparé ces estimations à celle réalisée en l'absence d'action de prévention.

Mots-clés. Infarctus du myocarde, projection, tabac, scénario, incidence, prévalence, modèle markovien

Abstract. In France, myocardial infarction is a significant cause of morbidity, healthcare use, quality of life impairment, and mortality. In a first work, we projected several indicators of myocardial infarction (number of prevalent cases, prevalence, average age of incident cases) until 2035, showing an increase in prevalence with a near doubling of cases in men and women in 2035 compared to 2015 and an increase in the number of cases in women under 65. In this second work, we quantified the impact of the decrease in smoking in France on the same epidemiological indicators as in the first work. To do this, we created different prevention scenarios and modeled the smoking status of individuals and smoking cessation for each one. We then estimated the impact of these scenarios on several epidemiological indicators in 2035 and compared these estimates to the one made in the absence of prevention action.

Keywords. Myocardial infarction, projection, tobacco, scenario, incidence, prevalence, Markov model

1 Introduction

En France, les maladies cardiovasculaires représentent un fardeau extrêmement lourd puisqu'elles sont à l'origine annuellement de 140 000 décès (seconde cause de décès tous sexes confondus, mais première pour les femmes) et de 1,2 million de patients hospitalisés. De plus, elles exposent à des séquelles lourdes et invalidantes et impactent la qualité de vie des patients. Parmi elles, les cardiopathies ischémiques dont l'infarctus du myocarde ont été identifiées comme la seconde cause d'années de vie perdues en France (Vos et al., 2020).

Dans un précédent travail, nous avons projeté plusieurs indicateurs épidémiologiques (incidence, prévalence, nombre de cas prévalents et âge moyen des cas incidents) pour l'infarctus du myocarde (IDM) chez les hommes et les femmes jusqu'en 2035 en France. Notre méthode de projection se basait sur un modèle « *illness-death* » qui modélisait les possibles transitions d'un individu entre trois états : état 0 (vivant, sans IDM), état 1 (vivant, avec IDM), état 2 (décédé). La transition α_{01} représentant le taux d'incidence de l'IDM a été estimée à partir d'un modèle âge-cohorte en utilisant des splines naturelles et les données d'hospitalisation du PMSI (Programme de Médicalisation des Systèmes d'Information). La transition α_{02} représentant le taux de mortalité des individus sans IDM a été estimée avec un modèle de Gompertz-Makeham en utilisant les données démographiques de l'INSEE. Enfin α_{12}, le taux de mortalité des individus ayant développé un IDM, a été estimé à partir de α_{02} multiplié par un risque relatif dépendant du temps depuis l'apparition de la maladie qui a été estimé à partir des données du SNDS (Système National des Données de Santé). Nos projections ont montré une augmentation de la prévalence et du nombre de cas prévalents avec un quasi-doublement des cas chez les hommes et les femmes entre 2015 et 2035. Un autre résultat intéressant était l'augmentation du nombre de cas chez les femmes relativement jeunes (en dessous de l'âge moyen de l'IDM).

Le tabac est l'un des facteurs de risque les plus importants d'IDM (Yusuf et al., 2020) et bien documenté dans la littérature permettant de disposer d'informations nécessaires à la modélisation. Malgré une diminution de la proportion de fumeurs quotidiens estimée à 30.0 % (33.7 % pour les hommes, 26.4 % pour les femmes) en 2000, celle-ci n'est jamais descendue en dessous des 20 % entre 2000 et 2020 (Pasquereau et al., 2021). Pour faire face à ce facteur de risque encore prépondérant, le gouvernement a mis en place en 2014 le programme national de réduction du tabac (PNRT). C'est à l'issue de ce programme que plusieurs actions emblématiques ont été mises en place (paquet neutre, droit de prescription des traitements de substitution nicotinique élargie, mois sans tabac, etc.), ce qui a entraîné des premiers résultats très encourageants sur la diminution du tabagisme. Pour la période 2018-2022, un nouveau programme appelé programme national de lutte contre le tabac (PNLT) a été mis en place avec des objectifs concernant la diminution de proportion de fumeurs quotidiens pour 2027 et jusqu'en 2032 (*Lutte contre le tabagisme - Ministère des Solidarités et de la Santé*, 2021.).

Dans ce travail, nous avons d'abord modélisé les taux d'incidence de l'IDM selon le statut tabagique des individus (non-fumeur, ancien fumeur et fumeur) en prenant en compte l'arrêt « naturel ». En effet, le taux d'incidence de l'IDM de la population générale est une moyenne pondérée des taux d'incidence associés au statut tabagique. Ensuite, nous avons modélisé différents scénarios simulant une baisse de la proportion de fumeurs à partir de 2023 jusqu'en 2035. Enfin, nous avons projeté les indicateurs épidémiologiques (nombre de cas prévalents, prévalence, âge moyen des cas incidents) en population générale jusqu'à 2035 afin de mesurer l'impact des scénarios sur ces indicateurs.

Nous allons d'abord décrire les données du tabagisme en France issues du Baromètre de Santé publique France. Le choix des hypothèses, la description du modèle et la méthode d'estimation seront expliqués dans la section 3. Dans la section 4, nous détaillerons les formules des probabilités de transition et du nombre de cas prévalents. Enfin, nous terminerons par une discussion sur l'objectif

de ce travail.

2 Données

Depuis plus de 25 ans, la compréhension et la connaissance des attitudes et des comportements de santé des personnes vivant en France sont réalisées à partir d'enquêtes afin d'orienter les politiques de prévention et d'information de la population. Ces enquêtes, regroupées dans le Baromètre de Santé publique France, sont répétées et ont pour but de surveiller les principaux comportements, attitudes et perceptions liés à la prise de risque et à l'état de santé comme le tabagisme (*Baromètres de Santé publique France, 2022*.). Ces enquêtes sont basées sur l'interview d'échantillons représentatifs de la population française, à l'aide d'un système d'interview téléphonique assisté par ordinateur. Les enquêtes sont transversales car elles fournissent un instantané à un moment donné d'un état de santé donné.

Entre 2000 et 2020, nous avons utilisé des estimations de la proportion de non-fumeurs, d'anciens fumeurs et de fumeurs par classe d'âge de cinq ans entre 35 et 75 ans chez les hommes et les femmes en France à partir du Baromètre de Santé publique France. Nous avons fait l'hypothèse que les proportions estimées entre 70 et 75 ans étaient les mêmes jusqu'à 95 ans. Nous avons également considéré que les proportions des individus selon leur statut tabagique restaient constantes après 2020. Les données démographiques (taille de la population) entre 1955 et 2035 par âge (35 à 95 ans) et par sexe sont issues de l'INSEE.

3 Modèle

Dans nos modèles, les sujets entrent à 35 ans. À cet âge, ils sont considérés soit comme non-fumeur, soit comme ancien fumeur, soit comme fumeur. Nous avons considéré un modèle pour chacun de ces cas. Les non-fumeurs ont été définis comme les individus n'ayant jamais fumé de leur vie, les anciens fumeurs comme les individus ayant arrêté de fumer et les fumeurs comme les individus fumant quotidiennement ou occasionnellement.

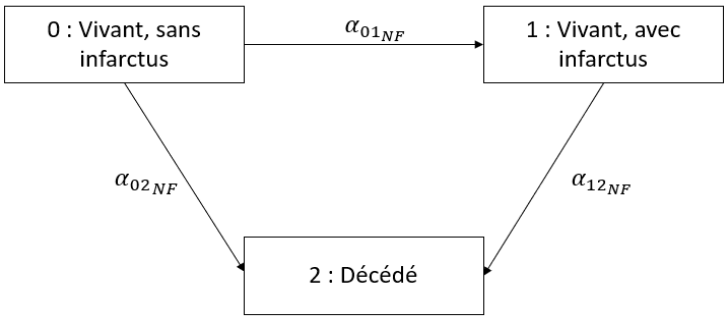


Figure 1 : Modèle illness-death pour les non-fumeurs

Avec $\alpha_{01_{NF}}$ le taux d'incidence d'infarctus du myocarde chez les non-fumeurs, $\alpha_{02_{NF}}$ le taux de mortalité pour les non-fumeurs n'ayant pas développé un infarctus et $\alpha_{12_{NF}}$ le taux de mortalité des non-fumeurs ayant développé un infarctus. De la même manière, nous avons utilisé ce modèle pour les anciens fumeurs avec des taux spécifiques notés $\alpha_{01_{AF}}$, $\alpha_{02_{AF}}$ et $\alpha_{12_{AF}}$.

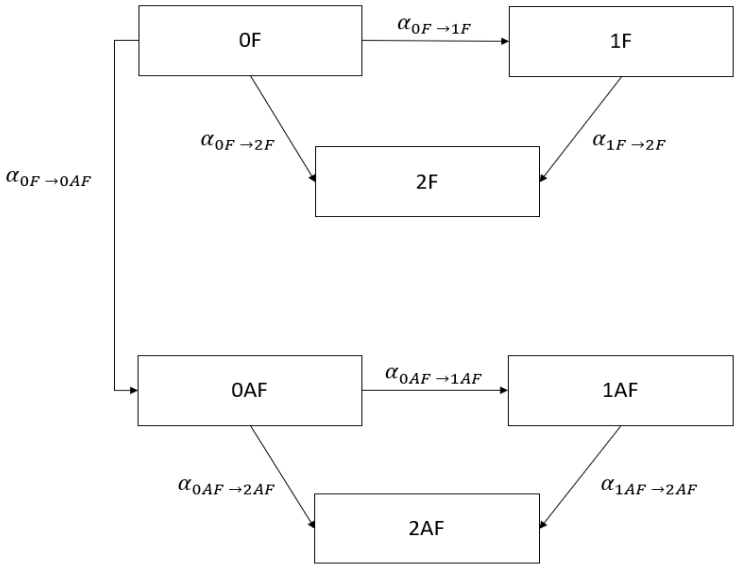


Figure 2 : Modèle à 6 états pour les fumeurs à 35 ans

On note également $\alpha_{0F \to 1F}$ et $\alpha_{0AF \to 1AF}$ le taux d'incidence d'infarctus du myocarde chez les fumeurs et les anciens fumeurs, $\alpha_{0F \to 2F}$ et $\alpha_{0AF \to 2AF}$ le taux de mortalité pour les fumeurs et les anciens fumeurs n'ayant pas développé un infarctus, $\alpha_{1F \to 2F}$ et $\alpha_{1AF \to 2AF}$ le taux de mortalité des fumeurs et anciens fumeurs ayant développé un infarctus, respectivement. Ici, $\alpha_{0F \to 0AF}$ représente la proportion de fumeurs s'arrêtant de fumer.

À noter que les intensités de transition chez les nouveaux anciens fumeurs sont égales à celles des anciens fumeurs : $\alpha_{0AF \to 1AF} = \alpha_{01_{AF}}$, $\alpha_{0AF \to 2AF} = \alpha_{02_{AF}}$, $\alpha_{1AF \to 2AF} = \alpha_{12_{AF}}$.

3.1 Hypothèses

Nous avons considéré qu'un non-fumeur ou un ancien fumeur ne pouvait pas (re)commencer à fumer : cette hypothèse nous semble plausible car nous travaillons avec des personnes âgées de 35 ans ou plus. Par contre, un fumeur pouvait s'arrêter de fumer tout au long de sa vie (à partir de 35 ans et jusqu'à ses 95 ans).

3.2 Calcul du taux d'incidence de l'IDM chez les fumeurs, anciens fumeurs et non-fumeurs à partir du taux d'incidence marginale de l'IDM.

Le taux d'incidence marginale de l'IDM est le taux d'incidence de l'IDM dans la population générale que l'on note α_{01}^{M}. On suppose qu'elle dépend de l'âge a et de la cohorte c. Nous pouvons écrire la relation entre le taux d'incidence marginale α_{01}^{M} et celui des non-fumeurs $\alpha_{01_{NF}}$, des anciens fumeurs $\alpha_{01_{AF}}$ et des fumeurs α_{01_F} $(= \alpha_{0F \to 1F})$ par :

$$\alpha_{01}^{M}(a,c) = p_{NF}\,(a,c)\alpha_{01_{NF}}(a,c) + p_{AF}(a,c)\alpha_{01_{AF}}(a,c) + p_F(a,c)\,\alpha_{01_F}(a,c)$$

avec :

$$\alpha_{01_{AF}}(a,c) = \alpha_{01_{NF}}(a,c) \times \theta_{01_{AF}}(a)$$
$$\alpha_{01_F}(a,c) = \alpha_{01_{NF}}(a,c) \times \theta_{01_F}(a)$$
$$p_{NF}(a,c) + p_{AF}(a,c) + p_F(a,c) = 1,$$

où $p_{NF}(a,c)$, $p_{AF}(a,c)$, $p_F(a,c)$ sont les proportions pour un âge a et une cohorte c chez les non-fumeurs, anciens fumeurs et fumeurs, respectivement. Nous notons aussi $\theta_{01_{AF}}(a)$ et $\theta_{01_F}(a)$ comme

les sur-risques de développer un IDM d'un ancien fumeur par rapport à un non-fumeur et d'un fumeur par rapport à un non-fumeur, respectivement. Nous pouvons alors écrire :

$$\alpha_{01}^{M}(a,c) = p_{NF}\,(a,c)\,\alpha_{01_{NF}}(a,c) + p_{AF}(a,c)\left[\alpha_{01_{NF}}(a,c) \times \theta_{01_{AF}}(a)\right] + p_{F}(a,c)\left[\alpha_{01_{NF}}(a,c) \times \theta_{01_{F}}(a)\right]$$

Qui permet d'en déduire $\alpha_{01_{NF}}$ en fonction des autres quantités connues :

$$\alpha_{01_{NF}}(a,c) = \frac{\alpha_{01}^{M}(a,c)}{p_{NF}\,(a,c) + p_{AF}(a,c)\,\theta_{01_{AF}}(a) + p_{F}(a,c)\,\theta_{01_{F}}(a)}$$

La même méthode est utilisée pour le taux de mortalité chez les personnes n'ayant pas développé un IDM dans la population générale α_{02}^{M}.
Les valeurs de θ_{01_F}, $\theta_{01_{AF}}$ (Teo et al., 2006), θ_{02_F} (Banks et al., 2015), $\theta_{02_{AF}}$ (Jha et al., 2013) sont issues de la littérature.

3.3 Modélisation de l'arrêt du tabac

L'arrêt du tabac est modélisé avec la quantité $\alpha_{0F\to 0AF}$. Nous avons fait l'hypothèse qu'il y avait environ 1 % des fumeurs par an qui s'arrêtaient de fumer en France. Nous avons donc modélisé $\alpha_{0F\to 0AF}$ par une loi exponentielle de paramètre $\lambda = 0.01$. C'est cette quantité que nous avons modifié pour simuler un scénario entraînant une baisse plus forte de la proportion de fumeurs.

4 Indicateurs épidémiologiques

Les indicateurs épidémiologiques s'écrivent à partir des probabilités de transition qui eux-mêmes dépendent des intensités de transition. Pour alléger les notations, nous nous plaçons dans le cadre d'une cohorte c fixée. Ainsi, les formules ne sont écrites qu'en fonction de l'âge. De plus, pour des raisons de place, nous ne développons ici que la façon de calculer le nombre de cas prévalents de l'infarctus du myocarde.

On note les intensités de transition cumulées entre deux âges a_1 et a_2 ($a_1 \leq a_2$) par :

$$A_{ij}(a_1,a_2) = \int_{a_1}^{a_2} \alpha_{ij}(u)\,du$$

La probabilité $P_{01}(a_1,a_2)$ est la probabilité pour un individu d'être en vie et de ne pas avoir développé d'infarctus à l'âge a_1 (état 0) et de développer un infarctus entre les âges a_1 et a_2 et de rester en vie jusqu'à l'âge a_2 (état 1) :

$$P_{01}(a_1,a_2) = \int_{a_1}^{a_2} e^{-A_{01}(a_1,u)-A_{02}(a_1,u)}\,\alpha_{01}(u)e^{-A_{12}(u,a_2)}\,du$$

Nous pouvons écrire P_{01} en fonction du statut tabagique d'un individu.

Pour un non-fumeur :

$$P_{01_{NF}}(a_1,a_2) = \int_{a_1}^{a_2} e^{-A_{01_{NF}}(a_1,u)-A_{02_{NF}}(a_1,u)}\,\alpha_{01_{NF}}(u)e^{-A_{12_{NF}}(u,a_2)}\,du$$

Pour un ancien fumeur :

$$P_{01_{AF}}(a_1, a_2) = \int_{a_1}^{a_2} e^{-A_{01_{AF}}(a_1,u)-A_{02_{AF}}(a_1,u)}\ \alpha_{01_{AF}}(u) e^{-A_{12_{AF}}(u,a_2)}\ du$$

Pour un fumeur :

$$P_{0F\to 1F}(a_1, a_2) = \int_{a_1}^{a_2} e^{-A_{0F\to 1F}(a_1,u)-A_{0F\to 2F}(a_1,u)-A_{0F\to 0AF}(a_1,u)}\ \alpha_{0F\to 1F}(u) e^{-A_{1F\to 2F}(u,a_2)}\ du$$

Pour un fumeur d'âge a_1 ayant arrêté de fumer avant de faire un infarctus tout en ayant survécu jusqu'en a_2 :

$$P_{0F\to 1AF}(a_1, a_2) = \int_{a_1}^{a_2}\int_{a_1}^{a_2} e^{-A_{0F\to 1F}(a_1,u)-A_{0F\to 2F}(a_1,u)-A_{0F\to 0AF}(a_1,u)} \alpha_{0F\to 0AF}(u)$$
$$e^{-A_{01_{AF}}(u,v)-A_{02_{AF}}(u_1,v)} \alpha_{0AF\to 1AF}(v) e^{-A_{1AF\to 2AF}(v,a_2)} du\ dv$$

Le nombre de cas prévalents pour une année t est le nombre de personnes malades de tous âges à l'année t considérée. Notons a_0 l'âge minimal et a_{max} l'âge maximal des individus (respectivement égaux à 35 et 95 ans), $\nu(a_0, t)$ la taille de la population pour un âge a_0 et une année t (INSEE) et c la cohorte, le nombre de cas prévalents s'écrit alors :

$$N_{prev}(t) = \sum_{k=a_0}^{a_{max}} \nu(a_0, t - a_{max} + k) \times P_{01}(a_0, a_{max} - k + a_0|c)$$

Pour calculer le nombre de cas prévalents dans la population générale en prenant compte le statut tabagique des individus, il suffit de sommer tous les cas selon le statut tabagique (non-fumeur, ancien fumeur, fumeur et fumeur ayant arrêté de fumer). Le nombre de cas prévalents selon un statut tabagique s'obtient en utilisant la formule ci-dessus en multipliant la probabilité P_{01}, associée au statut tabagique, par le nombre de personnes ayant ce même statut tabagique. Cela revient à multiplier la taille de la population pour une année t par la proportion du statut tabagique choisi pour cette même année.

Par exemple, si nous prenons le cas des non-fumeurs :

$$N_{prev_{non-fumeur}}(t)$$
$$= \sum_{k=a_0}^{a_{max}} p_{non-fumeur}(a_0, t - a_{max} + k)\ \nu(a_0, t - a_{max} + k)\ P_{01_{NF}}(a_0, a_{max} - k + a_0|c)$$

Avec : $p_{non-fumeur}(a_0, t - a_{max} + k)$ la proportion de non-fumeurs pour un âge a_0 pour l'année $t - a_{max} + k$ et $p_{non-fumeur}(a_0, t - a_{max} + k)\ \nu(a_0, t - a_{max} + k)$ le nombre de non-fumeurs d'âge a_0 pour l'année $t - a_{max} + k$.

5 Discussion

Dans ce travail, nous avons pris en compte le statut tabagique des individus et avons modélisé l'arrêt du tabac chez les fumeurs, nécessitant certaines hypothèses fortes (comme l'arrêt du tabac à tout âge). Nous avons fait en sorte qu'elles restent plausibles et nous permettent de ne pas complexifier la

modélisation. En effet, une des difficultés dans ce travail est de trouver un compromis entre la complexification et la faisabilité de la modélisation. Néanmoins, nous avons validé la modélisation en vérifiant que nous retrouvions les résultats de notre premier travail.
Cette modélisation permet de créer des scénarios qui simulent une diminution plus importante de la proportion de fumeurs. Nous avons choisi plusieurs scénarios comme celui du PNLT qui se fixe un objectif de moins de 22 % de fumeurs d'ici 2027 chez les 18-75 ans ou un scénario extrême abaissant la proportion de fumeurs à 0 % en une année. Ces différents scénarios permettront de quantifier l'impact de la diminution du tabagisme en France sur les indicateurs épidémiologiques de l'infarctus du myocarde en les comparant avec nos premières projections. D'autres indicateurs seront calculés comme les espérances de vie en bonne santé, en étant malade, etc. afin de quantifier, par exemple, le gain en espérance de vie en bonne santé entre plusieurs scénarios.

Références

Banks, E., Joshy, G., Weber, M. F., Liu, B., Grenfell, R., Egger, S., Paige, E., Lopez, A. D., Sitas, F., & Beral, V. (2015). Tobacco smoking and all-cause mortality in a large Australian cohort study : Findings from a mature epidemic with current low smoking prevalence. *BMC Medicine*, *13*(1), 38. https://doi.org/10.1186/s12916-015-0281-z

Baromètres de Santé publique France. (s. d.). Consulté 6 mai 2022, à l'adresse https://www.santepubliquefrance.fr/etudes-et-enquetes/barometres-de-sante-publique-france

Jha, P., Ramasundarahettige, C., Landsman, V., Rostron, B., Thun, M., Anderson, R. N., McAfee, T., & Peto, R. (2013). 21st-Century Hazards of Smoking and Benefits of Cessation in the United States. *New England Journal of Medicine*, *368*(4), 341-350. https://doi.org/10.1056/NEJMsa1211128

Lutte contre le tabagisme—Ministère des Solidarités et de la Santé. (s. d.). Consulté 01 mai 2022, à l'adresse https://solidarites-sante.gouv.fr/prevention-en-sante/addictions/article/lutte-contre-le-tabagisme

Pasquereau, A., Andler, R., Guignard, R., Soullier, N., Gautier, A., Richard, J., & Nguyen-Thanh, V. (2021). Consommation de tabac parmi les adultes en 2020 : Résultats du Baromètre de Santé publique France. *Bull Epidémiol Hebd*, *8*, 132-139.

Teo, K. K., Ounpuu, S., Hawken, S., Pandey, M., Valentin, V., Hunt, D., Diaz, R., Rashed, W., Freeman, R., Jiang, L., Zhang, X., & Yusuf, S. (2006). Tobacco use and risk of myocardial infarction in 52 countries in the INTERHEART study : A case-control study. *The Lancet*, *368*(9536), 647-658. https://doi.org/10.1016/S0140-6736(06)69249-0

Vos, T., Lim, S. S., Abbafati, C., Abbas, K. M., Abbasi, M., Abbasifard, M., Abbasi-Kangevari, M., Abbastabar, H., Abd-Allah, F., Abdelalim, A., Abdollahi, M., Abdollahpour, I., Abolhassani, H., Aboyans, V., Abrams, E. M., Abreu, L. G., Abrigo, M. R. M., Abu-Raddad,

L. J., Abushouk, A. I., … Murray, C. J. L. (2020). Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019 : A systematic analysis for the Global Burden of Disease Study 2019. *The Lancet*, *396*(10258), 1204-1222. https://doi.org/10.1016/S0140-6736(20)30925-9

Yusuf, S., Joseph, P., Rangarajan, S., Islam, S., Mente, A., Hystad, P., Brauer, M., Kutty, V. R., Gupta, R., Wielgosz, A., AlHabib, K. F., Dans, A., Lopez-Jaramillo, P., Avezum, A., Lanas, F., Oguz, A., Kruger, I. M., Diaz, R., Yusoff, K., … Dagenais, G. (2020). Modifiable risk factors, cardiovascular disease, and mortality in 155 722 individuals from 21 high-income, middle-income, and low-income countries (PURE) : A prospective cohort study. *The Lancet*, *395*(10226), 795-808. https://doi.org/10.1016/S0140-6736(19)32008-2

Planification – Plans d'expérience

Simulation d'événements rares par échantillonnage préférentiel adaptatif pour des processus de Markov déterministes par morceaux

Guillaume CHENNETIER[1,2] & Hassane CHRAIBI[1] & Anne DUTFOY[1] & Josselin GARNIER[2]

[1] *EDF Lab Paris-Saclay, Boulevard Gaspard Monge, 91120 Palaiseau, France.*
guillaume.chennetier@edf.fr ; hassane.chraibi@edf.fr ; anne.dutfoy@edf.fr
[2] *CMAP, École Polytechnique, Institut Polytechnique de Paris, 91120 Palaiseau, France.*
guillaume.chennetier@polytechnique.edu ; josselin.garnier@polytechnique.edu

Résumé. On souhaite estimer la probabilité de défaillance de systèmes industriels multi-composants hautement fiables modélisés par des processus de Markov déterministes par morceaux. Pour réduire le nombre de simulations nécessaires par rapport à une méthode de Monte-Carlo standard, nous proposons une méthode d'échantillonnage préférentiel adaptatif par entropie croisée. La paramétrisation de cette méthode, délicate pour des systèmes de grande dimension, repose sur la notion fiabiliste de "chemin minimal".

Mots-clés. PDMP, échantillonnage préférentiel, simulation d'événements rares, entropie croisée, chemins minimaux.

Abstract. We wish to estimate the failure probability of highly reliable multi-component industrial systems modeled by piecewise deterministic Markov processes. To reduce the number of simulations required compared to a standard Monte-Carlo method, we propose an adaptive importance sampling method using cross-entropy. The parameterization of this method, which is delicate for high-dimensional systems, is based on the reliability notion of "minimal path set".

Keywords. PDMP, importance sampling, rare event simulation, cross entropy, minimal path sets.

1 Introduction

Nous souhaitons évaluer la fiabilité de systèmes industriels dont la défaillance intervient lorsqu'une ou plusieurs variables physiques continues (température, pression, etc.) dépassent un seuil critique, seuil qui ne peut être atteint qu'après la détérioration de certains groupes de composants. Ces variables physiques évoluent selon des équations différentielles déterministes qui dépendent de l'état des composants du système, modifié quant à lui à la suite d'événements aléatoires discrets (pannes, réparations, mécanismes de contrôle, etc.).

L'estimation de la probabilité de défaillance du système repose sur la simulation de trajectoires de processus de Markov déterministes par morceaux, abrégés PDMP, modélisant le fonctionnement hybride du système. Ces simulations étant coûteuses et les systèmes hautement fiables, une estimation Monte-Carlo standard n'est pas envisageable. Nous présentons une méthode d'échantillonnage préférentiel adaptatif articulée en deux étapes :

1. définir une famille de distributions instrumentales à partir d'une famille d'approximations de la fonction "committor" du système. Ces approximations sont basées sur la notion fiabiliste de chemins minimaux du système (voir section 3),
2. conjointement déterminer le meilleur candidat au sein de cette famille et estimer la probabilité de défaillance à l'aide d'une procédure d'entropie croisée couplée à un schéma de recyclage des échantillons passés (voir section 4).

2 Processus de Markov déterministes par morceaux

L'état d'un PDMP à un instant t est noté $Z_t = (X_t, M_t) \in E$ où $X_t \in \mathbb{X} \subset \mathbb{R}^{d_x}$ est la position et $M_t \in \mathbb{M}$ est le mode du PDMP (avec $\mathbb{M}$ fini ou dénombrable et $E = \mathbb{X} \times \mathbb{M}$). On s'intéresse à des trajectoires $\boldsymbol{Z} := (Z_t)_{t\in[0,t_{\max}]}$ de durée $t_{\max} > 0$.

Flot. Chaque mode $m \in \mathbb{M}$ correspond à un système d'équations différentielles de solution associée ϕ_m. Entre deux sauts, le mode du PDMP reste constant et la position du PDMP suit ϕ_m. Le flot $\Phi_z = \Phi_{x,m} : t \longmapsto (\phi_m(x,t), m)$ indique donc en l'absence de saut l'état du processus après un temps t en partant d'un état z.

Sauts déterministes. Si l'espace d'états E a des frontières (notées ∂E), le processus saute lorsqu'il les atteint. On définit pour tout état de départ z le temps déterministe d'atteinte de la frontière $t_z^\partial := \inf\{t > 0 : \Phi_z(t) \in \partial E\}$.

Intensité de saut. Partant d'un état z, le processus peut également sauter à un instant aléatoire $T_z < t_z^\partial$ dont la loi dépend de l'intensité de saut λ. Cette intensité est une fonction qui associe un poids à chaque état : plus ce poids est grand, plus le processus a de chances de sauter en passant par cet état.

$$\mathbb{P}\left(T_z > t \mid Z_s = z\right) = \mathbb{1}_{t<t_z^\partial} \exp\left(-\int_0^t \lambda\left(\Phi_z(u)\right) du\right). \tag{1}$$

Noyau de saut. Connaissant l'instant du saut et l'état z^- qui précède le saut, l'état d'arrivée après le saut est choisi aléatoirement parmi les autres états par le noyau de saut $\mathcal{K}_{z^-}$ de densité $z \mapsto K(z^-, z)$ par rapport à une certaine mesure ν_{z^-}.

Densité d'une trajectoire. Une mesure dominante par rapport à laquelle une trajectoire de PDMP de durée fixée admet une densité de probabilité a été explicitée dans [1]. En notant $n_{\boldsymbol{z}}$ le nombre de saut de la trajectoire $\boldsymbol{z}$, et (t_k, z_k) les k-èmes temps de saut et état d'arrivée post-saut de la trajectoire pour $k = 1, \ldots, n_{\boldsymbol{z}}$, cette densité peut s'écrire :

$$\pi(\boldsymbol{z}) = \prod_{k=0}^{n_{\boldsymbol{z}}} \left[\lambda\left(\Phi_{z_k}(t_k)\right)\right]^{\mathbb{1}_{t_k < t^{\partial}_{z_k}}} \times \exp\left[-\int_0^{t_k} \lambda\left(\Phi_{z_k}(u)\right) \mathrm{d}u\right] \times \prod_{k=0}^{n_{\boldsymbol{z}}-1} K\left(\Phi_{z_k}(t_k), z_{k+1}\right). \quad (2)$$

On notera par la suite $\pi_{\boldsymbol{0}}$ la distribution du PDMP dont on souhaite estimer la probabilité de défaillance et $P = \mathbb{P}_{\pi_{\boldsymbol{0}}}(\boldsymbol{Z} \in \boldsymbol{D})$ cette probabilité (avec $\boldsymbol{D}$ le domaine correspondant à l'ensemble des trajectoires défaillantes).

3 Échantillonnage préférentiel pour des PDMP

L'échantillonnage préférentiel est l'une des méthodes de réduction de variance les plus populaires pour la simulation d'événements rares. On simule nos trajectoires non pas selon la distribution initiale $\pi_{\boldsymbol{0}}$ mais selon une distribution auxiliaire $\widetilde{\pi}$ plus susceptible de produire des défaillances et on rectifie le biais en introduisant le bon rapport de vraisemblance. Avec $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_N \overset{\text{i.i.d.}}{\sim} \widetilde{\pi}$ et en supposant $\mathrm{supp}(\widetilde{\pi}) \subset \mathrm{supp}(\pi_{\boldsymbol{0}})$:

$$\sum_{k=1}^{N} \mathbb{1}_{\boldsymbol{Z}_k \in \boldsymbol{D}} \frac{\pi_{\boldsymbol{0}}(\boldsymbol{Z}_k)}{\widetilde{\pi}(\boldsymbol{Z}_k)} \xrightarrow[N\to\infty]{\text{p.s.}} \mathbb{E}_{\widetilde{\pi}}\left[\mathbb{1}_{\boldsymbol{Z}\in\boldsymbol{D}} \frac{\pi_{\boldsymbol{0}}(\boldsymbol{Z})}{\widetilde{\pi}(\boldsymbol{Z})}\right] = \mathbb{E}_{\pi_{\boldsymbol{0}}}\left[\mathbb{1}_{\boldsymbol{Z}\in\boldsymbol{D}}\right] = \mathbb{P}_{\pi_{\boldsymbol{0}}}(\boldsymbol{Z} \in \boldsymbol{D}). \quad (3)$$

Processus optimal. La distribution auxiliaire optimale π_{opt} permettant de construire un estimateur de variance nulle existe et correspond à la distribution d'un PDMP dont seuls l'intensité de saut et le noyau de saut diffèrent par rapport au PDMP de distribution $\pi_{\boldsymbol{0}}$, et que l'on connaît théoriquement [1] :

$$\lambda_{\text{opt}}(\Phi_z(t)\,;\, s) = \lambda(\Phi_z(t)) \times \frac{U^-_{\text{opt}}\left(\Phi_z(t), s+t\right)}{U_{\text{opt}}\left(\Phi_z(t), s+t\right)} \quad \text{et} \quad K_{\text{opt}}\left(z^-, z\,;\, s\right) = K\left(z^-, z\right) \times \frac{U_{\text{opt}}(z,s)}{U^-_{\text{opt}}(z^-, s)}. \quad (4)$$

Ici U_{opt} et U^-_{opt} sont des fonctions dites "committor" qui mesurent respectivement la probabilité d'atteindre la défaillance avant la fin de la simulation lorsque le processus se trouve dans un état donné à un instant fixé, et lorsque le processus saute depuis un état donné à un instant fixé :

$$U_{\text{opt}}(z,s) = \mathbb{P}_{\pi_0}\left(\boldsymbol{Z} \in \boldsymbol{D} \,|\, Z_s = z\right) \quad \text{et} \quad U^-_{\text{opt}}(z^-, s) = \int_E U_{\text{opt}}(z,s)\, K\left(z^-, z\right) d\nu(z). \quad (5)$$

Ainsi la connaissance de U_{opt} est suffisante pour construire le processus optimal. Celle-ci étant hors de portée, on se donne en pratique une famille $(U_{\boldsymbol{\alpha}})_{\boldsymbol{\alpha}\in\mathbb{A}}$ d'approximations de U_{opt} où $\mathbb{A} \subset \mathbb{R}^{d_{\boldsymbol{\alpha}}}$ et on construit $(\pi_{\boldsymbol{\alpha}})_{\boldsymbol{\alpha}\in\mathbb{A}}$ des distributions auxiliaires sur le modèle de π_{opt} mais où $U_{\boldsymbol{\alpha}}$ remplace U_{opt}.

Approximation de la fonction committor basée sur les chemins minimaux. On cherche une famille de fonctions $(U_{\boldsymbol{\alpha}})_{\boldsymbol{\alpha}\in\mathbb{A}}$ qui se comportent comme U_{opt} à une constante multiplicative près. Dans le cas de systèmes multi-composants, de telles fonctions doivent pénaliser les réparations et favoriser les pannes sur les composants clés.

Un *chemin* du système est un groupe de composants dont le fonctionnement empêche la défaillance du système quel que soit l'état des autres composants. Un chemin est dit *minimal* s'il ne contient pas d'autre chemin, on abrège MPS pour *minimal path set*. Si le système est *cohérent* (une panne ne peut pas empêcher la défaillance du système et une réparation ne peut pas la causer), alors il existe un nombre $d_{\text{MPS}} > 0$ de MPS et ceux-ci sont uniques. La défaillance du système n'est alors possible que lorsque $\beta^{(\text{MPS})}(z) = d_{\text{MPS}}$ où $\beta^{(\text{MPS})}(z)$ est le nombre de MPS avec au moins un composant cassé. Un bon choix de $U_{\boldsymbol{\alpha}}$ devrait donc être croissant en $\beta^{(\text{MPS})}$. Nous proposons la forme suivante :

$$U_{\boldsymbol{\alpha}}^{(\text{MPS})}(z,s) = \exp\Big\{\Big(\sum_{i=1}^{d_{\text{MPS}}} \alpha_i \mathbb{1}_{\beta^{(\text{MPS})}(z)\geq i}\Big)^2\Big\}, \qquad \boldsymbol{\alpha} = (\alpha_i)_{i=1}^{d_{\text{MPS}}}. \tag{6}$$

Quand d_{MPS} est grand, on peut imposer l'égalité de certaines coordonnées de $\boldsymbol{\alpha}$ pour réduire sa dimension effective et simplifier la recherche du meilleur candidat. La forme $x \mapsto \exp(x^2)$ garantit la croissance stricte des rapports $U_{\boldsymbol{\alpha}}^{-}/U_{\boldsymbol{\alpha}}$ en $\beta^{(\text{MPS})}$ et permet ainsi les pannes en cascade que l'on observe dans des trajectoires défaillantes réalistes.

4 Entropie croisée avec schéma de recyclage

Une fois choisie la famille paramétrique $(U_{\boldsymbol{\alpha}})_{\boldsymbol{\alpha}\in\mathbb{A}}$ et la famille de lois associées $(\pi_{\boldsymbol{\alpha}})_{\boldsymbol{\alpha}\in\mathbb{A}}$, il reste à déterminer le meilleur candidat au sein de cette famille.

Entropie croisée. En suivant l'idée de [5], on minimise la divergence de Kullback-Leibler entre π_{opt} et $\pi_{\boldsymbol{\alpha}}$:

$$\arg\min_{\boldsymbol{\alpha}\in\mathbb{A}} \mathcal{D}_{\text{KL}}\left(\pi_{\text{opt}} \| \pi_{\boldsymbol{\alpha}}\right) = \arg\min_{\boldsymbol{\alpha}\in\mathbb{A}} \mathbb{E}_{\pi_{\mathbf{0}}}\left[-\mathbb{1}_{\boldsymbol{z}\in\boldsymbol{D}} \log\left(\pi_{\boldsymbol{\alpha}}\left(\boldsymbol{Z}\right)\right)\right]. \tag{7}$$

La méthode est séquentielle : à l'itération q on simule n_q nouvelles trajectoires $\left(\boldsymbol{Z}_k^{(q)}\right)_{k=1}^{n_q}$ selon la loi $\pi_{\boldsymbol{\alpha}^{(q)}}$ pour former un total de $N_q = \sum_{r=1}^{q} n_r$ trajectoires. On détermine le prochain itéré $\boldsymbol{\alpha}^{(q+1)}$ en minimisant une estimation de (7) par échantillonnage préférentiel.

$$\boldsymbol{\alpha}^{(q+1)} = \arg\min_{\boldsymbol{\alpha}\in\mathbb{A}} \left\{ \sum_{r=1}^{q}\sum_{k=1}^{n_r} \mathbb{1}_{\boldsymbol{Z}_k^{(r)}\in\boldsymbol{D}} \frac{\pi_0\left(\boldsymbol{Z}_k^{(r)}\right)}{\pi_{\boldsymbol{\alpha}^{(r)}}\left(\boldsymbol{Z}_k^{(r)}\right)} \log\left[\pi_{\boldsymbol{\alpha}}\left(\boldsymbol{Z}_k^{(r)}\right)\right] \right\}. \tag{8}$$

À l'itération finale Q, on réutilise toutes les trajectoires tirées pour estimer P :

$$\widehat{P}_{N_Q} = \frac{1}{N_Q} \sum_{q=1}^{Q}\sum_{k=1}^{n_q} \mathbb{1}_{\boldsymbol{Z}_k^{(q)}\in\boldsymbol{D}} \frac{\pi_0\left(\boldsymbol{Z}_k^{(q)}\right)}{\pi_{\boldsymbol{\alpha}^{(q)}}\left(\boldsymbol{Z}_k^{(q)}\right)}, \tag{9}$$

Optimalité asymptotique. En utilisant [4], nous avons déterminé des critères suffisants pour assurer la consistance et la normalité asymptotique de l'estimateur (9).

Théorème 1. *Si $\mathbb{A}$ est compact, si $\boldsymbol{\alpha}_{opt} \in \mathbb{A}$ est l'unique minimiseur de (7) et si de plus :*

1. *les fonctions λ, K, et $(U_{\boldsymbol{\alpha}})_{\boldsymbol{\alpha}\in\mathbb{A}}$ sont bornées inférieurement et supérieurement sur leur support par des constantes strictement positives,*
2. *il existe $t_\varepsilon > 0$ tel que $t_z^\partial \geq t_\varepsilon$ pour tout $z^- \in \partial E$ et tout $z \in supp\, K(z^-, \cdot)$,*

alors, en notant $V(\boldsymbol{\alpha}) = \mathbb{E}_{\pi_{\mathbf{0}}}\left[\mathbb{1}_{\boldsymbol{Z}\in\boldsymbol{D}}\frac{\pi_{\mathbf{0}}(\boldsymbol{Z})}{\pi_{\boldsymbol{\alpha}}(\boldsymbol{Z})}\right] - P^2$, on a :

$$\boldsymbol{\alpha}^{(Q)} \xrightarrow[Q\to\infty]{p.s} \boldsymbol{\alpha}_{opt} \quad et \quad \sqrt{N_Q}\left(\widehat{P}_{N_Q} - P\right) \xrightarrow[Q\to\infty]{\mathcal{L}} \mathcal{N}\left(0, V\left(\boldsymbol{\alpha}_{opt}\right)\right). \tag{10}$$

Intervalle de confiance asymptotique de niveau $1-a$ pour P. En notant $q_{1-a/2}$ le quantile d'ordre $1-a/2$ de la loi $\mathcal{N}(0,1)$, et $\widehat{\sigma}^2_{N_Q} = \frac{1}{N_Q}\sum_{q=1}^{Q}\sum_{k=1}^{n_q}\mathbb{1}_{\boldsymbol{Z}_k^{(q)}\in\boldsymbol{D}}\frac{\pi_0\left(\boldsymbol{Z}_k^{(q)}\right)^2}{\pi_{\boldsymbol{\alpha}^{(q)}}\left(\boldsymbol{Z}_k^{(q)}\right)^2} - \widehat{P}^2_{N_Q}$ l'estimateur de la variance asymptotique $V(\boldsymbol{\alpha}_{\mathrm{opt}})$, on a :

$$\mathbb{P}\left(P \in \left[\widehat{P}_{N_Q} - q_{1-\alpha/2}\sqrt{\frac{\widehat{\sigma}^2_{N_Q}}{N_Q}}\,;\, \widehat{P}_{N_Q} + q_{1-\alpha/2}\sqrt{\frac{\widehat{\sigma}^2_{N_Q}}{N_Q}}\right]\right) \xrightarrow[Q\to\infty]{} 1-a. \tag{11}$$

5 Résultats numériques

La méthode a été testée sur un système décrivant l'évolution d'une piscine de combustible nucléaire usagé inspiré de [2] (voir Figure 1).

Méthode	N	$\widehat{P}$	Intervalle de confiance à 95%
CMC	10^6	2×10^{-6}	$[-7.72\times10^{-7}\,;\,4.78\times10^{-6}]$
CMC	10^7	3.9×10^{-6}	$[2.67\times10^{-6}\,;\,5.12\times10^{-6}]$
CMC	10^8	3.50×10^{-6}	$[3.13\times10^{-6}\,;\,3.87\times10^{-6}]$
IS	10^2	3.15×10^{-6}	$[2.07\times10^{-6}\,;\,4.23\times10^{-6}]$
IS	10^3	3.17×10^{-6}	$[2.76\times10^{-6}\,;\,3.58\times10^{-6}]$
IS	10^4	3.39×10^{-6}	$[3.15\times10^{-6}\,;\,3.62\times10^{-6}]$

TABLE 1 – Monte-Carlo standard (CMC) versus échantillonnage préférentiel (IS).

L'eau de la piscine est chauffée par le combustible entreposé au fond. La défaillance a lieu lorsqu'un volume critique d'eau s'est évaporé. Pour l'éviter, on transfère la température d'une source d'eau froide à l'eau de la piscine via des circuits étanches reliés par des

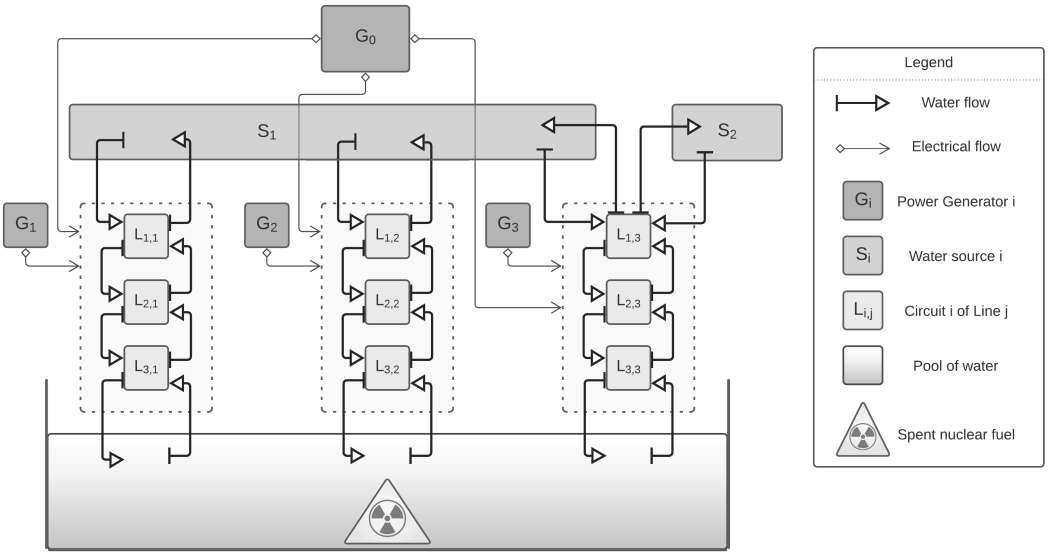


FIGURE 1 – Représentation de la piscine de combustible usagé.

échangeurs de chaleur formant une ligne reliée à une alimentation électrique générale. En secours, le système dispose de deux autres lignes, d'un générateur électrique par ligne et d'une seconde source accessible pour la troisième ligne. Le tableau 1 montre que notre méthode (IS) fait aussi bien sur ce cas test qu'une approche Monte-Carlo standard (CMC) avec 10^4 fois moins de simulations.

Références

[1] H. Chraibi, A. Dutfoy, T. Galtier et J. Garnier (2019), On the optimal importance process for piecewise deterministic Markov process. *ESAIM : PS. 23 :893–921.*

[2] H. Chraibi, J.-C. Houdebine et A. Sibler (2016), PyCATSHOO : Toward a new platform dedicated to dynamic reliability assessments of hybrid systems. *Proceedings PSAM13.*

[3] M. H. A. Davis (1984), Piecewise-Deterministic Markov Processes : A general class of non-diffusion stochastic models. *J. R. Stat. Soc. B 46 :353-388.*

[4] B. Delyon et F. Portier (2018), Asymptotic optimality of adaptive importance sampling. *NIPS 2018.*

[5] R. Rubinstein (1999), The Cross-Entropy method for combinatorial and continuous optimization. *Methodology and Computing in Applied Probability 1 :127–190.*

Estimation de courbes de fragilité sismique par planification séquentielle d'expériences.

Clément Gauchy [12] & Cyril Feau [1] & Josselin Garnier [2]

[1] *Universite Paris-Saclay, CEA, Service d'Études Mécaniques et Thermiques, 91191, Gif-sur-Yvette, France, E-mail: clement.gauchy@cea.fr, cyril.feau@cea.fr*
[2] *CMAP, Ecole Polytechnique, Institut Polytechnique de Paris, 91128 Palaiseau Cedex, France, E-mail: josselin.garnier@polytechnique.edu*

Résumé. Les études probabilistes de sûreté sismique consistent à évaluer les probabilités de défaillance de structures mécaniques soumises à des excitations sismiques. Ces études nécessitent l'estimation de courbes de fragilité sismique, qui sont la probabilité de défaillance de la structure conditionnellement à une mesure d'intensité du signal sismique. Cependant, leur estimation requiert de nombreuses expériences numériques qui peuvent être très coûteuses en temps de calcul, ce qui rend l'estimation par une méthode Monte Carlo inappropriée. Nous proposons dans ce papier de construire un algorithme de planification séquentielle d'expériences en supposant un *a priori* de processus Gaussien sur la réponse du code de calcul mécanique.

Mots-clés. processus Gaussien, planification séquentielle, courbes de fragilité.

Abstract. Seismic probabilistic risk assessment studies consist in evaluating the probabilities of failure of mechanical structures when submitted to seismic ground motions. These studies are often concentrated on fragility curve estimation. The fragility curve is the probability of failure of the structure conditionally to a seismic intensity measure. However, its estimation requires computer experiments involving huge computation time. Such a computational burden makes crude Monte Carlo methods untractable, fragility curves estimation must then be economical in terms of sample size. We propose an algorithm of sequential planning of experiments by supposing a Gaussian process *prior* on the output of the mechanical computer model.

Keywords. Gaussian process, sequential planning of experiments, fragility curves.

1 Introduction

The estimation of so-called fragility curves is a crucial part of seismic probabilistic risk assessment (SPRA) or probabilistic based earthquake engineering (PBEE). The fragility curve is a way to evaluate structural reliability using a sample of seismic ground motions, which are characterized by their intensity measure (IM) (e.g maximum acceleration of the seimic signal given a time frame). It consists in the conditional probability $\mathbb{P}(Y > C|\mathrm{IM} = a)$ that a specific mechanical demand Y exceeds a threshold of acceptable structural behaviour C for a given seismic intensity of value a. The mechanical demand Y often

needs time consuming computation using complex computer codes, making the estimation of fragility curves intractable. This article proposes a Bayesian algorithm of sequential design of experiments, inspired from Bayesian algorithms of the same kind for probability of failure estimation [1] or quantile estimation [2]. The proposed Bayesian algorithm for sequential design of experiments is presented in Section 2. The Section 3 details the practical implementation of the algorithm. Finally, Section 4 presents the performance of the algorithm on a real industrial application concerning a simplified part of a piping system of a French Pressurized Water Reactor.

2 Bayesian decision theory framework

Let $\mathcal{X}$ be a compact set of $\mathbb{R}$ and X an $\mathcal{X}$-valued random variable, in SPRA studies $X = \log \mathrm{IM}$. We define the following nonparametric regression model linking the seismic intensity $x \in \mathcal{X}$ to the log structural mechanical demand $y(x)$:

$$y(x) = g(x) + \varepsilon \ , \tag{1}$$

where ε is a Gaussian homoscedastic noise such that $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2)$. The usual regression model used in SPRA studies is the log linear model [3], we keep the homoscedasticity assumption but we propose a non-parametric regression function g instead of a linear regression. The problem to be considered here is to estimate the *fragility curve* expressed as

$$\Psi(x; g) = \mathbb{P}_\varepsilon \left(y(X) > C | X = x \right) = \Phi \left(\frac{g(x) - C}{\sigma_\varepsilon} \right) \ , \tag{2}$$

which corresponds to the probability of exceedance of the level C given an input parameter x. The noise variance σ_ε^2 is supposed known while g is unknown. Based on a sequential budget of evaluation points $\mathcal{D}_N = (X_i)_{1 \leq i \leq N}$ with corresponding observed values $(y(X_i))_{1 \leq i \leq N}$, it is possible to provide an estimator $\widehat{\Psi}_N$ of Ψ. After choosing an estimator $\widehat{\Psi}_N$, one has to focus on the good choices of a sequential strategy of evaluation $(X_i)_{1 \leq i \leq N}$. For that matter, we follow the lines of [1] to propose a Bayesian decision-theoretic framework to solve the decision problem of choosing $\mathcal{D}_N$. Given a loss function $\ell : \ \mathbb{R} \times \mathbb{R} \to \mathbb{R}^+$, we measure the quality of the design of experiments $\mathcal{D}_N$ by the approximation error $e(\mathcal{D}_N, g) = \ell(\Psi, \widehat{\Psi}_N)$. In this paper we consider the Integrated Squared Error (ISE):

$$e(\mathcal{D}_N, g) = \int_{\mathcal{X}} (\Psi(t; g) - \widehat{\Psi}_N(t))^2 \eta(t) dt \ , \tag{3}$$

where η is the pdf of a probability measure with respect to the Lebesgue measure on $\mathcal{X}$. We consider a Bayesian framework to model the uncertainty about the regression function g, we assume that g is a realization of a real-valued Gaussian process G defined on a probability space $(\Omega, \mathcal{B}, \mathbb{P}_0)$. We define the random observation tainted by the Gaussian

process uncertainty

$$Y(x) = G(x) + \varepsilon \ . \tag{4}$$

Denote by $\mathcal{F}_n$ the σ-algebra generated by $(X_i, Y(X_i))_{1\leq i\leq n}$ for $1 \leq n \leq N$. Given a Gaussian process G and a set of evaluation points $\mathcal{D}_N$, the optimal estimator $\widehat{\Psi}_n$ that minimizes $\mathbb{E}_{\mathbb{P}_0}\left[\int_{\mathcal{X}}(\Psi(t;G) - \widehat{\Psi}_n(t))^2\eta(t)dt\middle|\mathcal{F}_n\right]$ among the $\mathcal{F}_n$ - measurable estimators is:

$$\widehat{\Psi}_n(t) = \mathbb{E}_{\mathbb{P}_0}\left[\Psi(t;G)\middle|\mathcal{F}_n\right] = \Phi\left(\frac{\widehat{G}_n(t) - C}{\sigma_n(t)}\right) \ , \tag{5}$$

where $\sigma_n(x)^2 = \widehat{\sigma}_n(x)^2 + \sigma_\varepsilon^2$ and $\mathbb{E}_{\mathbb{P}_0}$ is the expectation with respect to $\mathbb{P}_0$, given that $(G(t)|\mathcal{F}_n) \sim \mathcal{N}(\widehat{G}_n(t), \widehat{\sigma}_n(t)^2)$ using kriging equations. We choose the optimal strategy $\mathcal{D}_N^* = (X_i^*)_{1\leq i\leq N}$ as the minimizer of the *Bayes risk*:

$$\mathcal{D}_N^* = \underset{\mathcal{D}_N\in\mathcal{X}^N}{\operatorname{argmin}}\, \mathbb{E}_{\mathbb{P}_0}[e(\mathcal{D}_N, G)] \ , \tag{6}$$

The optimal strategy can be formally obtained. We define $R_N = \mathbb{E}_{\mathbb{P}_0}\left[\int_{\mathcal{X}}(\Psi(t;G) - \widehat{\Psi}_N(t))^2\eta(t)dt\middle|\mathcal{F}_N\right]$ the terminal risk and using backward induction:

$$R_n = \min_{x\in\mathcal{X}} \mathbb{E}_{\mathbb{P}_0}\left[R_{n+1}|X_{n+1} = x, \mathcal{F}_n\right] \ , \tag{7}$$

$$X_{n+1}^* = \underset{x\in\mathcal{X}}{\operatorname{argmin}}\, \mathbb{E}_{\mathbb{P}_0}\left[R_{n+1}|X_{n+1} = x, \mathcal{F}_n\right] \ , \tag{8}$$

for $0 \leq n \leq N-1$. However, the optimal evaluation points for a finite horizon N suffers from the curse of dimensionality. So, in the rest of the paper, we build a sub-optimal set of evaluation points by resolving the optimization problem (8) for $N = 1$. We thus define a $\mathcal{F}_n$-measurable *sampling criterion* $J_n(x)$:

$$J_n(x) = \mathbb{E}_{\mathbb{P}_0}\left[\int_{\mathcal{X}}(\Psi(t;G) - \widehat{\Psi}_{n+1}(t))^2\eta(t)dt\middle|X_{n+1} = x, \mathcal{F}_n\right] \ . \tag{9}$$

We define the *Stepwise Uncertainty Reduction* (SUR) - based set of evaluation points $\mathcal{D}_N^{\mathrm{SUR}} = (X_i)_{1\leq i\leq N}$ such that for $1 \leq n \leq N-1$:

$$X_{n+1}^{\mathrm{SUR}} = \underset{x\in\mathcal{X}}{\operatorname{argmin}}\, J_n(x) \ . \tag{10}$$

3 Practical computation of the sampling criterion

This section is devoted to derive a numerical approximation of the sampling criterion $J_n(x)$ defined in Equation 9. First of all remark that

$$\begin{aligned} J_n(x) &= \int_{\mathcal{X}} \mathbb{E}_{\mathbb{P}_0}\left[\mathbb{E}_{\mathbb{P}_0}\left[\left(\Psi(t;G)-\widehat{\Psi}_{n+1}(t)\right)^2\middle|\mathcal{F}_{n+1}\right]\middle|X_{n+1}=x,\mathcal{F}_n\right]\eta(t)dt \\ &= \int_{\mathcal{X}} \mathbb{E}_{\mathbb{P}_0}\left[\mathbb{E}_{\mathbb{P}_0}\left[\Psi(t;G)^2-\widehat{\Psi}_{n+1}(t)^2\middle|\mathcal{F}_{n+1}\right]\middle|X_{n+1}=x,\mathcal{F}_n\right]\eta(t)dt \\ &= \int_{\mathcal{X}} \mathbb{E}_{\mathbb{P}_0}\left[\Psi(t;G)^2\middle|X_{n+1}=x,\mathcal{F}_n\right]-\mathbb{E}_{\mathbb{P}_0}\left[\widehat{\Psi}_{n+1}(t)^2\middle|X_{n+1}=x,\mathcal{F}_n\right]\eta(t)dt \end{aligned}$$

We can now precise the two expectations inside the integrands: For the first term, we have

$$\mathbb{E}_{\mathbb{P}_0}\left[\Psi(t;G)^2\middle|X_{n+1}=x,\mathcal{F}_n\right]=\mathbb{E}_{Z\sim\mathcal{N}(\widehat{G}_n(t),\widehat{\sigma}_n(t)^2)}\left[\Phi\left(\frac{Z-C}{\sigma_\varepsilon}\right)^2\right] \tag{11}$$

For the second term, we have

$$\mathbb{E}_{\mathbb{P}_0}\left[\widehat{\Psi}_{n+1}(t)^2\middle|X_{n+1}=x,\mathcal{F}_n\right]=\mathbb{E}_{Z\sim\mathcal{N}(\widehat{G}_n(x),\widehat{\sigma}_n(x)^2)}\left[\Phi\left(\frac{\widehat{G}_{n+1}(t;Z)-C}{\sigma_{n+1}(t)}\right)^2\right] \tag{12}$$

where $\widehat{G}_{n+1}(t;Z)$ and $\widehat{\sigma}_{n+1}(t)^2$ are respectively the conditional mean and variance of the GP G knowing the observations $(X_i,Y(X_i))_{1\leq i\leq n}$ and the virtual output Z at point $X_{n+1}=x$. Remark that $\widehat{G}_{n+1}$ is an affine function in Z and $\widehat{\sigma}_{n+1}$ depends only on $(X_i)_{1\leq i\leq n+1}$. The two expectations can be approximated using Gauss-Hermite quadrature (for the heat kernel e^{-x^2}) with Q points with quadrature points $(u_q)_{1\leq q\leq Q}$ and quadrature weights $(\omega_q)_{1\leq i\leq Q}$. The integral with respect to η can be approximated by a Riemann sum with regular grid $(t_i)_{1\leq i\leq T}$ of interval length ΔT (Gaussian quadrature could be used as well):

$$\begin{cases} J_n(x) & \approx \frac{1}{\sqrt{\pi}}\sum_{i=1}^{T}\sum_{q=1}^{Q}\Delta T\omega_q\left(\Phi\left(\frac{z_{n,q}(t_i)-C}{\sigma_\varepsilon}\right)^2-\Phi\left(\frac{\widehat{G}_{n+1}(t_i;z_{n,q}(x))-C}{\sigma_{n+1}(t_i)}\right)^2\right)\eta(t_i) \\ z_{n,q}(x) & = \widehat{G}_n(x)+\widehat{\sigma}_n(x)u_q\sqrt{2} \end{cases} \tag{13}$$

4 Industrial application: safety water pipe of a French Pressurized Reactor

The following test case corresponds to a piping system which is a simplified part of a secondary line of a French Pressurized Water Reactor [4]. The structural mechanical

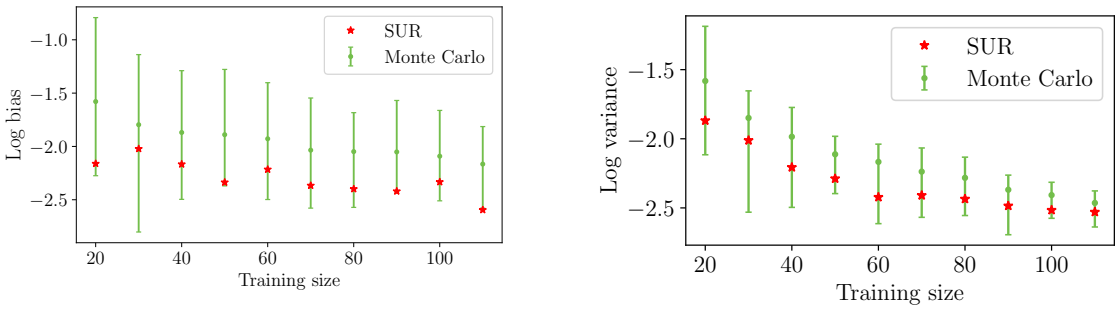
demand in this case is the out-of-plane rotation R of a specific elbow. A single degree of freedom nonlinear oscillator is calibrated on the Finite Element Model of the pipe implemented in CAST3M [5] and is used to validate the proposed sequential planning of experiments. The synthetic seismic signals dataset – developed and used in [6] – is filtered by a fictitious linear single-mode building at 5 Hz and damped at 2 %. The fragility curve of the piping system is here expressed as a function of the pseudo-spectral acceleration of the initial set of the synthetic signals (i.e not filtered signals), calculated at 5 Hz (first eigenfrequency of the pipe) and 1 % damping ratio. The rotation threshold considered here is $C = 1.8°$. This is the 90 %-level quantile of the full dataset of simulations based on 10^5 synthetic seismic ground motions. The Gaussian process used has a zero mean function and stationary Matèrn 5/2 covariance function. The performance of the proposed sequential planning strategy is evaluated by comparing the estimator $\widehat{\Psi}_n$ with a non-parametric estimation of the fragility curve Ψ_{ref} with the full dataset of 10^5 synthetic signals (see e.g. [6]). We compute the *bias* b_n and the *posterior variance* v_n of a design $\mathcal{D}_n$ by:

$$b_n = \int (\widehat{\Psi}_n(t) - \Psi_{\text{ref}}(t))^2 \eta(t) dt \ , \quad v_n = \mathbb{E}_{\mathbb{P}_0} \left[\int (\Psi(t;G) - \widehat{\Psi}_n(t))^2 \eta(t) dt \middle| \mathcal{F}_n \right] \ . \tag{14}$$

For this case we choose $\eta(t) = \mathbb{1}_{[a,b]}(t)$ ($a = 3,1\ m/s^2,\ b = 31,2\ m/s^2$). The SUR procedure starts by choosing 10 synthetic seismic ground motions in a stratified manner, by dividing the full dataset into 10 partitions of same size using the $(j/10)_{1\leq j\leq 9}$-level quantiles of the seismic intensity measure. They are the first points of the design set and are used to optimize the Gaussian process hyperparameters (including σ_ε). After that, we propose a brute force algorithm to solve the optimization problem (10). A set of candidate seismic signals with log intensity measure $(U_i)_{1\leq i\leq\ m}$ ($m = 200$) are chosen in the dataset of 10^5 synthetic seismic ground motions. The candidate set is built by using the same partition into 10 subdomains as for initialization. 20 seismic signals are chosen at random for each subdomain. We then evaluate the sampling criterion using Eq. (13) for each seismic signal in the set of candidates. We then take the seismic signal that minimizes the criterion $i^* = \text{argmin}_{1\leq i\leq\ m} J_{n-1}(U_i), \quad X_n = U_{i^*}$ and then we add X_n to our design $\mathcal{D}_n$ of size n and compute the log mechanical demand $y(X_n)$ using our computer code. The Gaussian process hyperparameters are optimized every 10 iterations.

Figure 1 compares the biases b_n and variances v_n as a function of the training size n for the SUR strategy and for 100 replications of Monte-Carlo sampling in the 10^5-sized dataset of seismic signals. The SUR strategy offers good performance in terms of bias between the estimated fragility curve $\widehat{\Psi}_n$ and the reference fragility curve Ψ_{ref} and is equivalent to Monte-Carlo sampling in terms of variance. These results motivate further research in this topic.

Acknowledgments: Clément Gauchy would like to thank Vincent Chabridon and Elias Fekhari for fruitful discussions about Bayesian decision theory.



(a) Log bias

(b) Log variance

Figure 1: Comparison of the bias b_n and variance v_n (in log) between Monte-Carlo sampling and our SUR strategy. The green error bars correspond to the interval between the 10% and 90% level quantile of 100 replications of Monte Carlo sampling.

References

[1] Julien Bect, David Ginsbourger, Ling Li, Victor Picheny, and Emmanuel Vazquez. Sequential design of computer experiments for the estimation of a probability of failure. *Statistics and Computing*, 22(3):773–793, April 2011. doi: 10.1007/s11222-011-9241-4.

[2] Aurélie Arnaud, Julien Bect, Mathieu Couplet, Alberto Pasanisi, and Emmanuel Vazquez. Évaluation d'un risque d'inondation fluviale par planification séquentielle d'expériences. *42èmes Journées de Statistique*, 2010.

[3] Bruce R. Ellingwood and Kursat Kinali. Quantifying and communicating uncertainty in seismic risk assessment. *Structural Safety*, 31(2):179–187, 2009. doi: https://doi.org/10.1016/j.strusafe.2008.06.001. Risk Acceptance and Risk Communication.

[4] F. Touboul, P. Sollogoub, and N. Blay. Seismic behaviour of piping systems with and without defects: experimental and numerical evaluations. *Nuclear Engineering and Design*, 192(2):243 – 260, 1999.

[5] URL `http://www-cast3m.cea.fr/`.

[6] Rémi Sainct, Cyril Feau, Jean-Marc Martinez, and Josselin Garnier. Efficient methodology for seismic fragility curves estimation by active learning on support vector machines. *Structural Safety*, 86:101972, 2020.

Design of Experiment for Bayesian transferred model

Loïc Iapteff [1], Julien Jacques [2], Benoit Celse [3] & Victor Costa [4]

[1] *loic.iapteff @ifpen.fr, Rond-point de l'échangeur de Solaize, 69360 Solaize*
[2] *julien.jacques @univ-lyon2.fr, 5 Avenue Pierre Mendès France, 69500 Bron*
[3] *benoit.celse @ifpen.fr, Rond-point de l'échangeur de Solaize, 69360 Solaize*
[4] *victor.costa @ifpen.fr, Rond-point de l'échangeur de Solaize, 69360 Solaize*

Résumé. Le groupe IFP commercialise des catalyseurs et doit s'engager sur leur performance. Il est donc nécessaire de disposer de modèles prédictifs fiables pour chaque nouvelle génération de catalyseurs. Ces modèles sont construits à partir de données expérimentales très couteuses. Afin d'optimiser les coûts, notre ambition est de réduire le nombre d'expérimentations nécessaires pour estimer un modèle associé à un nouveau type de catalyseur. De précédents travaux ont montré qu'une approche de transfert Bayésien permettait d'améliorer la qualité des modèles lorsque le nombre d'observations est réduit. Dans cet article, les plans d'expériences sont étudiés afin de déterminer comment sélectionner ce nombre réduit d'observations permettant d'obtenir les meilleurs modèles par transfert Bayésien. Cet article montre que l'algorithme Kennard and Stone peut, sous certaines conditions, offrir de meilleurs résultats que des plans optimaux.

Mots-clés. Plan d'expérience, Transfer learning, inférence bayésienne

Abstract. IFP group develops catalysts and has to guarantee their performances. It is therefore crucial to have good predictive models for all new catalysts. These models are built upon very expensive experimental data. In order to minimize costs, we aim at reducing the number of new data points to measure to fit a model on the new catalyst. Previous work has shown that a Bayesian transfer approach can improve the quality of models when the number of observations is reduced. In this paper, experimental designs are studied in order to determine how to select this reduced number of observations to obtain the best models by Bayesian transfer. This article shows that the Kennard and Stone algorithm can, under certain conditions, offer better results than optimal designs.

Keywords. Design of Experiment, Transfer learning, Bayesian inference

1 The challenge

IFP Group develops and sells catalysts to chemical and biochemical producers. Catalysts are solids that make the reaction feasible, faster, and/or at a lower temperature and pressure. Performance must be guaranteed and it is therefore crucial to have good predictive

models for all new catalysts. These models are built on very expensive experimental data, therefore the aim is to use a small number of data and select them carefully to build a good model.

For the modeling of hydrocracking process, previous work has shown that the Bayesian transfer approach significantly reduces the number of points needed to obtain robust models (Iapteff et al. 2020). The method consists in estimating a model for a new catalyst using Bayesian models, whose priors depend on the old catalyst generation, for which important amounts of data are available. Celse, J.-J. D. Costa, and V. Costa 2016 shows the effectiveness of D-optimal designs for modeling the hydrocracking process using the kinetic model.

This paper focuses on hydrocracking modeling using Bayesian linear model (1):

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{\beta X} + \boldsymbol{\epsilon}, \\ \boldsymbol{\epsilon} &\sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}), \\ \boldsymbol{\beta} &\sim \pi(\cdot), \end{aligned} \tag{1}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$ is the vector of model parameters, assumed to be a random variable of prior density $\pi(\cdot)$, $\boldsymbol{X}$ is the matrix of observations and $\boldsymbol{y}$ is the variable of interest. For the experiments, an industrial dataset from hydrocracking facilities is available. The prior $\pi(\cdot)$ is built with an old dataset that uses another catalyst.

The goal of this work is to show that when the sample size is not too small, basic geometric experimental design offer better results than optimal designs. This paper is composed of two parts: a first section presents the Design of Experiment (DoE) and a second section shows the results on our industrial dataset, using some of the most commonly employed DoE approaches. This section also explains why the results are not necessarily the best when using optimal designs.

2 Design of Experiment

The purpose of the Design of Experiments (DoE) is to organise the experiments carried out for scientific research or industrial studies in the best possible way. When experiments are used to study the relation between a quantity of interest and other features, DoE aims to maximise the information on this quantity from a reduced amount of data. The benefits of DoE are multiple, for instance it allows to control bias, to reduce random variation, to increase the precision of parameter estimates, to improve predictions of future observations, or to make a choice between competing models. In this section, we focus on most common DoE approaches for linear models.

2.1 Model independent approaches

The Kennard and Stone algorithm (Kennard and Stone 1969) is an iterative algorithm for selecting the best observations regardless of the type of model. This algorithm is a sequential construction for a minimax space-filling design and is based on a set of possible observations in order to extract the optimal points. It is initialized with a training dataset composed of the two most distant points of the complete dataset and a candidate set composed of the remaining points. Then, at each iteration, one observation $\boldsymbol{x}_i$ is moved from the candidate set to the training set such as $\boldsymbol{x}_i = \underset{\boldsymbol{x}_i \in \text{Candidate}}{argmax} \left(\underset{\boldsymbol{x}_{i'} \in \text{Training}}{min} dist(\boldsymbol{x}_i, \boldsymbol{x}_{i'}) \right)$ until the training dataset reaches a predefined size. Notice that in the original Kennard and Stone algorithm, the euclidean distance is considered. With such an approach, we try to obtain an uniform distribution over the features space for the training set.

Other DoE model-independent approaches are well studied, such as modified Kennard and Stone algorithm, Latin hypercube sampling or full factorial designs but are not developped in this paper. The model independent DoE approaches are also often used to initialise algorithms for obtaining optimal designs, allowing faster convergence.

2.2 Optimal DoE for Linear Model

In this section, DoE approaches for the linear model (1) is considered. For this model, the most commonly used criteria are the $\mathcal{D}/\mathcal{A}/\mathcal{E}/\mathcal{G}$-optimal designs (Fedorov 1972, De Aguiar et al. 1995, Wong 1994). When the objective is to improve the prediction's quality of the model, a popular criterion is the $\mathcal{G}$-optimal criterion, which aims to minimize the maximum prediction uncertainty: the design must minimize $\underset{\boldsymbol{x}_i}{max}(\boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{X})^{-1})\boldsymbol{x}_i)$. The $\mathcal{D}$, $\mathcal{A}$ and $\mathcal{E}$ optimal criteria use the dispersion matrix $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ which is proportional to the covariance matrix of parameters. The idea is that minimizing the uncertainty in the parameters helps to minimize the uncertainty in prediction.

The $\mathcal{D}$-optimal design is the design that minimizes $det((\boldsymbol{X}^T\boldsymbol{X})^{-1})$. It is the most widely used criterion due to its efficiency and ease of implementation. The Fedorov's algorithm (Fedorov 1972) gives an exact $\mathcal{D}$-optimal design with an easy implementation. This algorithm ensures convergence to a local minimum, so different initializations can be considered. An usual initialization is to take a design obtained via the Kennard and Stone algorithm.

The $\mathcal{A}$-optimal design and $\mathcal{E}$-optimal design minimize the trace of $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ and the largest eigenvalue of $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$, respectively.

2.3 Bayesian Experimental Design

Since the objective of DoE is often to correctly estimate the model parameters, it can be interesting to provide a prior information on the value of these parameters. This approach is called Bayesian Experimental Design (Chaloner and Verdinelli 1995) and also has the

advantage that a single observation can be sufficient to build the model, the information matrix being always non-singular with a proper prior. Bayesian Experimental Design is commonly based on an utility function $U(\boldsymbol{d}, \boldsymbol{\theta}, \boldsymbol{y})$, defining the contribution of the choice of design $\boldsymbol{d}$, yielding the $\boldsymbol{y}$-values for the dependent variable and with the model's parameters $\boldsymbol{\theta}$. The optimal design $\boldsymbol{d}_{opt}$ is thus obtained by maximizing the maximum expected utility:

$$\boldsymbol{d}_{opt} = \underset{d}{argmax} \int_{\mathcal{Y}} \int_{\Theta} U(\boldsymbol{d}, \boldsymbol{\theta}, \boldsymbol{y}) p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{d}) p(\boldsymbol{\theta}) d\boldsymbol{\theta} d\boldsymbol{y}, \tag{2}$$

where $p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{d})$ is the likelihood of the model and $p(\boldsymbol{\theta})$ the prior information, which is assumed not to be affected by the choice of the design.

In the Bayesian linear regression case, with a Gaussian prior for parameters $\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and σ fixed, the posterior distribution of the parameters is known:

$$\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X} \sim \mathcal{N}((\boldsymbol{X}^T\boldsymbol{X} + \sigma^2\boldsymbol{\Sigma}^{-1})^{-1}(\boldsymbol{X}^T\boldsymbol{y} + \sigma^2\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}), \sigma^2(\boldsymbol{X}^T\boldsymbol{X} + \sigma^2\boldsymbol{\Sigma}^{-1})^{-1}).$$

In our case of study, the parameters $\boldsymbol{\beta}$ are considered uncertain and σ is assumed to be known and identical to the parameter of the source model since we seek to model similar phenomena with identical tolerance.

Equivalently to the classical DoE framework, Bayesian alphabetical criteria are developed and use the matrix $(\boldsymbol{X}^T\boldsymbol{X}+\sigma^2\boldsymbol{\Sigma}^{-1})^{-1}$, proportional to the covariance of the Bayesian linear model parameters, instead of the matrix $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. Then, Bayes $\mathcal{D}/\mathcal{A}/\mathcal{E}/\mathcal{G}$-optimal designs minimize respectively $det((\boldsymbol{X}^T\boldsymbol{X} + \sigma^2\boldsymbol{\Sigma}^{-1})^{-1})$, the trace of $(\boldsymbol{X}^T\boldsymbol{X} + \sigma^2\boldsymbol{\Sigma}^{-1})^{-1}$, the largest eigenvalue of $(\boldsymbol{X}^T\boldsymbol{X} + \sigma^2\boldsymbol{\Sigma}^{-1})^{-1}$ and $\underset{\boldsymbol{x}_i}{max}(\boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{X} + \sigma^2\boldsymbol{\Sigma}^{-1})^{-1}\boldsymbol{x}_i)$.

The Bayesian $\mathcal{D}$-optimal design can be derived from the mutual information for the Gaussian linear model while the Bayesian $\mathcal{A}$-optimal design can be obtained with the quadratic loss. Concerning the $\mathcal{E}$-optimal design and $\mathcal{G}$-optimal design, no utility function has yet been found whose maximisation leads to the construction of such a model.

Notice that Bayesian DoE is not only used if one wishes to build a Bayesian predictive model but can also be used in a non-Bayesian framework.

3 Use case: hydrocraking dataset

The dataset used for this experiment is obtained from two refineries using a new generation catalyst and is composed of 1004 observations described with 12 features. Let $\boldsymbol{X}$ be the matrix of observations and $\boldsymbol{y}$ the vector of the actual value of Diesel density to be predicted. In this experiment, different DoE are tested for a varying size of sample n. The aim is to build a design that allows to obtain the most efficient models with the fewest possible points for learning. The objective is to combined the DoE approach with the Bayesian transfer method (Iapteff et al. 2020), which has proven its efficiency in

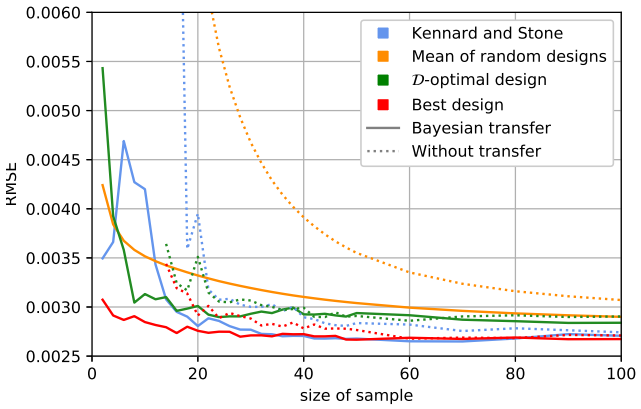


Figure 1: Comparison of the different designs for the Linear model. The red lines give the best possible design obtained by taking for each sample size n the best RMSE score among the thousand random samples.

improving the quality of the model. The model fitted with the Bayesian transfer approach is compared to a classical "non-transfer" model, for which $\boldsymbol{\beta}$ is assumed to be no longer random and estimated by maximum likelihood. The "Bayesian transfer model" is the model learned with prior on $\boldsymbol{\beta}$: $\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\beta}_{prior}, \boldsymbol{\Sigma}_{prior})$. As a reminder, for the Bayesian approach, the prior is built from a dataset from refineries using an old generation catalyst. To compare different methods, the target dataset is split: 600 observations are used as candidate for the design and the 404 remaining observations are used for testing the model. The RMSE score is thus always evaluated only on these 404 points whatever the size n of the sample.

For both models with and without transfer, three different designs are considered:

- Mean of random designs: samples are randomly selected from the candidates dataset, a thousand times, and scores are then averaged.
- Kennard and Stone,
- $\mathcal{D}$-optimal design: the Fedorov's algorithm is used in order to build the $\mathcal{D}$-optimal design for each sample size. For the model without transfer, classical $\mathcal{D}$-optimal design is considered while Bayesian $\mathcal{D}$-optimal design is considered for the model fit with Bayesian transfer.

Note that other alphabetical criteria have been tested and offer less good results than the $\mathcal{D}$-optimal criterion for our prediction task and are therefore not presented here. The results are plotted in Figure 1.

For all designs, the Bayesian transfer approach improves the results compared to a non-transfer approach. Nevertheless, the non-transfer approach with a $\mathcal{D}$-optimal design

provides very good results fairly quickly . The best results with a reduced number of points are obtained with a $\mathcal{D}$-optimal design coupled with the Bayesian transfer approach. As the number of points increases, above 15, the Bayesian transfer approach coupled with the design obtained via Kennard and Stone is better.

In the present case study, the good performance of the Kennard and Stone method in comparison with the $\mathcal{D}$-optimal approach can be explained by the fact that data came from two refineries. Indeed, the dataset is then not exactly homogeneous, and the discrete $\mathcal{D}$-optimal design may be far from the theoretical optimum. Additional experiments on simulated dataset confirmed this behavior.

Another important point is that the design obtained with Kennard and Stone algorithm allows to reach the best possible score when number of observation is sufficient, around 25 observations for our application with 12 features.

4 Conclusion

This study shows that in some practical cases, building optimal designs is not the best solution. It is the case when the features space distribution is not homogeneous and the chosen model does not perfectly describes the phenomenon. Then, the Kennard and Stone algorithm allows to keep the important information when the number of observation is sufficient, and become more effective than optimal design. However, when the number of observations is low, $\mathcal{D}$-optimality remains the most efficient.

References

Celse, Benoit, Jean-Jérôme Da Costa, and Victor Costa (2016). "Experimental Design in Nonlinear Case Applied to Hydrocracking Model: How Many Points Do We Need and Which Ones?" In: *International Journal of Chemical Kinetics* 48.11, pp. 660–670.

Chaloner, Kathryn and Isabella Verdinelli (1995). "Bayesian experimental design: A review". In: *Statistical Science*, pp. 273–304.

De Aguiar, P. F. et al. (1995). "D-optimal designs". In: *Chemometrics and intelligent laboratory systems* 30, pp. 199–210.

Fedorov, Valerii Vadimovich (1972). *Theory of Optimal Experiments Designs*. Academic Press.

Iapteff, Loıc et al. (2020). "Reducing the number of experiments required for modelling the hydrocracking process with kriging through Bayesian transfer learning". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)*.

Kennard, R.W. and L.A. Stone (1969). "Computer aided design of experiments". In: *Technometrics*, pp. 137–148.

Wong, W. K. (1994). "Comparing robust properties of A, D, E and G-optimal designs". In: *Computational statistics & data analysis* 18, pp. 441–448.

Données expérimentales à bas coût pour la prévision de la distance parcourue par une avalanche de pierres

Frédérique Leblanc [1] & Vincent Couallier [2] & Dominique Daudon [3]

[1] *LJK-DATA UGA, 700 avenue centrale, IMAG - CS 40700, 38058 GRENOBLE cedex 9*
frederique.leblanc@univ-grenoble-alpes.fr
[2] *IMB-OPTIMAL Université de Bordeaux*
vincent.couallier@u-bordeaux.fr
[3] *3SR-Géomécanique UGA, BP53, 38041 GRENOBLE cedex 9*
dominique.daudon@univ-grenoble-alpes.fr

Résumé. Les chutes de bloc rocheux constituent un risque important dans les zones montagneuses. La trajectoire de chaque bloc est aléatoire et dépend de sa forme, de la topographie, et de nombreux autres facteurs. Les résultats présentés ici qui portent sur la distance de propagation maximale de l'un des blocs sont obtenus à partir de données acquises sur des essais, très peu coûteux, réalisés en laboratoire sur un modèle réduit et dans des environnements caractérisés par des facteurs contrôlés.

Mots-clés : données expérimentales à bas coût, distance de propagation, survie, modèle de Weibull, quantiles extrêmes

Abstract. In montain areas boulder falls have a significant risk of occuring and can cause heavy human and material damages. Hence it is important to estimate the maximal propagation distance of a rockslide. Herein are presented results obtained with experimental low coast data acquired on a reduced model, where main factors are controlled.

Keywords. Low coast experimental data, propagation distance, survival analysis, Weibull model, extreme quantiles

1 Contexte et objectifs

Dans les régions montagneuses et compte tenu du réchauffement climatique les éboulements rocheux tiennent et tiendront une grande place parmi les risques naturels (environ

30%). Leurs conséquences peuvent être lourdes tant du point de vue humain que matériel et le coût des dispositifs de protection très élevé. Dans les zones les plus sensibles une cartographie des risques est un outils essentiel à la gestion du territoire et est une aide à la décision pour la mise en place des ouvrages de protection ou la délimitation des zones les plus dangereuses, notamment autour des axes de communications. Cette question n'est pas nouvelle et a fait l'objet de nombreuses études ou projets tels que C2ROP (projet national) ou GEORAMP (projet européen) impliquant différents acteurs : entreprises, décideurs de l'état et organismes de recherche. Les résultats disponibles sont souvent propres à la zone géographique observée et rarement transposables dans une autre zone. Il est par ailleurs exclus d'attendre d'avoir des observations d'incidents réels (dont les occurrences sont de l'ordre de la décennie voire du siècle) pour pouvoir cartographier un secteur et anticiper autant que possible un éventuel évènement.

Pour évaluer les risques et les prévenir, plusieurs éléments sont importants à estimer : la sensibilité d'une partie de falaise au risque de décrochage de blocs, les points de premier impact des blocs et la distance maximale parcourue depuis le point de décrochage (zone de propagation). Cette dernière est l'objet d'intérêt ici. Elle est aléatoire et dépend de très nombreux facteurs structurels ou environnementaux. De nombreux modèles statistiques, incluant des facteurs observables, permettent de faire de la prévision sur les distances d'arrêt des blocs mais très peu de données réelles et cohérentes sont disponibles pour un ajustement de ces modèles.

D'une part, les données historiques sont rares et peu documentées et d'autre part les données réelles acquises dans le cas d'essais *in situ* (par ex en carrière) sont peu nombreuses (insuffisamment pour proposer des prévisions fiables et précises) et concernent des lâchers de bloc individuels.

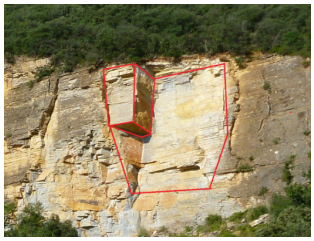
Deux options sont possibles pour acquérir des données en plus grande quantité :

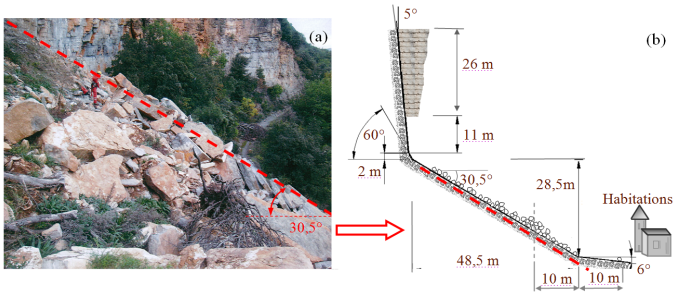
- On peut les simuler avec des algorithmes numériques de propagation mais ils sont coûteux en temps de calculs et posent la question de la calibration des paramètres d'entrée. De plus ces algorithmes qui décrivent la trajectoire d'un seul bloc ne prennent pas en compte les intéractions possibles entre plusieurs blocs chutant en même temps.

- On peut aussi, produire des jeux de données en répliquant des expériences de chute d'objets de caractéristiques variées sur modèle réduit en laboratoire. L'intérêt majeur de cette stratégie est son bas coût et la possibilité d'acquérir autant de données que nécessaire pour l'obtention d'une précision fixée pour la prévision des quantiles extrêmes qui permettront la cartographie des risques.

2 Le dispositif expérimental

Le premier modèle réduit construit par D. Daudon a été conçu pour ressembler à la situation topographique réelle de l'éboulement de Millau (décrite dans *Cuervo 2015*). Ci-dessous à gauche une photo de la masse rocheuse qui s'est décrochée et au centre et à droite une photo et un croquis du plan de coupe correspondant.





Afin de quantifier les quantiles extrêmes de la distance maximale de propagation la maquette suivante a été construite (Figure 1 à gauche) où différentes valeurs de pente ou hauteur de chute sont possibles. Chaque résultat de laché de n blocs est acquis avec un cliché de la zone de propagation (au centre) puis traité pour récupérer les positions (x, y) de chacun des n objets lancés (à droite).

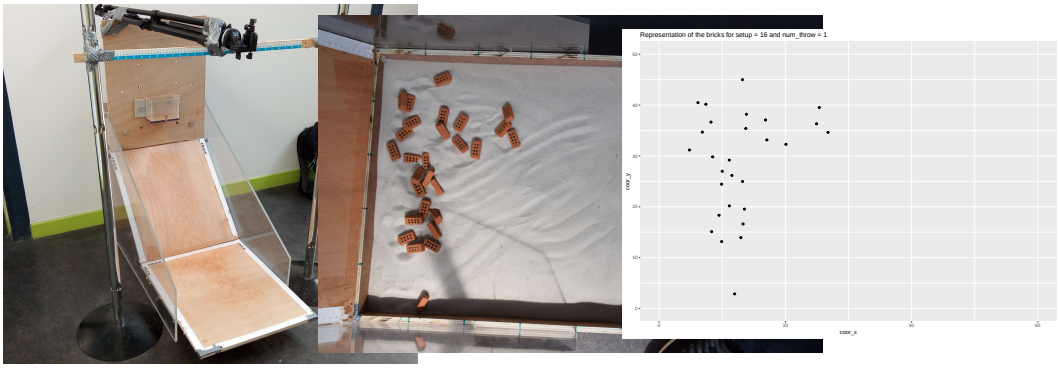


Figure 1 : maquette (gauche) - cliché d'un lancé (centre) - données (droite)

3 Les données disponibles

Les conditions environnementales de l'expérience sont qualifiées par un assez grand nombres de facteurs : pente, hauteur de chute, nombre d'objets lancés, substrats de la pente et du plan horizontal, nombre n d'objets (identiques) lancés, forme, densité, sphéricité des

objets... Les modalités de chaque facteur, sont pour l'instant peu nombreuses, le processus d'acquisition des données n'étant pas automatisé. Ainsi 22 combinaisons différentes de niveaux de facteurs que souhaitaient investiguer les mécaniciens sont pour l'instant utilisées. Chacune de ses 22 combinaisons est appelée setup (numérotés de 1 à 22). Dans chaque setup on dispose de $N = 50$ lancés de n objets.

On s'intéresse ici à la distance X de propagation sur le plan horizontal dans la direction de la pente (voir Figure 1 à droite).

La taille du plan horizontal de la maquette étant insuffisante dans la plupart des setups il y a une proportion plus ou moins significative de données censurées à droite.

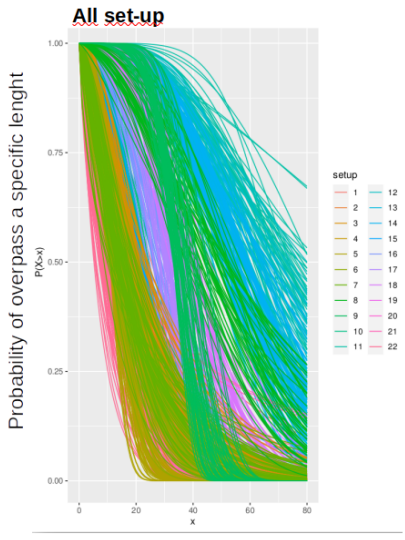
Ces données ont été organisées dans une base de donnée SQLite compilant tous les résultats des essais déjà réalisés (voir *Piras 2019* et *Benlhabib 2021*).

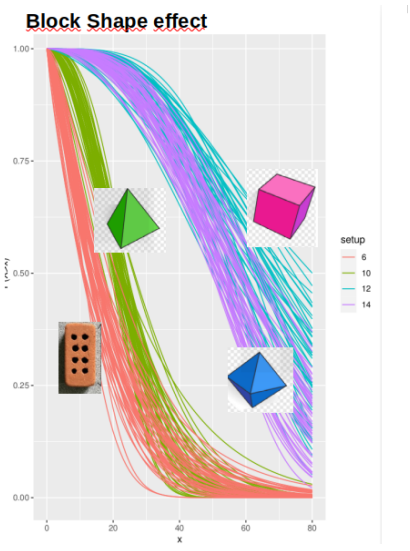
4 traitement statistique

Les répartitions observées de $(x_1, ..., x_n)$ (très dissymétriques) dans chaque environnement suggèrent une loi de Weibull pour décrire X : $\mathcal{W}(\alpha_{i,j}, \lambda_{i,j})$ pour chaque setup $i = 1, ..., 22$ et numéro de lancé $j = 1, ..., 50$ (shape : α et scale : λ).

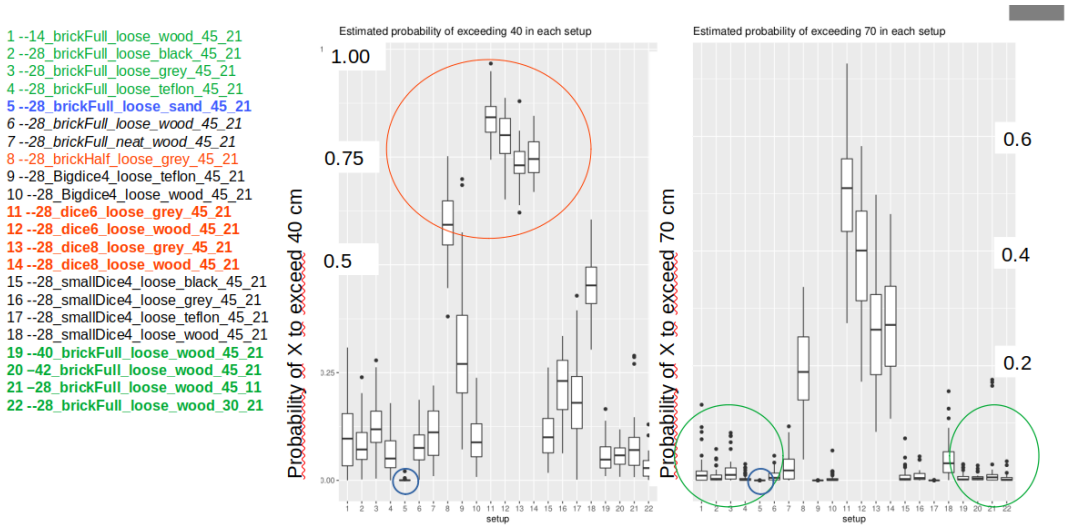
De plus la comparaison de l'estimateur non paramétrique de Kaplan-Meyer et de celui obtenu par maximum de vraissemblance pénalisée (selon le niveau de censure) montrent que le modèle de Weibull est ici un bon choix (voir *Mafhoud 2020*). C'est celui qui a été retenu pour ajuster les deux paramètres de forme et d'échelle dans chacun des environnements et pour chaque lâcher de n blocs (voir *Daudon & al 2021*).

Dans la figure ci dessous sont représentées toutes les fonctions de survie (à gauche) et à droite seules celles qui diffèrent le plus selon le type d'objets lancés et principalement leur sphéricité.





Chaque couple de paramètre $(\alpha_{i,j}, \lambda_{i,j})$ étant estimés on peut accéder aux estimations des quantiles extrêmes ou des probabilités de dépassement de seuil via le modèle ajusté et, s'agissant des probabilités de dépassement des distances (40 et 70 cm), les graphiques suivants permettent d'identifier les facteurs les plus déterminants mais aussi d'apprécier les variabilités intra-setup.



Les situations où la distance de propagation est la plus longue sont celles ou la sphéricité des objets lancés est la plus grande (paquet entouré en rouge dans la figure centrale ci-dessus). La qualité d'amortissement du substrat de réception est elle aussi déterminante comme le montre les résultats obtenus dans le setup où le matériel de réception était du sable (setup 5 pour le sable et setups 1 à 7 pour les autres substrats, toutes choses égales par ailleurs).

L'effet du nombre d'objets lancés (14, 28, 40 ou 42) semble peu significatif. Est-ce le cas pour tout n ? Probablement pas. Aussi chercher à comprendre un peu plus finement l'effet de n sur la distance de propagation est important et sera possible dès qu'auront été acquises de nouvelles données.

5 Travail en cours et perspectives

La seule identification des facteurs les plus déterminants est intéressante mais doit être poursuivie par l'ajustement d'un modèle de régression permettant d'expliquer la probabilité de dépassement ou un quantile selon les niveaux des facteurs qualitatifs et les valeurs des facteurs quantitatifs.

Cela suppose d'acquérir rapidement et efficacement de nouvelles données dans des environnements ou un plus grand nombre de niveaux (au moins cinq pour des facteurs

quantitatifs) et combinaisons de facteurs peuvent être essayés. Or pour l'instant le traitement des images permettant d'extraire les positions des blocs n'est pas complètement automatisé et performant. Un travail de traitement des clichés et de mise à jour de la base de données automatique est en cours. Une semi-automatisation du traitement des images a été déjà été implémenté dans *Benlhabib 2021*.

La prévision sur modèle réduit n'étant pas une fin en soi, il s'agira ensuite pour les mécaniciens d'opérer des transformations d'échelles permettant d'en déduire des distances de propagation à l'échelle réelle des sites sensibles.

C'est un point clef car une trajectoire est une succession de phénomènes mécaniques dynamiques non linéaires lors du contact (vol libre, glissement, roulement, rebond...) mal connus. La littérature concernant le respect des lois de similitudes d'échelles, est peu développée pour les chutes de blocs rocheux . Elle se concentre surtout sur des calibrations d'expériences en centrifugeuses où l'on peut agir sur la gravité pour favoriser le respect des similitudes d'échelles (voir *Chikartamarla & al 2006*, *Bowman & al 2010* et *Asteriou & al 2013*). Quelques illustrations de ce problème sont aussi proposées dans *Daudon & al.*

Bibliographie

Asteriou P., H. Saroglou H. & G. Tsiambaos (2013), *Rockfall : Scaling factors for the Coefficient of Restitution*, Rock Mechanics for Resources, Energy and Environment – Kwasniewski & Lydzba (eds), © 2013 Taylor & Francis Group, London, ISBN 978-1-138-00080-3

Benlhabib Y. (sept. 2021), *Localisation d'objets dans une image et organisation des résultats*, rapport de stage de Master 1, ENSIMAG.

Bowman E.T., Laue J., Imre B. & Springman S.M. (2010), *Experimental modelling of debris flow behaviour using a geotechnical centrifuge*, Canadian Geotechnical Journal, 47(7): 742–762. doi:10.1139/T09-141.

Chikatamarla R., Laue J. & Springman S.M. (2006), *"Centrifuge scaling laws for guided free fall events including rockfalls*, International Journal of Physical Modelling in Geotechnics Vol. 6, No. 2, pp. 14-25.

Cuervo Y. (2015), *Modélisation des éboulements rocheux par la méthode des éléments discrets : application aux évènements réels*, Thèse de doctorat IMEP-2. Université Grenoble Alpes. Français. ⟨NNT : 2015GREAI066⟩.

Daudon D., Couallier V. et Leblanc F. (nov. 2021), *Small scale rock fall experiments and statistical laws for run-out*, Rock Slope Stability 2021, Chambery.

Mafhoud A. (sept. 2020), *: Boulder rockfalls run-outs : statistical analysis of data collected on a scale model*, Rapport de stage, Master 2 SSM, Université de Bordeaux.

Piras T. (juil. 2019), *chute de blocs rocheux : organisation et analyse statistique de données acquises sur modèle réduit*, rapport de stage de l'Université Grenoble Alpes.

PLANS D'EXPERIENCE POUR MELANGES AVEC UNE DISTRIBUTION DE DIRICHLET

Astrid Jourdan[1]

[1] *Laboratoire ETIS, CY Paris Université, aj@cy-tech.fr*

Résumé. Le support de la distribution de Dirichlet est un simplexe tel que les coordonnées appartiennent à [0,1] et leur somme égale 1. Cette loi est donc appropriée pour définir des plans d'expérience pour mélanges. En fonction de ses paramètres, la distribution de Dirichlet permet d'obtenir une distribution des points du plan, symétrique ou asymétrique, uniforme ou concentrée à l'intérieur du domaine expérimental. Le cas de la distribution uniforme est largement utilisé pour les plans d'expérience sur les mélanges. L'uniformité des points de plan est généralement évaluée à l'aide d'un critère de discrépance. Dans ce travail, nous proposons un nouveau critère. Nous utilisons la divergence de Kullback-Leibler pour mesurer l'écart entre la distribution empirique des points du plan et une distribution de Dirichlet. Son estimation est faite, soit avec une méthode plug-in en remplaçant la fonction de densité par une estimation par noyau, soit par plus proches voisins.

Mots-clés. Space-filling design, plan d'expérience pour mélange, Kullback-Leibler divergence, estimation par plus proches voisins, estimation par noyau.

Abstract. The support of the Dirichlet distribution is a simplex such that the coordinates belong to [0,1] and their sum equals 1. It is perfectly suitable to define experimental designs for mixtures. Depending on its parameters, the Dirichlet distribution allows a symmetrical or asymmetrical, uniform or more concentrated point distribution within the experimental domain. The case of a uniform distribution is widely used for mixture experiments. The uniformity of the design points is usually evaluated using a discrepancy criterion. In this paper, we propose a new criterion. We use the Kullback-Leibler discrepancy to measure the difference between the empirical distribution of the design points and a Dirichlet distribution. Its estimation is done either with the plug-in estimate and the nearest-neighbor estimate.

Keywords. Space-filling design, mixture experiments, Kullback-leibler divergence, nearest neighbor density estimation, kernel density estimation

1. Introduction

Dans le cas de mélanges, l'expérience consiste à faire varier les proportions de certains composants impliqués dans un phénomène physico-chimique, et à observer le changement qui en résulte sur la réponse. Les proportions des composants du mélange varient entre 0 et 1 et leur somme doit être égale à 1. La région expérimentale est réduite à un simplexe de (d-1)-dimensions,

$$S^{d-1} = \{(x_1, \ldots, x_d) \mid x_1 + \ldots + x_d = 1,\ x_k \geq 0\},$$

où x_k est la proportion de la kème composante, $k = 1, \ldots, d$.

Depuis Scheffé (1958), de nombreux auteurs ont étudié les plans d'expérience pour mélanges. Les pionniers (Scheffé, 1958, Kiefer, 1961, Cornell, 1981) ont défini des plans optimaux pour les modèles

de mélanges linéaires et quadratiques. Une approche alternative, proposée par Wang et Fang (1990) et Fang et Wang (1994), consiste à couvrir uniformément le domaine expérimental indépendamment de tous modèles. L'idée principale est de générer un plan uniforme sur le cube unitaire de dimension (d-1) comme expliqué dans Hickernell (1998). Ensuite, une fonction de mapping est appliquée pour placer les points dans le simplexe S^{d-1}. Suivant ce principe, de nombreux articles ont proposé des améliorations notamment pour prendre en compte des contraintes complexes sur les composants, Fang et Yang (2000), Prescott (2008), Borkowski et Piepe (2009), Ning et al. (2011), Liu et Liu (2016). Le plan initial dans le cube unité est uniforme mais rien ne garantit qu'il conserve cette distribution uniforme après l'application de la fonction de mapping. Certains auteurs ont défini des critères pour évaluer l'uniformité d'un plan d'expérience pour mélange : La distance quadratique moyenne (Fang et Wang, 1994, Borkowski et Piepe, 2009), la discrépance composite centrale (Chuang et Hung, 2010) ou la discrépance étoile généralisée (Ning et al., 2011).

Dans ce travail, nous proposons un nouveau critère basé sur la distribution de Dirichlet définie par la densité de probabilité,

$$g(x) = \frac{1}{B(\alpha)} \prod_{k=1}^{d} (x_k)^{\alpha_k - 1},$$

où x appartient au simplexe S^{d-1}, $\alpha = (\alpha_1, \dots, \alpha_d)$ avec $\alpha_i > 0$, et $B(\alpha)$ est la constante de normalisation. Cette distribution permet de prendre en compte la distribution uniforme avec le vecteur $\alpha = (1, \dots, 1)$ et ainsi définir un critère pour évaluer l'uniformité d'un plan d'expérience au même titre que les références citées précédemment. Mais elle permet aussi de construire des plans d'expérience avec une distribution asymétrique ou concentrée à l'intérieur du domaine expérimental suivant le paramétrage α choisi.

Le paragraphe 2 explique comment estimer l'écart entre la distribution empirique des points du plan d'expérience et la distribution de Dirichlet, et comment en déduire un nouveau critère.

Dans le paragraphe 3, le nouveau critère est utilisé pour construire des plans d'expérience. Des tests numériques permettent d'étudier son comportement.

2. Définition d'un nouveau critère

Supposons que les points du plan d'expérience pour mélange, $(x_1, \dots, x_d)$, soient n réalisations indépendantes d'un vecteur aléatoire $X = (X_1, \dots, X_d)$ de fonction de densité continue f de support le simplexe S^{d-1}. La divergence de Kullback-Leibler permet de mesurer l'écart entre f et la fonction de densité de la loi de Dirichlet g,

$$I(f, g) = \int_{S^{d-1}} f(x) ln\left(\frac{f(x)}{g(x)}\right) dx.$$

La méthode de Monte Carlo donne un estimateur sans biais et convergent de cette intégrale,

$$\hat{I}(f, g) = \frac{1}{n} \sum_{i=1}^{n} ln\left(\frac{f(x_i)}{g(x_i)}\right). \quad (1)$$

Dans cette expression, il est nécessaire de remplacer la fonction inconnue f par une estimation. Nous proposons d'utiliser la méthode plug-in qui consiste à utiliser une estimation par noyau, ou une estimation par plus proches voisins. Dans les deux cas, l'estimateur de l'intégrale I n'est plus sans biais. Cependant, avoir une estimation biaisée ne pose pas de problème dans notre application. Le but n'est pas d'obtenir une estimation précise de la divergence mais un critère de comparaison entre deux plans d'expérience.

2.1. La méthode plug-in

La fonction de densité inconnue f est estimée par une méthode à noyau à partir des points du plan $D = (x_1, \ldots, x_d)$,

$$\hat{f} = \frac{1}{n|H|^{1/2}} \sum_{i=1}^{n} K\left(H^{1/2}(x - x_i)\right)$$

avec H la matrice des fenêtres et K un noyau gaussien, $K(z) = (2\pi)^{-d/2} e^{-\frac{1}{2}\|z\|^2}$.

Le biais de l'estimateur de $I(f, g)$ ainsi obtenu dépend de n, d et la matrice des fenêtres H (Joe, 1989). Afin de fixer le biais pour comparer des plans de même taille à l'aide de ce critère, la matrice des fenêtres a été choisie indépendante des points du plan telle que $H = diag(h^2{}_1, \ldots, h^2{}_d)$ où $h_k = n^{-1/(d+4)}\sigma_k$ avec σ_k l'écart-type de la loi de Dirichlet,

$$\sigma_k = \frac{1}{\alpha_0}\sqrt{\frac{\alpha_k(\alpha_0 - \alpha_k)}{(\alpha_0 + 1)}} \text{ et } \alpha_0 = \sum_{k=1}^{d} \alpha_k.$$

Pour le cas symétrique et en supprimant les termes qui ne dépendent pas des points du plan, on obtient un critère simplifié,

$$C_{ker\,n}(D) = \sum_{i=1}^{n}\left[ln\left(\sum_{j=1}^{n} e^{-\frac{1}{2}\left\|\frac{x_j - x_i}{h}\right\|^2}\right)\right] - (\alpha - 1)\sum_{i=1}^{n}\sum_{k=1}^{d} ln(x_{ik})$$

où

$$h = n^{-1/(d+4)}\,\frac{1}{d}\sqrt{\frac{d-1}{d\alpha+1}}.$$

2.2. Estimation par plus proches voisins

Wang *et al.* (2006) et Leonenko *et al.* (2008) proposent d'estimer la divergence de Kullback-Leibler par la méthode des plus proches voisins.

Soit $\rho(x, y)$ la distance euclidienne entre deux points x et y de $\mathbb{R}^d$. On note $\rho^{(1)}(x, S) \leq \rho^{(2)}(x, S) \leq \cdots \leq \rho^{(m)}(x, S)$, les distances ordonnées par ordre croissant entre un point x et les points d'un ensemble S tel que $x \notin S$. Dans l'expression (1), en remplaçant la fonction de densité inconnue f par

$$\hat{f}(x) = (n-1)e^{-\Psi(k)}V_d\left(\rho^{(k)}(x, D\backslash x)\right)^d$$

avec Ψ la fonction digamma, V_d le volume de la boule unité de $\mathbb{R}^d$, on obtient un estimateur convergent et asymptotiquement sans biais de $I(f, g)$.

Pour le cas symétrique et en supprimant les termes qui ne dépendent pas des points du plan, on obtient un critère simplifié,

$$C_{nn}(D) = -\sum_{i=1}^{n} ln\left\{\left(\rho^{(k)}(x_i, D\backslash x_i\right)^d\right\} - (\alpha - 1)\sum_{i=1}^{n}\sum_{k=1}^{d} ln(x_{ik}).$$

Naturellement la valeur du critère augmente quand k diminue. Il est donc nécessaire de fixer k pour comparer deux plans d'expérience.

Remarque : Dans le cas d'une distribution uniforme (α=1), les deux critères sont réduits à leur premier terme.

3. Construction des plans

Les deux critères définis précédemment ont été utilisés dans un algorithme d'optimisation afin de

construire des plans d'expérience ayant une distribution de Dirichlet. L'algorithme est initialisé aléatoirement avec un tirage de n points indépendants suivant une loi de Dirichlet. On procède ensuite par remplacement des points avec un simple échange comme défini dans Jin *et al.* (2005).

La figure 1 illustre les plans obtenus avec les deux critères à partir d'une même initialisation de l'algorithme. On constate sur la colonne de gauche qu'un simple tirage aléatoire ne suffit pas pour obtenir des points bien répartis dans le domaine expérimental : les grappes de points donne une information redondante alors que certaines zones restent inexplorées. L'algorithme d'optimisation avec les deux critères corrige les défaillances de la répartition initiale. On note une tendance du critère C_{kern} à repousser les points sur les bords tandis que le critère C_{nn} donne une répartition plus concentrée à l'intérieur du domaine expérimental. Cela est confirmé avec la figure 2 dans laquelle on trouve les boxplots des valeurs du critère DM2 (Ning *et al.*, 2011) calculées sur 30 plans construits par l'algorithme avec les deux critères C_{kern} et C_{nn}. Le critère DM2 mesure l'uniformité de la distribution des points du plan. Il correspond donc au cas particulier des critères C_{kern} et C_{nn} avec $\alpha=1$. En faisant varier les valeurs de α, on note que le critère DM2 atteint bien son minimum pour $\alpha=1$ dans le cas du critère C_{nn}. Dans le cas du critère C_{kern}, le minimum est atteint avec $\alpha=3$, ce qui correspond à une distribution de Dirichlet plus concentrée au centre. Le meilleur plan uniforme (au sens du critère DM2) est obtenu avec le critère C_{nn}.

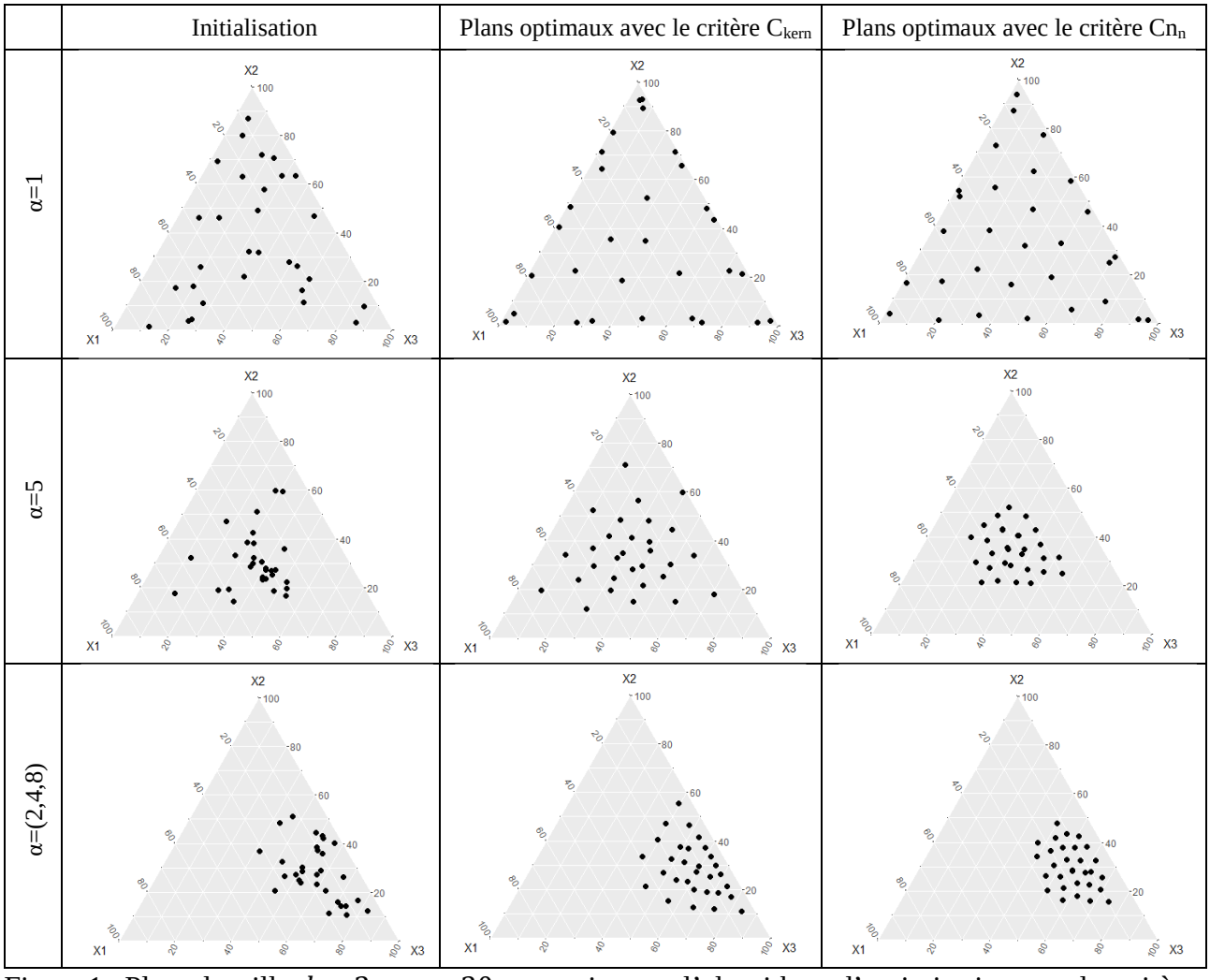


Figure 1 : Plans de taille $d = 3$ et $n = 30$ construits avec l'algorithme d'optimisation pour les critères C_{kern} et C_{nn} à partir de la même initialisation (colonne de gauche) et pour différents paramétrages α de la loi de Dirichlet.

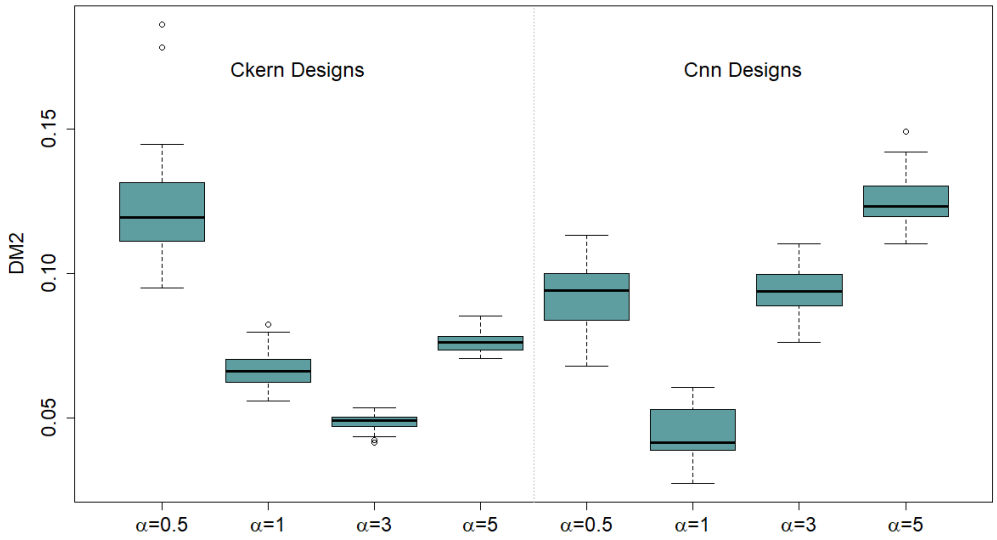


Figure 2 : Boxplot des valeurs du critère DM2 calculé pour 30 plans de taille $d = 3$ et $n = 30$ construits avec les critères C_{kern} et C_{nn} en fonction de différentes valeurs du paramètre α.

4. Conclusion

La loi de Dirichlet et la distance de Kulback-Leibler permettent de définir deux nouveaux critères pour construire des plans d'expérience pour mélanges. A l'issue des tests numériques, il semble que le critère basé sur l'estimation par plus proches voisins soit plus performant que celui basé sur l'estimation par noyau dans le cas où on souhaite construire un plan uniforme. L'avantage de ces deux critères au regard des critères existants dans la littérature est leur capacité à générer des plans ayant une distribution de points plus générale que la distribution uniforme.

Bibliographie

Borkowski, J.J., Piepel, G.F. (2009). Uniform designs for highly constrained mixture experiments. *Journal of Quality Technology*, 41 (1), 35–47.

Chuang S.C., Hung Y.C. (2010). Uniform design over general input domains with applications to target region estimation in computer experiments. *Computational Statistics & Data Analysis*, 54 (1), 219-232.

Cornell, J . A. (1981). *Experiments with Mixtures, designs, models, and the analysis of mixture data*, Wiley, New York.

Fang, K.T., Wang, Y. (1994). Number-theoretic Methods in Statistics. Chapman & Hall, London.

Hickernell, F.J. (1998). A generalized discrepancy and quadrature error bound, *Mathematics of Computation*, 67, 299-322.

Jin, R., Chen, W., Sudjianto, A. (2005). An efficient algorithm for constructing optimal design of computer experiments. Journal of Statistical Planning and Inference, 134, 268–287.

Joe H. (1989). Estimation of entropy and other functional of multivariate density. *Annals of the Institute of Statistical Mathematics*, 41, 683-697.

Kiefer, J. (1961). Optimal designs for regression model, II. *The Annals of Mathematical Statistics*, 32, 298–325.

Leonenko N, Pronzato L, Savani V. (2008). A class of Rényi information estimators for multidimensional densities. *The annals of Statistics*, 36 (5), 2153-2182.

Liu, Y., & Liu, M. (2016). Construction of uniform designs for mixture experiments with complex

constraints. *Communications in Statistics - Theory and Methods*, 45, 2172 - 2180.

Ning J.H., Zhou Y.D., Fang K.T. (2011). Discrepancy for uniform design of experiments with mixtures, *Journal of Statistical Planning and Inference*, 141 (4), 1487-1496.

Prescott, P. (2008). Nearly uniform designs for mixture experiments. Communication in Statistics - Theory and Methods, 37, 2095–2115.

Scheffé, H. (1958). Experiments with Mixtures. *Journal of the Royal Statistical Society: Series B*, 20 (2), 344-360.

Wang Q., Kulkarni R., Verdu S. (2006). A Nearest-Neighbor Approach to Estimating Divergence between Continuous Random Vectors. *2006 IEEE International Symposium on Information Theory*, 242-246.

Wang Y., Fang K.T. (1990). Number theoretic methods in applied statistics (II). *Chinese Annals of Mathematics - Series B*, 11, 384-394.

Prix ENSAI

Modélisation de la courbe de saturation en O2 de l'hémoglobine chez des patients en réanimation

Enora Alaoui*[1], Pierre Barbe[1], Felix Lucas[1], Victoria Mas[1]

[1] ENSAI – Ecole Nationale de la Statistique et de l'Analyse de l'Information – France

Modélisation de la courbe de saturation en O2 de l'hémoglobine chez des patients en réanimation.

14 juin 2022

Programme

PLENIERE : Francoise Berthoud **417**

SFdS - Environnement **418**

Comptage de macrodéchets en zone riparienne par apprentissage profond et modèle d'états, Mathis Chagneux 419

Bivariate Log-Gaussian Cox Process Model for Presence-only Data, Florian Lasgorceux [et al.] 428

When strong annotators lead to bad model choices, Maximilien Servajean [et al.] 434

SFdS - Enseignement **440**

Analyse des tendances de la qualité de l'air, Manon Andre 441

ANNULÉ : Hybridation Numérique : la réussite des étudiants à EPITECH, Nefeli Paparisteidi [et al.] 447

Scientific Reform and Visual Data Science: Retiring the EDA/CDA dichotomy, Roger Beecham 453

Comprendre l'échec des étudiants en vue d'un enseignement différencié, Sylvie Viguier-Pla [et al.] 458

Transposition didactique et enseignement de la statistique, Antoine Rolland . . . 464

Application : santé & société **469**

Application des réseaux de neurones aux données longitudinales de santé: cas du suivi de la fonction rénale aux patients transplantés rénaux, Clément Benoist [et al.] 470

Data visualisation of adverse events in randomised clinical trials, Victoria Cornelius476

Impact de la variabilité de la pression artérielle sur le risque d'AVC, Léonie Courcoul [et al.] 482

Utilisation des matrices emplois expositions pour les troubles musculo-squelettiques. Application aux douleurs sévères du genou, Guillaume Deltreil [et al.] 488

Optimal Individualized Treatment Regime of Plasma Transfusion for Severe Trauma, Pan Zhao [et al.] 494

Données Fonctionnelles **499**

Estimation adaptative dans le modèle linéaire fonctionnel à sortie fonctionnelle, Gaëlle Chagny [et al.] 500

Procédure de classification plug-in pour des données fonctionnelles, Eddy Ella Mintsa 506

Fusion tardive en analyse de données fonctionnelles élastique, Julien Ah-Pine [et al.] 507

Prédiction de la Qualité d'Expérience dans les Réseaux Mobiles : Cas de la VoIP, Jean Steve Tamo Tchomgui [et al.] . . . 513
Sélection d'intervalles pour des prédicteurs fonctionnels à partir de forêts aléatoires, Rémi Servien [et al.] . . . 519
Statistique mathématique . . . **525**
Multivariate Poisson Lognormal model: optimisation, inference and application to high dimensional data, Bastien Batardière [et al.] . . . 526
Les approximations à noyau à base de processus ponctuels déterminantaux, Ayoub Belhadji . . . 532
Un multi-critère pour contrôler la multicolinéarité dans les modèles linéaires de régression multiple, Christian Derquenne . . . 538
High-dimensional Two-sample Rank Statistics as Scalar Performance Criteria, Myrto Limnios [et al.] . . . 544
Clustering de modalités en régression logistique, Slimane Makhlouf [et al.] . . . 550
PLENIERE : Clémentine Prieur . . . **551**
PLENIERE ANNULÉE : Chloé-Agathe Azencott . . . **552**
PLENIERE ANNULEE : Nicolas Papadakis . . . **553**
SFdS - Fiabilité . . . **554**
Degradation Model Selection Using Depth Functions, Mitra Fouladirad [et al.] . . . 555
On misspecification of a Gamma with an inverse Gaussian process and its effects in terms of prognostic and related decision processes, Nicola Esposito [et al.] . . . 557
Sequential Interval Reliablity for semi-Markov systems, Vlad Stefan Barbu [et al.] 559
Building wind loading estimation under uncertain inflow parameters by means of lattice Boltzmann simulation., Jérôme Jacob [et al.] . . . 565
Méta-modélisation multi-fidélité par processus Gaussien en utilisant une base d'ondelettes, Baptiste Kerleguer [et al.] . . . 566
SFdS - Jeunes Statisticiens . . . **569**
Agrégation experts . . . **570**
Modélisation de la charge et de l'occupation des véhicules électriques, Yvenn Amara-Ouali [et al.] . . . 571
Functional mixture-of-experts for classification, Nhat Thien Pham [et al.] . . . 577
Ensemble Methods for Data Analytics, Camila Fernandez [et al.] . . . 583
Consensual aggregation on randomly projected high-dimensional features of predictions for regression, Sothea Has . . . 588
Prédictions Conformelles Adaptatives pour les Séries Temporelles, Margaux Zaffran [et al.] . . . 594
Model selection by penalization in mixture of experts models with a non-asymptotic approach, Trungtin Nguyen [et al.] . . . 600
Apprentissage non supervisé . . . **606**
Clustering Longitudinal Ordinal Data via Finite Mixture of Matrix-Variate Latent Gaussians, Francesco Amato [et al.] . . . 607
Process Mining : une nouvelle approche de l'évaluation des clusters à l'aide d'un Quantitative Model Checking, Emmanuelle Claeys [et al.] . . . 613
Model-Based Clustering by Greedy Maximization of the Integrated Classification Likelihood, Nicolas Jouvin [et al.] . . . 619

A Dynamic Latent Block Model for Co-clustering of Zero-Inflated Count Data Streams, Giulia Marchello [et al.] 625
False clustering rate control in mixture models, Ariane Marandon [et al.] 632
Transport optimal **639**
Transport optimal pour l'appariement automatique de données métabolomiques non-ciblées, Marie Breeur [et al.] 640
GAN Estimation of Lipschitz Optimal Transport, Lucas De Lara [et al.] 646
JDCOT : an Algorithm for Transfer Learning in Incomparable Domains using Optimal Transport, Marion Jeamart [et al.] 652
La Distance de Sliced-Wasserstein pour l'Apprentissage à Grande Échelle, Kimia Nadjahi 658

PLENIERE : Françoise Berthoud

Numérique et transition écologique : amis ou ennemis ?

Françoise Berthoud

CNRS, Grenoble

J'aborderai dans cet exposé la matérialité du monde numérique et examinerai quelques pistes proposées pour réduire son empreinte afin de rester dans les limites planétaires : économie circulaire, découplage, utilisation du numérique pour réduire les émissions de CO2 d'autres secteurs, efficacité, sobriété numérique etc. Je passerai par un petit voyage virtuel aux pays de la "naissance" et de la "mort" de ces bijoux technologiques, pays où les réalités dépassent l'imaginaire de nos dystopies.

SFdS - Environnement

Macrolitter video counting on river banks with state space models for moving cameras

Mathis Chagneux [1] & Charles Ollion [2] & Pierre Gloaguen [3] & Sylvain Le Corff [4] & Océane Lepâtre [5] & Antoine Bruge [6]

[1] *Telecom Paris, 19 place Marguerite Perey, 91120 Palaiseau, France*
mathis.chagneux@telecom-paris.fr
[2] *charges.ollion@gmail.com*
[3] *Agro Paristech, 16 Rue Claude Bernard, 75005 Paris, France,*
pierre.gloaguen@agroparistech.fr
[4] *Telecom SudParis, 9 Rue Charles Fourier, 91000 Courcouronnes, France,*
sylvain.lecorff@telecom-sudparis.fr
[5] *Surfrider Foundation Europe, 33 Allée du Moura, 64200 Biarritz France,*
olepatre@surfrider.eu
[6] *Surfrider Foundation Europe, 33 Allée du Moura, 64200 Biarritz France,*
abruge@surfrider.eu

Résumé.
Les déchets sont une cause connue de dégradation des environnements marins et la plupart d'entre eux voyagent dans les rivières avant d'atteindre les océans. Dans cet article, nous présentons un nouvel algorithme pour aider à la surveillance des déchets le long des cours d'eau. Alors que plusieurs tentatives ont été faites pour quantifier les déchets en utilisant la détection neuronale d'objets dans des photographies d'objets flottants, nous nous attaquons à la tâche plus difficile du comptage direct dans des vidéos en utilisant des caméras embarquées sur des bateaux. Nous nous appuyons sur le suivi d'objets multiples (MOT), mais nous nous concentrons sur la diminution comptages faux et/ou redondants. Notre système ne nécessite qu'une supervision au niveau de l'image et effectue un filtrage bayésien via un modèle d'état basé sur le flux optique. Nous présentons un nouvel jeu de données ouvert d'images recueillies via une campagne de crowdsourcing et utilisées pour entraîner un détecteur d'objets basé sur la prédiction d'une carte de scores où chaque pixel donne la probabilité d'être au centre d'un objet. Des séquences vidéo réalistes assemblées par des experts en surveillance de la pollution marine sont annotées et fournies pour évaluation. Les améliorations de la qualité du comptage sont démontrées en comparaison à des systèmes construits à partir de traqueurs multi-objets partageant les mêmes capacités de détection. Une décomposition précise des erreurs met en évidence les défis restants.

Mots-clés. Filtrage bayésien, poursuite multi cibles, détection d'objets

Abstract.
Litter is a known cause of degradation in marine environments and most of it travels in rivers before reaching the oceans. In this paper, we present a novel algorithm to assist

waste monitoring along watercourses. While several attempts have been made to quantify litter using neural object detection in photographs of floating items, we tackle the more challenging task of counting directly in videos using boat-embedded cameras. We rely on multi-object tracking (MOT) but focus on the key pitfalls of false and redundant counts which arise in typical scenarios of poor detection performance. Our system only requires supervision at the image level and performs Bayesian filtering via a state space model based on optical flow. We present a new open image dataset gathered through a crowdsourced campaign and used to train a center-based anchor-free object detector. Realistic video footage assembled by water monitoring experts is annotated and provided for evaluation. Improvements in count quality are demonstrated against systems built from state-of-the-art multi-object trackers sharing the same detection capabilities. A precise error decomposition allows clear analysis and highlights the remaining challenges.

Keywords. Bayesian filtering, multi-object tracking, object detection...

1 Overview

Litter pollution concerns every part of the globe. Each year, almost ten thousand million tons of plastic waste is generated, among which 80% ends up in landfills or in nature [Geyer et al., 2017], notably threatening all of the world's oceans, seas and aquatic environments [Welden, 2020, Gamage and Senevirathna, 2020]. Plastic pollution is known to already impact more than 3763 marine species worldwide,[1] with risk of proliferation through the whole food chain. This accumulation of waste is the endpoint of the largely misunderstood path of trash, mainly coming from land-based sources [Rochman et al., 2016], yet rivers have been identified as a major pathway for the introduction of waste into marine environments [Jambeck et al., 2015]. Therefore, field data on rivers and monitoring are strongly needed to improve decision making and assess the impact of measures that can be taken.

Different methods have already been tested to monitor waste in rivers: litter collection and sorting on riverbanks [Bruge et al., 2018], visual counting of drifting litter from bridges [González-Fernández et al., 2021], floating booms [Gasperi et al., 2014] and nets [Morritt et al., 2014]. All are helpful to understand the origin and typology of litter pollution yet hardly compatible with long term monitoring at country scales. Monitoring tools need to be reliable, easy to set up on various types of rivers, and should give an overview of plastic pollution during peak discharge to help locate hotspots and provide trends. Newer studies suggest that plastic debris transport could be better understood by counting litter trapped on river banks, providing a good indication of the local macrolitter pollution especially after increased river discharge [van Emmerik et al., 2019, van Emmerik and Schwarz, 2020]. Based on these findings, we propose a new method for litter monitoring which relies on videos of river banks directly captured from moving boats.

1. We provide a novel open-source image dataset of macro litter, which includes various objects seen from different rivers and different contexts. This dataset was produced with a new open-sourced platform for data gathering and annotation developed in conjunction with an Surfrider Foundation Europe, continuously growing with more data.

2. We propose a new algorithm specifically tailored to count in videos with fast camera movements. In a nutshell, DNN-based object detection is paired with a robust movement model which uses optical flow to perform Bayesian filtering, while confidence regions built on posterior predictive distributions are used for data association. This framework does not require video annotations at training time. It also fully leverages optical flow estimates and the uncertainty provided by Bayesian predictions to recover object identities even when detection recall is low. Contrary to existing

[1] see litterbase.awi.de

MOT solutions, this method ensures that tracks are stable enough to avoid repeated counting of the same object.

3. We provide a set of video sequences where litter counts are known and depicted in real conditions. For these videos only, litter positions are manually annotated at every frame in order to carefully analyze performance. This allows us to build new informative count metrics. We compare the count performance of our method against other MOT-based alternatives.

2 Optical flow-based counting via Bayesian filtering and confidence regions

Our counting method is divided into several interacting blocks. First, a detector outputs a set of predicted positions for objects in the current frame. The second block is a tracking module designing consistent trajectories of potential objects within the video. At each frame, a third block links the successive detections together using confidence regions provided by the tracking module, proposing distinct tracks for each object. A final postprocessing step only keeps the best tracks which are enumerated to yield the final count.

2.1 Detector

In most benchmarks, the prediction quality of object attributes like bounding boxes is often used to improve tracking. For counting, however, point detection is theoretically enough and advantageous in many ways. First, to build large datasets, a method which only requires point annotations may benefit from more data due to annotation ease. Second, contrary to previous popular methods [Ren et al., 2015] involving intricate mechanisms for bounding box prediction, center-based and anchor-free detectors [Zhou et al., 2019, Law and Deng, 2018] only use additional regression heads which can simply be removed for point detection. Adding to all this, [Zhang et al., 2021] highlight conceptual and experimental reasons to favor anchor-free detection in tracking-related tasks.

For these reasons, we use a stripped version of CenterNet [Zhou et al., 2019] where offset and bounding box regression heads are discarded to output bare estimates of center positions on a coarse grid. An encoder-decoder network takes an input image $I \in [0,1]^{w \times h \times 3}$ (an RGB image of width w and height h), and produces a heatmap $\hat{Y} \in [0,1]^{\lfloor w/p \rfloor \times \lfloor h/p \rfloor}$ such that $\hat{Y}_{xy}$ is the probability that (x, y) is the center of an object (p being a stride coefficient).

2.2 Bayesian tracking with optical flow

Optical flow Between two timesteps $n-1$ and n, the optical flow Δ_n is a mapping satisfying the following consistency constraint [Paragios et al., 2006]:

$$\widetilde{I}_n[u] = \widetilde{I}_{n-1}[u + \Delta_n(u)]\,,$$

where, in our case, $\widetilde{I}_n$ denotes the frame n downsampled to dimensions $\lfloor w/p \rfloor \times \lfloor h/p \rfloor$ and $u = (x, y)$ is a coordinate on that grid. To estimate Δ_n, we choose a simple unsupervised Gunner-Farneback algorithm which does not require further annotations, see [Farnebäck, 2003].

State-space model Using optical flow as a building block, we posit a state-space model where estimates of Δ_n are used as a time and state-dependent offset for the state transition. Let $(X_k)_{k\geq 1}$ and $(Z_k)_{k\geq 1}$ be the true (hidden) and observed (detected) positions of a target object in $\mathbb{R}^2$, respectively. Considering the optical flow value associated with X_{k-1} on the discrete grid of dimensions $\lfloor w/p \rfloor \times \lfloor h/p \rfloor$, write

$$X_k = X_{k-1} + \Delta_k(\lfloor X_{k-1} \rfloor) + \eta_k \tag{1}$$

and

$$Z_k = X_k + \varepsilon_k\,,$$

where $(\eta_k)_{k\geq 1}$ are i.i.d. centered Gaussian random variables with covariance matrix Q independent of $(\varepsilon_k)_{k\geq 1}$ i.i.d. centered Gaussian random variables with covariance matrix R. In the following, Q and R are assumed to be diagonal.

Approximations of the filtering distributions Denoting $u_{1:k} = (u_1, \ldots, u_k)$ for any k and sequence $(u_i)_{i\geq 0}$, Bayesian filtering aims at computing the conditional distribution of X_k given $Z_{1:k}$, referred to as the filtering distribution. In the case of linear and Gaussian state space models, this distribution is known to be Gaussian, and Kalman filtering allows to update exactly the posterior mean $\mathbb{E}[X_k|Z_{1:k}]$ and posterior variance matrix $\mathbb{V}[X_k|Z_{1:k}]$. This algorithm and its extensions are prevalent and used extensively in time-series and sequential-data analysis. As the transition model proposed in (1) is nonlinear, Kalman updates cannot be implemented and solving the target tracking task requires resorting to alternatives. Many solutions have been proposed in the literature such as extended Kalman filters (EKF), unscented Kalman filters (UKF) or Sequential Monte Carlo (SMC) methods, see [Särkkä, 2013] and references therein. In this paper, we propose to use an EKF approximation of (1) which is computationally cheaper and confirmed to be as robust as these alternatives. At each time step k, (1) is approximated by

$$X_k = X_{k-1} + \Delta_k(\lfloor \mu_{k-1} \rfloor) + \nabla_x \Delta_k(\lfloor \mu_{k-1} \rfloor)(X_{k-1} - \mu_{k-1}) + \eta_k\,,$$

which allows the implementation of Kalman updates. It is worth noting here that this state-space model naturally accounts for missing observations, as the contribution of Δ_k in every transition ensures that each filter can cope with arbitrary inter-frame motion to keep track of its target.

2.3 Data association using confidence regions

Throughout the video, depending on various conditions on the incoming detections, existing trackers must be updated (with or without a new observation) and others might need to be created. This traditional Multiple Hypothesis Tracking (MHT) setup requires a third party data association block to link the incoming detections with the correct filters. At the frame n, a set of L_n Bayesian filters track previously seen objects and a new set of detections $\mathcal{D}_n$ is provided by the detector. Denote by $1 \leq \ell \leq L_n$ the index of each filter at time n, and by convention write $Z_{1:n-1}^{\ell}$ the previous observed positions associated with index ℓ (even if no observation is available at some past times for that object). Let $\rho \in (0, 1)$ be a confidence level.

1. For every detected object $z_n^i \in \mathcal{D}_n$ and every filter ℓ, compute $P(i, \ell) = \mathbb{P}(Z_n^{\ell} \in V_\delta(z_n^i) \mid Z_{1:n-1}^{\ell})$ where $V_\delta(z)$ is the neighborhood of z defined as the squared area of width 2δ centered on z.

2. Using the Hungarian algorithm [Kuhn, 1955], compute the assignment between detections and filters with P as cost function, but discarding associations (i, ℓ) having $P(i, \ell) < \rho$. Formally, ρ represents the level of a confidence region centered on detections and we use $\rho = 0.5$. Denote a_ρ the resulting assignment map defined as $a_\rho(i) = \ell$ if z_n^i was associated with the ℓ-th filter, and $a_\rho(i) = 0$ if z_n^i was not associated with any filter.

3. For $1 \leq i \leq D_n$, if $a_\rho(i) = \ell$, use z_n^i as a new observation to update the ℓ-th filter. If $a_\rho(i) = 0$, create a new filter initialized from the prior distribution, i.e. sample the true location as a Gaussian random variable with mean z_n^i and variance R.

4. For all filters ℓ' which were not provided a new observation, update only the predictive law of $X_n^{\ell'}$ given $Z_{1:n-1}^{\ell'}$.

In other words, we seek to associate filters and detections by maximising a global cost built from the predictive distributions of the available filters, but an association is only valid if its corresponding predictive probability is high enough. Though the Hungarian algorithm is a very popular algorithm in MOT, it is often used with the Euclidean distance or an Intersection-over-Union (IoU) criterion. Using confidence regions for the distributions of Z_n given $Z_{1:(n-1)}$ instead allows to naturally include uncertainty in the decision process. Note that we deactivate filters whose posterior mean estimates lie outside the image subspace in $\mathbb{R}^2$.

3 Practical impacts and future goals

Conducted in coordination with Surfrider Foundation Europe specialized on water preservation, our work is part of a broader campaign for macrolitter monitoring and is already being used in a production version of a monitoring system. That said, large amounts of litter items are still not detected. Solving this problem is largely a question of augmenting the object detector training dataset through crowdsourced images. A specific annotation platform (https://www.trashroulette.com/) is online, thus the amount of annotated images is expected to continuously increase, while training is provided to volunteers collecting data on the field to ensure data quality. Finally, several expeditions on different rivers are already underway and new video footage is expected to be annotated in the near future for better evaluation. All data is made freely available. Future goals include downsizing the algorithm, a possibility given the architectural simplicity of anchor-free detection and the relatively low computational complexity of EKF. In a citizen science perspective, a fully embedded version for portable devices will allow a larger deployment. The resulting field data will help better understand litter origin, allowing to model and predict litter density in non surveyed areas. Correlations between macro litter density and environmental parameters will be studied (e.g., population density, catchment size, land use and hydromorphology). Finally, our work naturally benefits any extension of macrolitter monitoring in other areas (urban, coastal, etc) that may rely on a similar setup of moving cameras.

References

[Bruge et al., 2018] Bruge, A., Barreau, C., Carlot, J., Collin, H., Moreno, C., and Maison, P. (2018). Monitoring litter inputs from the Adour river (southwest France) to the marine environment. *Journal of Marine Science and Engineering*, 6(1).

[Farnebäck, 2003] Farnebäck, G. (2003). Two-frame motion estimation based on polynomial expansion. In *Scandinavian conference on Image analysis*, pages 363–370. Springer.

[Gamage and Senevirathna, 2020] Gamage, T. and Senevirathna, J. (2020). Plastic pollution in the marine environment. *Heliyon*, 6:e04709.

[Gasperi et al., 2014] Gasperi, J., Dris, R., Bonin, T., Rocher, V., and Tassin, B. (2014). Assessment of floating plastic debris in surface water along the seine river. *Environmental Pollution*, 195:163 – 166.

[Geyer et al., 2017] Geyer, R., Jambeck, J., and Law, K. (2017). Production, use, and fate of all plastics ever made. *Science Advances*, 3:e1700782.

[González-Fernández et al., 2021] González-Fernández, D., Cózar, A., Hanke, G., Viejo, J., Morales-Caselles, C., Bakiu, R., Barcelo, D., Bessa, F., Bruge, A., Cabrera, M.,

Castro-Jiménez, J., Constant, M., Crosti, R., Galletti, Y., Kideyş, A., Machitadze, N., de Brito, J. P., Pogojeva, M., Ratola, N., Rigueira, J., Rojo-Nieto, E., Savenko, O., Schöneich-Argent, R. I., Siedlewicz, G., Suaria, G., and Tourgeli, M. (2021). Floating macrolitter leaked from europe into the ocean. *Nature Sustainability*, 4:474 – 483.

[Jambeck et al., 2015] Jambeck, J., Geyer, R., Wilcox, C., Siegler, T., Perryman, M., Andrady, A., Narayan, R., and Law, K. (2015). Marine pollution. plastic waste inputs from land into the ocean. *Science (New York, N.Y.)*, 347:768–771.

[Kuhn, 1955] Kuhn, H. W. (1955). The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97.

[Law and Deng, 2018] Law, H. and Deng, J. (2018). Cornernet: Detecting objects as paired keypoints. In Ferrari, V., Hebert, M., Sminchisescu, C., and Weiss, Y., editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XIV*, volume 11218 of *Lecture Notes in Computer Science*, pages 765–781. Springer.

[Morritt et al., 2014] Morritt, D., Stefanoudis, P. V., Pearce, D., Crimmen, O. A., and Clark, P. F. (2014). Plastic in the thames: A river runs through it. *Marine Pollution Bulletin*, 78(1):196–200.

[Paragios et al., 2006] Paragios, N., Chen, Y., and Faugeras, O. D. (2006). *Handbook of mathematical models in computer vision*. Springer Science & Business Media.

[Ren et al., 2015] Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 91–99, Cambridge, MA, USA. MIT Press.

[Rochman et al., 2016] Rochman, C., Andrady, A., Dudas, S., Fabres, J., Galgani, F., lead, D., Hidalgo-Ruz, V., Hong, S., Kershaw, P., Lebreton, L., Lusher, A., Narayan, R., Pahl, S., Potemra, J., Rochman, C., Sherif, S., Seager, J., Shim, W., Sobral, P., and Amaral-Zettler, L. (2016). Sources, fate and effects of microplastics in the marine environment: Part 2 of a global assessment.

[Särkkä, 2013] Särkkä, S. (2013). *Bayesian Filtering and Smoothing*. Cambridge University Press, New York, NY, USA.

[van Emmerik and Schwarz, 2020] van Emmerik, T. and Schwarz, A. (2020). Plastic debris in rivers. *WIREs Water*, 7(1):e1398.

[van Emmerik et al., 2019] van Emmerik, T., Tramoy, R., van Calcar, C., Alligant, S., Treilles, R., Tassin, B., and Gasperi, J. (2019). Seine Plastic Debris Transport Tenfolded During Increased River Discharge. *Frontiers in Marine Science*, 6(October):1–7.

[Welden, 2020] Welden, N. (2020). *The environmental impacts of plastic pollution*, pages 195–222.

[Zhang et al., 2021] Zhang, Y., Wang, C., Wang, X., Zeng, W., and Liu, W. (2021). Fairmot: On the fairness of detection and re-identification in multiple object tracking. *International Journal of Computer Vision*, pages 1–19.

[Zhou et al., 2019] Zhou, X., Wang, D., and Krähenbühl, P. (2019). Objects as points.

Parameter identifiability in bivariate spatiotemporal point processes for ecological presence-only data under incomplete sampling

Florian Lasgorceux [1] & Julien Papaïx [1] & Yoann Bunz [2] & Damien Combrisson [2] & Thomas Opitz [1]

[1] *INRAE - Unité BioSP, 228 route de l'Aérodrome, CS 40509, Domaine Saint-Paul - Site Agroparc, 84914 Avignon Cedex 9.*
florian.lasgorceux@inrae.fr , thomas.opitz@inrae.fr, julien.papaix@inrae.fr.
[2] *Parc national des Ecrins, Domaine de Charance, 05000 Gap.*
yoann.bunz@ecrins-parcnational.fr , damien.combrisson@ecrins-parcnational.fr

Résumé. Les gestionnaires d'espaces naturels ont de plus en plus recours à des bases de données de comptage d'espèce alimentées de manière opportuniste (*données de présence seule*). Ces données offrent une meilleure couverture spatio-temporelle mais sont affectées de multiples biais d'acquisition du fait de l'absence de protocole d'observation. Ceci se traduit en particulier par un effort d'échantillonnage hétérogène et inconnu. De plus, la mise en évidence des interactions entre espèces à partir de ce type de données n'est pas aisée puisque la co-occurrence d'observations ne présage pas forcément d'une influence d'une espèce sur l'autre. Dans ce contexte, nous nous sommes intéressés à la modélisation jointe de la distribution d'espèces et à l'inférence d'un tel modèle à partir de données opportunistes. Nous développons, dans un premier temps, un modèle de processus ponctuel bivarié (pour deux espèces), à savoir un processus de Cox log-gaussien, dont les intensités intègrent des covariables environnementales et des effets aléatoires spatiaux, dont un commun aux deux variables. L'identifiabilité des paramètres de ce modèle est étudiée par simulations. Dans un deuxième temps, un effort d'échantillonnage hétérogène en espace est ajouté au modèle et nous discutons des conséquences du choix de modélisation de ce dernier. Les simulations et l'estimation des paramètres du modèle sont réalisées dans un cadre bayésien via la méthode INLA-SPDE. L'estimation est explorée pour différentes configurations des paramètres liés à l'abondance et la variabilité spatiale de chaque espèce, à l'interaction entre les deux espèces et à la structure de l'effort d'échantillonnage. Le cadre de simulation est adapté au contexte du Parc National des Ecrins pour lequel des données opportunistes des communautés aviaires sont disponibles.

Mots-clés. Processus de Cox log-gaussien, INLA, données de présence seule, effort d'échantillonage, données opportunistes, écologie des communautés, modèle de distribution d'espèce.

Abstract. Managers of natural parcs and protected areas are increasingly using databases of species counts, which are supplied in an opportunistic way (presence-only data). These data guarantee a better spatio-temporal coverage than available protocoled data but suffer from strong biases due to the lack of a proper observation protocol, which leads

to a heterogeneous and unknown sampling effort. Moreover, the study of species interactions through this kind of data is arduous since empirical co-occurrence patterns do not necessarily imply an influence of one species on another. Our study deals with the joint modeling of species distributions and the inference of such models from opportunistic data. We firstly establish a bivariate (for two species) point process model constructed as a log-gaussian Cox process. The intensity functions of the two species take into account environmental covariates and spatial random effects, including one that is common to both variables to account for species interaction. Our main goal here is to study the identifiability of parameters using simulations. Moreover, a sampling effort that may be heterogeneous in space is added to the model, and we discuss the consequences of modeling the latter. The simulations and the estimation of the model parameters are performed through Bayesian inference using the INLA-SPDE approach. The simulation framework is adapted to the Ecrins National Park, for which opportunistic data of bird communities are available.

Keywords. Log-Gaussian Cox Process, INLA, presence-only data, sampling effort, opportunistic data, ecological community, species distribution models.

1 Contexte d'application

Les modèles de distribution d'espèces permettent de prédire, dans l'espace et dans le temps, la présence et l'abondance d'espèces animales ou végétales. Ils sont donc un pré-requis à la compréhension des dynamiques écologiques et de leurs réponses aux changements climatiques et d'usage des sols. De plus, ils servent de base à la mise en place de nouvelles mesures de gestion ainsi qu'à l'évaluation de mesures déjà en place.

Du fait de l'étendue spatiale et temporelle nécessaire, les données permettant d'ajuster de tels modèles sont généralement fortement hétérogènes, les protocoles changent suivant les groupes taxonomiques, les lieux ou les périodes, et le nombre d'observateurs est grand. Parmi ces données, les gestionnaires d'espaces naturels ont de plus en plus recours aux données de contact opportuniste. Ces données, non protocolées, renseignent la présence d'un individu d'une espèce donnée en un lieu et un temps donné. L'avantage majeur de telles bases de données alimentées de manière opportuniste provient de leur importante couverture spatiale, de leur quasi-continuité temporelle et de l'information sur les comportements inhabituels (marge d'aire de répartition, décalage phénologique). Cependant, les données opportunistes sont difficiles à exploiter car fortement affectées par de multiples biais d'acquisition, en particulier l'échantillonnage inhomogène et préférentiel dans les dimensions de l'espace, du temps et des différentes espèces.

Dans ce travail, nous cherchons à caractériser la dynamique spatio-temporelle jointe de deux espèces à partir de données opportunistes. Nous développons un modèle de Cox log-Gaussien bivarié dont les intensités intègrent des covariables envionnementales et des

effets aléatoires spatiaux. Un des effets spatiaux est commun aux deux espèces et permet d'étudier les possibles interactions entre celles-ci. L'identifiabilité des paramètres est étudiée par simulations. Nous considérons ensuite un effort d'échantillonnage hétérogène en espace et nous discutons des conséquences du choix de modélisation de ce dernier.

2 Modélisation de la répartition spatiale de deux espèces

2.1 Du processus de Poisson au processus de Cox log-gaussien

L'ensemble des positions des individus d'une espèce observée dans une fenêtre d'observation spatiale peut être vu comme un semis de points aléatoires, et ce nuage de points peut être modélisé par un processus ponctuel. Le nombre d'individus se situant dans une unité de l'espace est un nombre entier aléatoire à valeurs dans $\mathbb{N} = \{0, 1, 2, \ldots\}$, et les nombres d'individus obtenus pour une partition de la fenêtre d'observation peuvent être modélisés via un processus de comptage. Les processus de comptage sont caractérisés par leur fonction d'intensité (ou simplement intensité) qui est définie comme le nombre moyen d'observation par unité de surface. L'intensité sera notée $\lambda(s)$ où $s \in \Omega \subset \mathbb{R}^2$ est point de l'espace. Comme le rappelle Renner et *al* (2015), le processus de comptage le plus naturel est le processus de Poisson inhomogène qui repose sur les hypothèses suivantes :

(i) Les points d'observations sont indépendants les uns des autres.

(ii) L'intensité $\lambda(s)$ varie dans l'espace.

L'hypothèse (ii) est souvent réduite à la modélisation de l'intensité par une fonction log-linéaire des covariables environnementales :

$$\ln(\lambda(s)) = \boldsymbol{x}(s)\boldsymbol{\beta},$$

où $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$ est le vecteur des paramètres correspondant aux p covariables environnementales $\boldsymbol{x}(s) = (x_1(s), \ldots, x_p(s))$. On notera que la fonction de lien logarithme assure la positivité de l'intensité.
L'hypothèse (i), trop restrictive pour nos données de biodiversité, peut être contournée en modélisant les tendances d'agrégation ou de fragmentation des observations au delà de structures spatiales induites par les covariables à l'aide d'un champ aléatoire. On appelle un tel processus un processus de Cox. Ainsi, conditionnellement à la réalisation de ce champ, les observations forment un processus de Poisson inhomogène. Le processus de Cox qui nous intéresse a été proposé par Møller et *al.* (1998) et est appelé processus de Cox log-Gaussien (*Log-Gaussian Cox process*, LGCP en abrégé). Concrètement, l'intensité que l'on notera Λ puisqu'elle est aléatoire, est toujours fonction des covariables environnementales mais aussi désormais d'un champ aléatoire gaussien W :

$$\ln(\Lambda(s)) = \boldsymbol{x}(s)\boldsymbol{\beta} + W(s).$$

Ici, W est un champ gaussien de moyenne nulle et de fonction de covariance fixée arbitrairement, évalué en les points $s \in \Omega$. Le champ W décrit la variabilité, structurée dans l'espace, non expliquée par les covariables environnementales.

2.2 Modèle mis en oeuvre dans cette étude

Nous avons fait le choix de modéliser l'intensité du processus d'observation de l'espèce $i \in \{1, 2\}$, notée Λ_i^{obs}, comme le produit d'un effort d'échantillonage dépendant de l'espace et de l'espèce $e_i(s)$ et d'un champ aléatoire, $\Lambda_i(s)$ représentant l'abondance de l'espèce i. C'est-à-dire,

$$\Lambda_i^{obs}(s) = e_i(s)\Lambda_i(s). \tag{1}$$

Le champ aléatoire Λ_i est fonction des covariables environnementales et de deux champs gaussiens, afin d'exprimer la variabilité spatiale non expliquée en une partie propre à l'espèce, et une autre issue de l'interaction entre les espèces. Si l'on note Λ_1 (*resp.* Λ_2) l'intensité du processus associé à l'espèce 1 (*resp.* à l'espèce 2), W_1 (*resp.* W_2) le champ gaussien propre à l'espèce 1 (*resp.* à l'espèce 2), et W_0 le champ gaussien partagé par les deux espèces, on obtient :

$$\ln(\Lambda_1(s)) = \boldsymbol{x}(s)\boldsymbol{\beta}_1 + W_1(s) + W_0(s), \tag{2}$$
$$\ln(\Lambda_2(s)) = \boldsymbol{x}(s)\boldsymbol{\beta}_2 + W_2(s) + \gamma W_0(s), \tag{3}$$

où γ (équation (3)) est un paramètre de corrélation entre les répartitions des deux espèces au travers du champ W_0. Ainsi, conditionnellement aux covariables environnementales et aux réalisations des champs gaussiens, un paramètre γ positif aura tendance à agréger les observations des deux espèces tandis qu'un paramètre γ négatif aura tendance à fragmenter les observations entre les deux espèces. Les vecteurs de coefficients $\boldsymbol{\beta}_1$ (*resp.* $\boldsymbol{\beta}_2$) sont quant à eux les paramètres correspondant aux p variables environnementales $\boldsymbol{x}(s)$ pour l'espèce 1 (*resp.* pour l'espèce 2).

3 Estimation des paramètres

Nous nous concentrons dans cette partie sur l'identifiabilité des paramètres mis en jeu dans le modèle décrit par les équations (2) et (3). On y présente la méthode INLA (Integrated Nested Laplace Approximation) et les différents scénarios de simulation.

3.1 La méthode INLA

Supposons pour le moment l'effort d'échantillonnage constant égal à 1 (voir équation (1)) afin de se concentrer sur l'abondance des espèces $\Lambda_i(s)$ pour $i \in \{1, 2\}$. Supposons

que nous faisions intervenir p covariables environnementales dans notre modèle et que les fonctions de covariance de chaque processus latent (W_0, W_1 et W_2) sont des covariances de Matérn (ici avec le paramètre de régularité fixé à $\nu = 1$). Chaque fonction de covariance est alors contrôlée par deux paramètres, la variance et la portée spatiale. Au total, nous avons besoin d'estimer $2 \times (p + 1)$ coefficients linéaires et 7 hyperparamètres (trois variances, trois portées spatiales, et le paramètre d'interaction γ). Pour cela nous utilisons une méthode d'inférence bayésienne reposant sur les approximations de Laplace appelée INLA développée par Rue, Martino et Chopin (2009). L'avantage d'INLA par rapport aux autres méthodes d'estimation bayésiennes du type Markov Chain Monte Carlo (MCMC) provient du gain en temps de calcul pour les modèles pouvant être exprimés comme des champs gaussien latents, et ce gain est particulièrement important en utilisant des champs gaussiens markoviens (souvent abrégés *GMRF* pour Gauss–Markov Random Fields). Avec INLA, nous utilisons l'approche des équations aux dérivées partielles stochastiques (*Stochastic Partial Differential Equations*, SPDE) pour simuler et estimer les champs gaussiens W_0, W_1 et W_2 grâce à des fonctions de base du type des éléments finis. Cette approche décrite par Lindgren, Rue et Lindström (2011), consiste en une approximation numériquement avantageuse de la fonction de covariance de Matérn via des matrices de précision creuses $\boldsymbol{Q}_j$ (donc de type GMRF) pour le vecteur des coefficients et mène à la représentation suivante :

$$W_j(s) = \sum_{\ell=1}^{L_j} w_{j,\ell}\psi_{j,\ell}(s), \quad \boldsymbol{w}_j = (w_{j,1}, \ldots, w_{j,L_j}) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{Q}_j^{-1}),$$

où $\psi_{j,\ell}$ sont les éléments finis fixés grâce à une triangulation de l'espace.

3.2 Scénarios de simulation

Afin d'étudier les capacités d'inférence de nos paramètres à l'aide d'INLA, différents scénarios de simulations suivant le modèle bivarié (équations (2) et (3)) seront présentés. Le premier scénario portera sur la simulation de données gaussiennes dans un carré unité. Un intérêt particulier sera porté à la qualité d'estimation des paramètres lié au champ partagé W_0 et au paramètre γ de corrélation entre les répartitions des deux espèces. Nous nous attacherons en particulier à l'estimation du signe de ce paramètre d'interaction ainsi qu'à sa significativité, où nous considérons un effet comme significatif si son intervalle de crédibilité ne contient pas la valeur zéro. En effet, au-delà de sa valeur quantitative, du signe du paramètre γ dépendent les interprétations écologiques.
Le deuxième scénario de simulation présentera l'adaptation de ces simulations de données gaussiennes au domaine du Parc National des Ecrins. Ce jeu de simulation permettra d'étudier l'importance de la résolution spatiale du maillage choisi pour la simulation des champs gaussiens à l'échelle des Ecrins en abordant le compromis entre le temps de calcul et la précision des estimations.

Enfin, le troisième scénario de simulation portera sur les simulations de données de comptage dans le parc national des Ecrins. En outre, la qualité des estimations des différents paramètres dans le cas de données de comptage (relativement peu informatives pour les champs gaussiens latents) sera comparée à celle obtenue pour des données gaussiennes avec une erreur de mesure (amenant plus d'information utile pour l'estimation des champs gaussiens latents).

4 Modélisation spatio-temporelle et effort d'échantillonnage

Une fois les résultats des simulations obtenus dans le cadre spatial, nous présenterons l'extension du modèle dans le cadre spatio-temporel à l'aide d'une structure autorégressive. Pour tout $i \in \{1, 2, 3\}$, le champ W_i s'écrit en le point $(s, t) \in \Omega \times \mathbb{R}_+$ comme

$$W_i(s,t) = W_i(s) + \varepsilon(t) \qquad \text{avec} \qquad \varepsilon(t+1) = \rho\varepsilon(t) + \sqrt{1-\rho^2}\tilde{\varepsilon}(t)$$

où les $\tilde{\varepsilon}(t)$ sont *i.i.d* de loi $\mathcal{N}(0, \sigma^2)$ et $\varepsilon(1) = \tilde{\varepsilon}(1)$.
Enfin on abordera la question de la modélisation de l'effort d'échantillonnage et de son estimation. Une première étude consistera à intégrer un effort d'échantillonnage dépendant de l'espace et du temps comme un offset de l'abondance en considérant les observations de l'ensemble des espèces. Dans une seconde étude, à l'instar de Botella et al. (2019), nous modéliserons et estimerons de façon explicite un champs spatial décrivant cet effort d'échantillonnage.

Bibliographie

Botella, C. and Joly, A. and Monestiez, P. and Munoz, F. (2019). Bias in presence-only niche models related to sampling effort and species niches : Lessons for background point selection. *PLoS ONE*, 15(5) : e023207, pp. 1-18.
Lindgren, F. and Rue, H. and Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields : the stochastic partial differential equation approach. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 73(4), pp.423-498.
Møller, J. and Syversveen, A. R. and Waagepetersen, R. P. (1998). Log Gaussian Cox processes. *Scandinavian Journal of Statistics*, 25, pp. 451-482.
Renner, I.W and Elith, J. and Baddeley, A. and Fithian, W. and Hastie, T. and Phillips, S. and Popovic, G. and Warton, D.I. (2015). Point process models for presence-only analysis - a review, *Methods in Ecology and Evolution*, 6, pp. 366-379.
Rue, H. and Martino, S. and Chopin, N. (2009) Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 71(2), pp. 319-392.

When strong annotators lead to bad model choices

Maximilien Servajean[1,2] & Waleed Ragheb[3] & Valentin Leveau[3] & Alexis Joly[1,4]

[1] *LIRMM, CNRS, University of Montpellier, Montpellier, France - prenom.nom@lirmm.fr*
[2] *AMIS, Paul Valery University, Montpellier, France*
[3] *Seenovate - prenom.nom@seenovate.com*
[4] *Inria, Montpellier, France - prenom.nom@inria.fr*

Résumé. Les contextes participatifs tels que les sciences citoyennes impliquent souvent des procédures d'annotation où les labels d'un jeu de données sont produits de manière collaborative. De telles procédures ont été étudiées dans la littérature et leurs performances mesurées au travers de métriques telles que l'*accuracy*. Cependant, peu de travaux ont connecté la qualité des annotations aux performances des modèles de *machine learning* entraînés sur ces dernières. Dans cette étude, nous montrons que l'*accuracy* n'est pas nécessairement un bon critère. En particulier, nous montrons que sous la présence de bruit, même des labels obtenus via le classifieur optimal de Bayes ne permettent pas de garantir que le modèle obtenu minimise bien l'erreur souhaitée.

Mots-clés. Apprentissage automatique, crowdsourcing, convergence.

Abstract. Participatory contexts such as citizen science often involve annotation procedures where the labels of a dataset are produced in a collaborative way. Such procedures have been extensively studied in the literature and their performance is measured through metrics such as *accuracy*. However, few studies devoted attention to linking the quality of annotations to the performance of the underlying machine learning models trained on these datasets. In this paper, we show that accuracy is not an necessarily a good criterion. In particular, we prove that under the presence of noise, labels obtained via the Bayes classifier do not guarantee that the model actually minimizes the error.

Keywords. Machine learning, crowdsourcing, convergence.

1 Introduction

Participatory science often involves a positive feedback loop where the more participants, the better the models built from the data collected, and the better the reward for the contributors. For instance, Figure 1 illustrates this in the context of Pl@ntnet [1, 8]. Pl@ntnet is a participatory app that aims at identifying plant species from observations. The machine learning model benefits from more data to obtain better performances which drives

the users to share their contributions. The annotation is done via several tools including a serious game[1], an expert platform[2] and the Pl@ntnet application. The produced data provides a rich source of information for research in ecology. However, it raises several essential challenges: data is biased (i.e. some regions are easier to visit than others) and the ecological models should correct it [2]. Regarding the machine learning predictions themselves, as the number of classes grows and as the ambiguity increases, predicting only the top-1 class makes less sense. Typically, one would reach for top-K or average-K predictors [7] and optimize the model for such metrics [5].

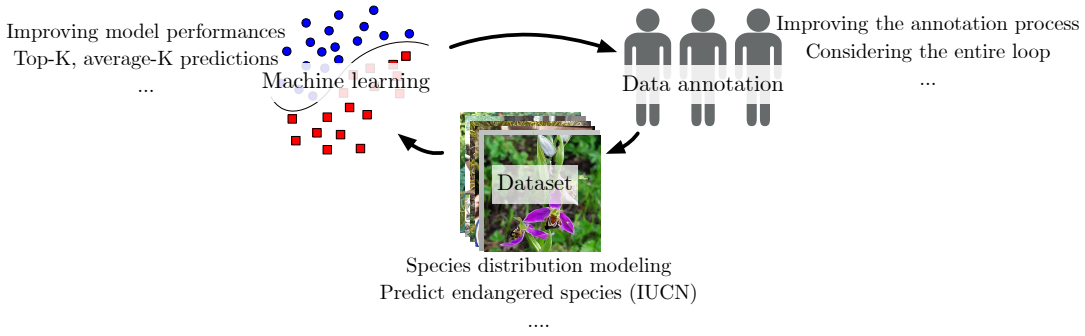


Figure 1: The positive participative loop.

Even though large efforts have been made to produce high-quality crowd-sourced annotations, little has been made to analyze these procedures within a machine learning loop. In this study, we show theoretically that some non intuitive bias may arise and that even high-quality annotations may lead to irremediable failure in the choice of the machine learning model.

The paper is organized as follows. Section 2 introduces the setup. Section 3 shows some results in the context of coherent learning with crowd-sourced labels. Finally, Section 4 concludes and presents some open questions.

2 Setup

For simplicity[3], we consider binary classification problems where $\mathcal{X}$ refers to the input space and $\mathcal{Y} = \{0, 1\}$ to the label space. Given a pair of random variables $(X, Y) \in \mathcal{X} \times \mathcal{Y}$, define μ as the marginal of X and $\eta(x) = \mathbb{P}(Y = 1 | X = x)$ as the label conditional probability. The pair (μ, η) defines the joint probability of (X, Y), noted $\mathbb{P}$. Our objective is to find a mapping $h : \mathcal{X} \to \mathcal{Y}$ such that the risk $L(h) \triangleq \mathbb{P}\big(h(X) \neq Y\big)$ is minimized based on a dataset $\mathcal{S}_n \sim \mathbb{P}^{\otimes n}$ of size n. Define the empirical risk as follows $L_n(h) =$

[1] `http://theplantgame.com`
[2] `https://www.tela-botanica.org/en/`
[3] All the results generalize easily to multi-class classification.

$\frac{1}{n}\sum_{X,Y\in\mathcal{S}_n}\mathbf{1}\{h(X)\neq Y\}$. Given a set of maps $\mathcal{H}$ from $\mathcal{X}$ to $\mathcal{Y}$, the empirical risk minimizer is defined as $h_{\text{erm}}=\arg\min_{h\in\mathcal{H}}L_n(h)$. The Bayes classifier $g^\star(x)=\mathbf{1}\{\eta(X)\geq 0.5\}$ minimises the error among all measurable functions. It's error is called the Bayes error and is obtained by $L^\star=\mathbb{E}\big[\min(\eta(X);1-\eta(X))\big]$.

Typically, constructing the dataset $\mathcal{S}_n$ requires having the corresponding labels. The common procedure to obtain them is through an annotation procedure. An annotator, either human, machine or crowd, takes an input a sample $x\in\mathcal{X}$ and outputs a label in $\mathcal{Y}$. We describe this process by the annotation map $\Pi:\mathcal{X}\mapsto[0,1]$ defined as $\Pi(x)=\pi_1(x)\eta(x)+(1-\pi_0(x))(1-\eta(x))$ where $\boldsymbol{\pi}=(\pi_0,\pi_1)^\top$ relates to the annotator's confusion matrix[4] [6]. Given $x\in\mathcal{X}$ the annotation procedure consists in sampling a label from $\text{Ber}(\Pi(x))$. The annotation map (Π) can model deterministic[5] to stochastic behaviors.

Suppose that we have a pool of annotators. Each time we need some label, we first select the annotator with some probability and then ask for the label. Let Y^a define the random variable describing the label obtained. Then we have $\eta^a(x)=\mathbb{P}\left(Y^a=1|X=x\right)=\eta(x)(\mathbb{E}\big[\pi_1(x)+\pi_0(x)\big]-1)+1-\mathbb{E}\big[\pi_0(x)\big]$ where the expectation is taken over the pool of annotators. Obviously, even if the annotators are better than a random classifier, η^a might be very different from η.

The pair (μ,η^a) defines the joint probability of (X,Y^a) and similarly, one may obtain the same quantities: $\mathcal{S}_n^a$, L^a, L_n^a, h_{erm}^a, $g^{a,\star}$ and $L^{a,\star}$ where the exponent a acknowledges the annotated origin of the label. We call L^a the proxy task and L the original task. We say that the proxy task permits coherent learning if $L(h_{\text{erm}})\xrightarrow{n\to\infty}\inf_{h\in\mathcal{H}}L(h)$ implies $L(h_{\text{erm}}^a)\xrightarrow{n\to\infty}\inf_{h\in\mathcal{H}}L(h)$. In other words, the best classifier on the task resulting from the annotations is also the best classifier on the original task.

3 Data Annotation

3.1 Majority voting and bayesian combination

In most crowdsourcing platforms, each sample receives multiple label propositions before being assembled to get the dataset $\mathcal{S}_n^a$ [6]. Define $Y_{i,j}^a$ as the j^{th} label obtained for the i^{th} item. Given a vector $(Y_{i,1}^a,\ldots,Y_{i,k}^a)^\top$, the **majority voting** selects the label that corresponds to the statistical mode. We propose here a close alternative where the annotations are kept $(Y_{i,1}^a,\ldots,Y_{i,k}^a)^\top$ and where optimization is done through the **soft empirical risk** defined as $L_{n,k}^a(h)=1/(kn)\sum_{i,j}\mathbf{1}\{h(X)\neq Y_{i,j}^a\}$. The **weighted majority voting** consists in weighing each vote by the annotator's accuracy which is optimized through an Expectation Maximization (EM) procedure. Finally, the **Bayesian annotator combination** is an improvement of the weighted majority voting where the annotator's confusion is modeled and priors are set on the classes proportions. It is typically optimized through

[4]In multi-classes problems, $\boldsymbol{\pi}$ is actually a right stochastic matrix called the confusion matrix.

[5]The Bayes classifier is defined by $\boldsymbol{\pi}=\mathbf{1}\{\eta(x)\geq 0.5\}(1,0)^\top+\mathbf{1}\{\eta(x)<0.5\}(0,1)^\top$.

approximated inference or Monte Carlo methods.

3.2 Convergence in noise-free settings ($L^\star = 0$)

First, we are interested in how fast learning on the proxy task actually reduces the error on the original task. This can be done by upper bounding the error gap on the proxy task and relating this error to the original task. It is known that [4] $L(h_{\text{erm}}) - \inf_{h\in\mathcal{H}} L(h) \leq 2\sup_{h\in\mathcal{H}} |L_n(h) - L(h)|$. Thus, bounding the right hand side is sufficient. Due to a lack of space, we will consider only a fixed $h \in \mathcal{H}$ as bounding the sup is more demanding and requires to consider the VC dimension [3].

Proposition 1 *Fix $\delta \in [0;1]$ and $h \in \mathcal{H}$. Consider the soft empirical risk. Then, we have with probability at least $1-\delta$: $\left|L^a(h) - L^a_{n,k}(h)\right| \leq \sqrt{\frac{2L^a(h)\log\frac{2}{\delta}}{kn}} + \sqrt{\frac{2L^a(h)\log\frac{2}{\delta}}{n}} + \frac{2\log\frac{2}{\delta}}{3n}$.*

Proof 1 *Set $Z_i = \sum_{j=1}^k \mathbf{1}\{h(X_i) \neq Y^a_{i,k}\}/k - L^a(h)$. Apply Bernstein inequality and a union bound, to obtain $\mathbb{P}\left(|\sum_{i=1}^n Z_i| \geq n\epsilon\right) \leq 2exp\left\{-\frac{n\epsilon^2/2}{L^a(h)/k+\frac{1}{3}\epsilon}\right\}$. Finally, equate the right hand side to δ and observe that for $a, b \geq 0$ we have $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ and for an equation $x^2 + bx - c = 0$, $b, c \geq 0$, we have $x \leq b + \sqrt{c}$.*

The upper bound of the soft empirical risk in Proposition 1 is decomposed into three terms relating to i) the label variability, ii) the error variability on $\mathcal{X}$ and iii) a classical term, respectively. Only the label variability term is associated with the number of annotators (k), while all the three terms are associated with the dataset size itself (n). Therefore, Proposition 1 suggests that it seems better to construct a larger dataset, than to multiply the number of annotations. This idea generalizes to the majority voting, weighted majority voting and Bayesian annotation. Obviously, the proxy task needs to permit coherent learning in order to minimize the risk of the original task as well. The next proposition shows that under classical hypotheses from the literature, the annotators make coherent learning possible[6].

Proposition 2 *Suppose that $\mathbb{E}\big[\pi_0\big] = \mathbb{E}\big[\pi_1\big] > 0.5$. Then, we have that $L(h_{erm}) \xrightarrow{n\to\infty} \inf_{h\in\mathcal{H}} L(h) \Leftrightarrow L(h^a_{erm}) \xrightarrow{n\to\infty} \inf_{h\in\mathcal{H}} L(h)$.*

Proof 2 *We have $\eta^a(x) = \eta(x)(2\mathbb{E}\big[\pi_1(x)\big] - 1) + 1 - \mathbb{E}\big[\pi_1(x)\big]$. The risk is given by $L^a(h) = (2\mathbb{E}\big[\pi_1(x)\big] - 1)L(h) + 1 - \mathbb{E}\big[\pi_1(x)\big]$ where $2\mathbb{E}\big[\pi_1(x)\big] - 1 > 0$.*

In other words, even though the convergence speed from the proxy to the original task depends on the noise added by the annotation procedure and described by $\boldsymbol{\pi}$, minimizing the former actually minimizes the latter as well. Unfortunately, once $\mathbb{E}\big[\pi_0\big] \neq \mathbb{E}\big[\pi_1\big]$, the situation becomes more difficult.

[6]Note that here the symmetry is required in expectation and not for each annotator. Task assignment policies could enforce the property.

Proposition 3 *Suppose that* $L^\star = 0$. *Then* $\exists \mu, \eta^a, \mathcal{H}$ *with* $|\mathcal{H}| > 1$ *such that* $g^\star(x) = g^{a,\star}(x)$ *a.s. and* $L(\arg\min_{h\in\mathcal{H}} L^a(h)) = \max_{h\in\mathcal{H}} L(h)$.

Proposition 3 shows that even if the annotators are better than random, leading to the same Bayes classifier on both the proxy and the original task, the way they make mistakes can lead to incoherent learning. Moreover, although $\mathbb{E}\big[\pi_0\big] = \mathbb{E}\big[\pi_1\big]$ is sufficient to prevent this behavior, it is not a necessary hypothesis. But this is beyond the scope of this paper.

Proof 3 *Set* $\mathcal{X} = \{x_1, x_2\}$, *note* $\mu(x_1) = \rho$ *and* $\mu(x_2) = 1-\rho$. *Set* $\eta^a(x_1) > \eta^a(x_2) > 0.5$. *Consider* $\mathcal{H} = \{(0,1),(1,0)\}$ *and set* $0.5 > \rho > \frac{\eta^a(x_2)-0.5}{\eta^a(x_1)+\eta^a(x_2)-1}$. *We obtain* $L(h_2) < L(h_1)$ *while* $L^a(h_2) > L^a(h_1)$.

Fortunately, it is still possible to minimize the gap between the errors obtained on the proxy and original tasks. Intuitively, the closer the risk estimation on the proxy to the original task, the smaller the gap. As annotators are better than random and $L^\star = 0$, it suffices to collect more annotations to better decide which label should be returned.

Proposition 4 *Suppose* $g^\star(x) = g^{a,\star}(x)$ *a.s.,* $L^\star = 0$ *and consider majority voting (MV). Note* L_k^a *and* $L_k^{a,\star}$ *the MV risks. Then, we have* $|L_k^a(h) - L(h)| \leq \mathbb{E}\big[e^{-2k(0.5-\min(\eta^a(X);1-\eta^a(X)))^2}\big]$.

Proof 4 *We have* $|L_k^a(h) - L(h)| \leq \mathbb{E}\big[\,|\eta(X) - \eta_k^a(X)|\,\big]$. *Then, considering Hoeffding gives the desired inequality.*

It is worth mentioning that if the annotators are extremely noisy, the convergence can be arbitrarily slow.

3.3 Convergence in the noisy settings ($L^\star \neq 0$)

In noisy settings, the problem becomes more complicated. Propositions 1 and 2 apply immediately, showing that coherent learning is achievable even in noisy settings. However, solving incoherence is not that easy. Let us consider a crowdsourcing platform that always outputs the most likely label. It minimizes the labeling error such as the Bayes classifier. However, as shown below, these platforms do not necessarily permit coherent learning.

Proposition 5 *Suppose that* $L^\star \neq 0$. *Then* $\exists \mathcal{H}$, $|\mathcal{H}| > 1$, η *and* μ *such that* $L^{a,\star} = 0$ *(the annotators are Bayes classifiers* $g^{a,\star}$*) and* $L(\arg\min_{h\in\mathcal{H}} L^a(h)) = \max_{h\in\mathcal{H}} L(h)$.

Proof 5 *The proof is similar to Proposition 3.*

Proposition 4 solved incoherence by converging toward the Bayes classifier which can be incoherent in the noisy setting. Thus, it no longer solves incoherency.

Soft labeling may resolve and limit the effect of this problem. However, and unfortunately, humans are incapable of estimating probabilities accurately (or even in expectation), this workaround is not trivial.

4 Conclusion and open questions

Crowdsourcing and annotators combination strategies have been extensively studied in the literature. The quality of the combination is typically evaluated in terms of classification accuracy. Even though minimizing this accuracy proves to be a coherent strategy in noise-free settings, the situation becomes more difficult in the noisy setting. Labels obtained through a Bayes classifier are no longer a guarantee to obtain coherent learning afterwards. Additionally, we showed that in a coherent context, obtaining a larger dataset rather than a dataset with more annotations is likely to provide a reduced estimation gap. This gap is directly related to the original task risk.

The difficulties described above pose many open questions. Most significantly, how to find a compromise between dataset size and coherence in order to minimize the risk ? How to guarantee coherent labeling in the noisy setting ? To what extent do these problems affect the likelihood estimation of the statistical models – which is at the basis of methods such as average-K predictors ?

References

[1] A Affouard, H Goëau, P Bonnet, JC Lombardo, and A Joly. Pl@ntnet app in the era of deep learning. In *ICLR - Workshop Track*, 2017.

[2] C Botella, A Joly, P Monestiez, P Bonnet, and F Munoz. Bias in presence-only niche models related to sampling effort and species niches: Lessons for background point selection. *Plos one*, 15(5):e0232078, 2020.

[3] S Boucheron, O Bousquet, and G Lugosi. Theory of classification: A survey of some recent advances. *ESAIM: probability and statistics*, 9:323–375, 2005.

[4] L Devroye, L Györfi, and G Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013.

[5] C Garcin, A Joly, P Bonnet, A Affouard, JC Lombardo, M Chouet, M Servajean, and J Salmon. Pl@ntNet-300K: a new plant image dataset for the evaluation of set-valued classifiers. In *NeurIPS Datasets and Benchmarks*, 2021.

[6] HC Kim and Z Ghahramani. Bayesian classifier combination. In *Artificial Intelligence and Statistics*, pages 619–627. PMLR, 2012.

[7] T Lorieul, A Joly, and D Shasha. Classification under ambiguity: When is average-k better than top-k?, 2021.

[8] M Servajean, A Joly, D Shasha, J Champ, and E Pacitti. Crowdsourcing thousands of specialized labels: a bayesian active training approach. *IEEE Transactions on Multimedia*, 19(6):1376–1391, 2017.

SFdS - Enseignement

ANALYSE DES TENDANCES DE LA QUALITE DE L'AIR

Manon ANDRE[1&2]

[1] *Département STID, IUT 2, Université Grenoble-Alpes, andre_manon@yahoo.fr*

[2] *Institut des Géosciences de l'Environnement, Université Grenoble-Alpes, 38400 Saint-Martin-d'Hères*

Résumé. Au cours de ma seconde année de DUT STID, j'ai eu l'opportunité d'effectuer mon stage au sein de l'Institut des Géosciences de l'Environnement, sur le campus de Saint-Martin-d'Hères. L'objectif de ce stage était d'observer les tendances des concentrations massiques des éléments chimiques composant les particules fines et de les quantifier afin de comparer un site urbain et un site agricole, mais également de comparer les différentes tailles de particules au sein d'un même site.

Les données ont été nettoyées puis agrégées par mois avant de tester la présence ou non de tendance puis en déduire la quantification de celle-ci. L'idée est d'étudier les tendances en confrontant deux méthodes statistiques ayant des hypothèses différentes, l'une se basant sur la moyenne et l'autre sur la médiane. Pour cela, l'étude a été réalisée avec le langage python.

Il est apparu que la concentration massique des éléments des particules fines suivent majoritairement une diminution significative sur les deux typologies de sites (urbain et rural). Ces résultats indiquent premièrement que la mise en place de politiques publiques larges échelles ont un impact sur le trafic routier notamment avec le carbone élémentaire (EC), le cuivre (Cu) et le zinc (Zn). Cette constatation est confirmée par une diminution également présente sur les fractions grossière et fine du site rural, qui est éloigné des sources locales. Deuxièmement, on observe que sur Grenoble les politiques locales semblent fonctionner avec une accélération de la baisse de carbone élémentaire (EC) sur les cinq dernières années.

Mots-clés. Qualité de l'air, analyse de tendance

L'impact environnemental et sociétal des particules atmosphériques (PM pour Particulate Matter en anglais) est aujourd'hui établi par la communauté scientifique. Il a été en effet prouvé qu'une exposition aux particules fines accroît notablement les risques de maladies cardio-vasculaires et respiratoires. Pour répondre à cet enjeu, la législation européenne et française impose des seuils de concentrations réglementaires et proposent différentes actions pour réduire la quantité ou nocivité des particules fines.

J'ai eu l'opportunité d'effectuer mon stage au sein de l'Institut des Géosciences de l'Environnement, sur le campus de Saint-Martin-d'Hères. Le premier objectif de ce stage était d'observer les tendances des concentrations massiques des éléments chimiques composant les particules fines. Dans un second temps, nous avons cherché à les quantifier et à déterminer leur évolution afin de comparer un site urbain et un site rural, mais également comparer les différentes tailles de particules au sein d'un même site à travers des mesures de longues durées. Pour cela, deux sites ont été sélectionnés : Grenoble (site urbain), et le site de l'Observatoire Pérenne de l'environnement « OPE » , proche d'Houdelaincourt (site rural).

Les mesures s'étendent de 2012 à aujourd'hui ou de 2009 à aujourd'hui selon les espèces chimiques considérées. Une observation de notre jeu de données représente des concentrations massiques mesurées pour une trentaine d'éléments (parmi plusieurs milliers, voire centaines de milliers composant les particules fines) pour un jour donné. Néanmoins, nous nous sommes focalisés sur neuf variables centrales (Masse totale des particules fines, concentration massique en Carbone Organique, Carbone Élémentaire, Lévoglucosan, Sulfate, Nitrate, Ammonium, Cuivre, Zinc) car elles sont émises par des sources variées et pertinentes pour les deux sites étudiés. Ces séries, uniques de par l'étendue des éléments analysés sur une si longue période, permettent une analyse des politiques publiques larges échelles (renouvellement du parc automobile avec moins de diesel, mais également par des véhicules plus lourds) et locales (« Prim'Air bois » ou zone à émissions restreintes dans le cadre de Grenoble).

La concentration de chaque élément chimique ne présente pas la même saisonnalité. On peut remarquer notamment sur la Figure 1, avec l'élément Carbone élémentaire qu'une régression simple n'était pas adéquate.

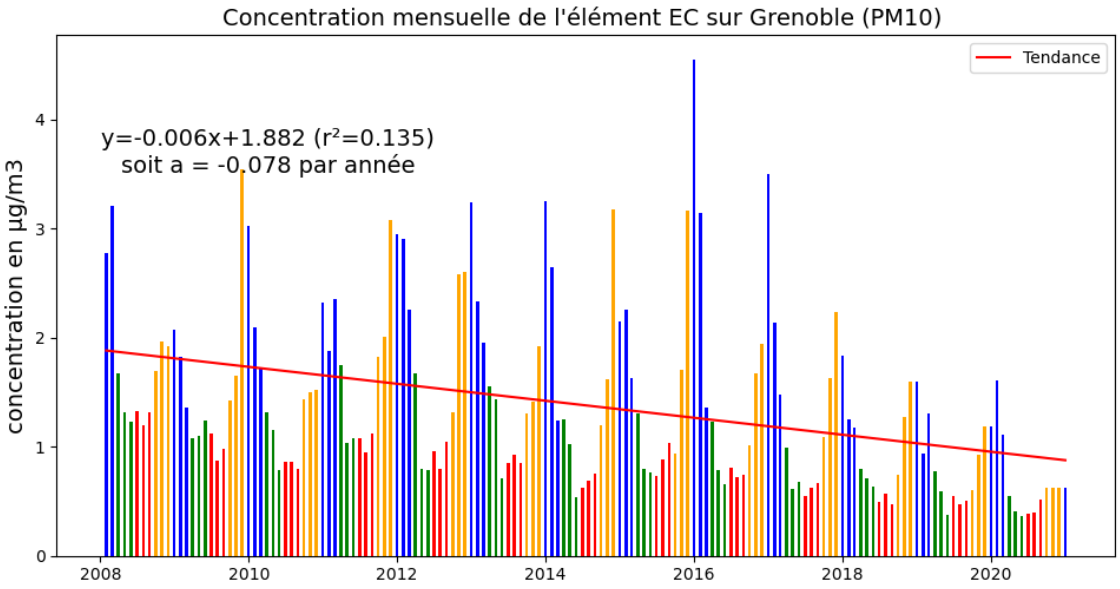


Figure 1 : Variation mensuelle de l'élément EC sur Grenoble

On en a conclu qu'au vu d'une forte variabilité saisonnière et des différences de sources potentielles de particules fines, il était nécessaire d'élaborer une méthode plus adaptée qu'une régression linéaire simple sur les données initiales journalières. Les données ont donc été nettoyées puis agrégées par mois, avant de tester la présence ou non d'une tendance puis d'en déduire ou non la quantification de celles-ci via des méthodes plus adéquates.

L'idée était d'étudier les tendances en confrontant deux méthodes statistiques ayant des hypothèses différentes, l'une se basant sur la moyenne et l'autre sur la médiane pour assurer une fiabilité à nos résultats. Nous avons pour cela mené notre étude avec le langage Python via différents packages tels que « Pandas », « Statsmodels », « Matplotlib », « Seaborn » ou encore « Pymannkendall ».

La première méthode est une décomposition du signal selon le modèle additif (Équation 1) : une tendance, une composante saisonnière et une composante résiduelle en utilisant la méthode du LOESS.

Équation 1 : Modèle additif

$$Y(t) = T(t) + S(t) + R(t)$$

Le LOESS est un acronyme de « LOcally Estimated Scatterplot Smoothing » qui est une méthode de régression locale basée sur les k plus proches voisins.

La seconde méthode est l'utilisation du test de Mann-Kendall non-paramétrique permettant de détecter si une tendance est significative dans une série temporelle. Ce test est utilisé conjointement à la pente de Sen (Équation 2) qui permet une estimation de cette dernière. Ici, la pente est la médiane de toutes les estimations Q de pentes entre chaque paire de points.

Équation 2 : Estimation de pente entre une paire de points

$$Q = \frac{Yi' - Yi}{i' - i}$$

Où Q est une estimation de la pente. Y_i et $Y_{i'}$ sont les valeurs aux instants i et i', où i' est supérieur à i.

Nos résultats s'articulent en trois temps. Premièrement, nous avons étudié la tendance globale des espèces. Puis, nous avons quantifié ces tendances comme nous pouvons le voir sur la Figure 2.

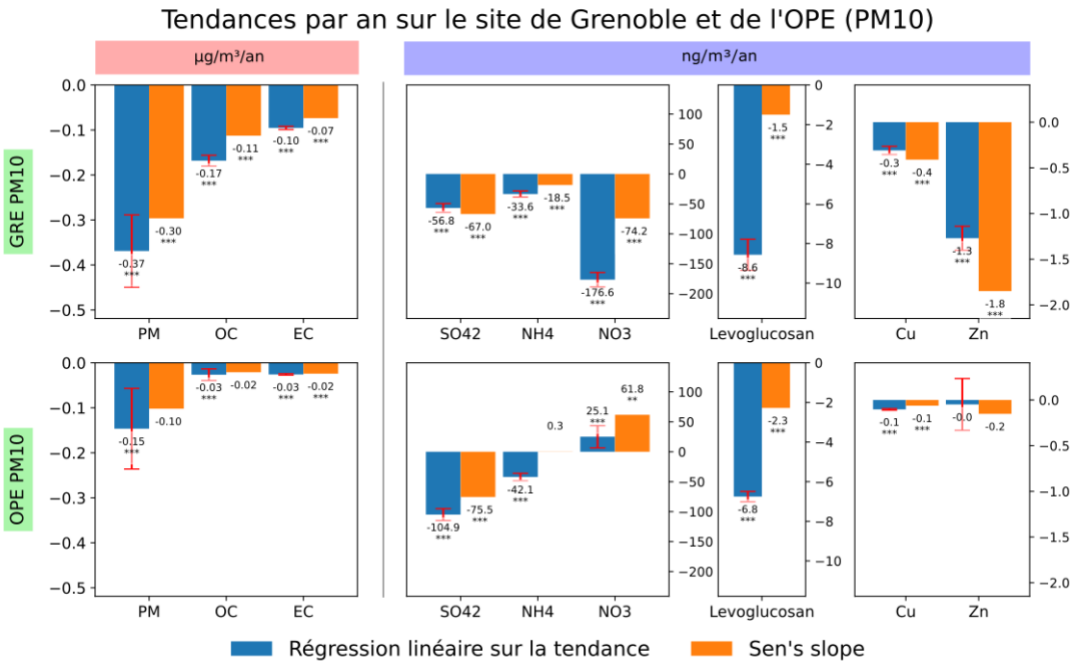


Figure 2 : Comparaison des coefficients de tendances en valeur absolue calculées pour le mode grossier de Grenoble et de l'OPE

Enfin, nous avons développé notre étude en amorçant l'idée d'effectuer des tests de ruptures sur nos données (Figure 3), notamment sur l'élément Carbone Élémentaire. Aucun travail de test ayant été documenté dans la littérature pour ce type de données spécifiques, nous nous sommes orientés en guise d'introduction vers une évolution des tendances.

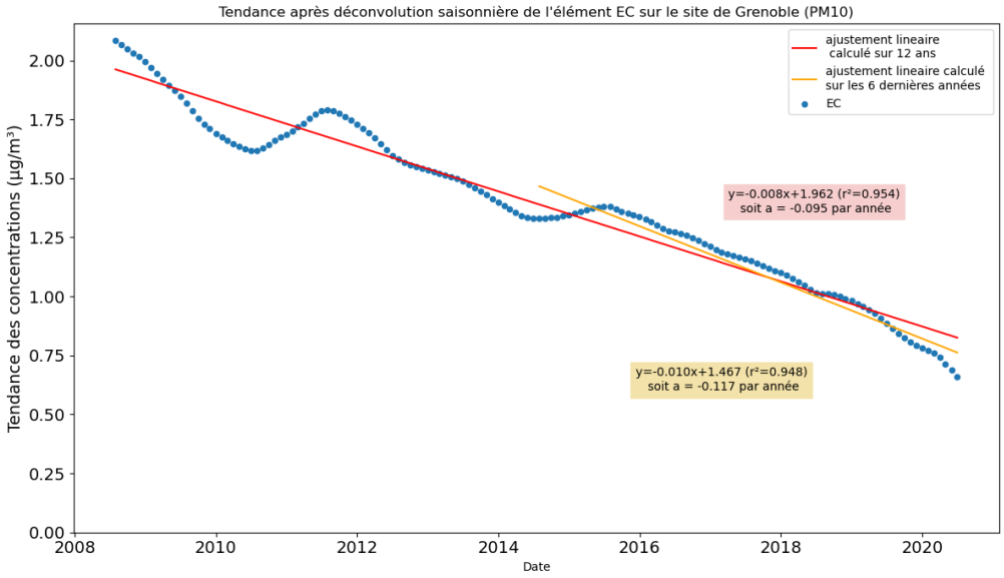


Figure 3 : Régressions linéaires de différentes échelles sur la tendance de la concentration de l'EC sur Grenoble

Au vu des résultats, on constate que la concentration massique en éléments des particules fines suit principalement une diminution significative sur les deux typologies de sites (urbain et rural). Ces résultats montrent d'abord que la mise en œuvre de politiques publiques à grande échelle aurait un impact sur le trafic routier. En effet, les éléments traceurs de cette source comme le carbone élémentaire, le cuivre ou le zinc ont connu des diminutions suite à ces législations. Les tendances sur le site rural, isolé de sources locales, ont également été réduites au niveau des deux fractions de particules, ce qui confirme ce résultat. Deuxièmement, nous observons qu'à Grenoble, les politiques locales semblent accélérer le déclin du carbone élémentaire au cours des cinq dernières années.

Cependant, il a été constaté que l'analyse de l'évolution de la tendance récente nécessite d'être plus approfondie avec l'utilisation de méthodes statistiques appropriées afin de pouvoir tirer des conclusions plus fines. Il serait également possible d'effectuer une analyse en comparant nos données de mesures avec des données de politiques publiques (comme par exemple le nombre de renouvellement d'anciens mode de chauffage). Enfin, après avoir constaté une variabilité inter-annuelle, il est aussi envisageable d'observer si la météorologie aurait une influence sur les particules fines.

Bibliographie

ASEF-Association Santé Environnement France. s. d. « Les particules fines – la synthèse de l'ASEF ». Consulté le 6 mai 2021. https://www.asef-asso.fr/production/les-particules-fines-la-synthese-de-lasef/.

Collaud Coen, Martine, Elisabeth Andrews, Andrés Alastuey, Todor Petkov Arsov, John Backman, Benjamin T. Brem, Nicolas Bukowiecki, et al. 2020. « Multidecadal Trend Analysis of in Situ Aerosol Radiative Properties around the World ». *Atmospheric Chemistry and Physics* 20 (14): 8867-8908. https://doi.org/10.5194/acp-20-8867-2020.

Hussain, Md., et Ishtiak Mahmud. 2019. « PyMannKendall: A Python Package for Non Parametric Mann Kendall Family of Trend Tests. » *Journal of Open Source Software* 4 (39): 1556. https://doi.org/10.21105/joss.01556.

Hussain, MM. s. d. *pymannkendall · PyPI* (version 1.4.2). Consulté le 6 mai 2021. https://pypi.org/project/pymannkendall/.

Hyndman, R.J., & Athanasopoulos, G. 2018. *6.6 STL decomposition | Forecasting: Principles and Practice (2nd ed)*. https://Otexts.com/fpp2/.

John Hunter, Darren Dale, Eric Firing, Michael Droettboom and the Matplotlib development team. 2021. *Matplotlib* (version 3.4.2). https://matplotlib.org/stable/index.html.

Michael Waskom. 2021. *Seaborn* (version 0.11.1).

Milan Gocic, Slavisa Trajkovic,. 2013. « Analysis of Changes in Meteorological Variables Using Mann-Kendall and Sen's Slope Estimator Statistical Tests in Serbia - ScienceDirect ». Serbia. https://www.sciencedirect.com/science/article/abs/pii/S0921818112002032.

RobertB. Cleveland, William S. Cleveland, Jea E. McRae, Irma Terpenning. 1990. « STL: A Seasonal-Trend Decomposition Procedure Based on Loess.Pdf ». Sweden: Journal of OfficialStatistics. https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/stl-a-seasonal-trend-decomposition-procedure-based-on-loess.pdf.

Seabold, Skipper and Perktold, Josef. 2010. *statsmodels: Econometric and statistical modeling with python* (version 0.12.2). Python.

https://www.statsmodels.org/stable/generated/statsmodels.tsa.seasonal.STL.html#statsmodels.tsa.seasonal.STL.

Sylvie NGO, Jean-Luc JAFFREZO, Emmanuel MOUSSU, Olivier FAVEZ, Morgane SALOMON, Alexandre ALBINET, Alexandre THOMASSON, et Jean-Luc BESOMBES. 2018. « LCSQA_2018_Metrologie_PMbiomasse Grenoble.pdf ». https://www.lcsqa.org/system/files/media/documents/LCSQA_2018_Metrologie_PMbiomasse%20Grenoble.pdf.

The pandas development team. 2020. *Pandas* (version 1.2.4). Python. https://doi.org/10.5281/zenodo.3509134.

Unknown. 2021. « Theil–Sen Estimator ». In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Theil%E2%80%93Sen_estimator&oldid=1007270203.

———. s. d. « Sen's estimator of slope ». Consulté le 31 mai 2021a. https://influentialpoints.com/Training/sens_estimator_of_slope.htm.

———. s. d. « Tests de tendance de Mann-Kendall ». XLSTAT, Your data analysis solution. Consulté le 31 mai 2021b. https://www.xlstat.com/fr/solutions/fonctionnalites/tests-de-tendance-de-mann-kendall.

Zaiontz, Charles. s. d. « Sen's Slope | Real Statistics Using Excel ». Consulté le 21 mai 2021. https://www.real-statistics.com/time-series-analysis/time-series-miscellaneous/sens-slope/.

Annulé

HYBRIDATION NUMERIQUE : LA REUSSITE DES ETUDIANTS A EPITECH

Nefeli Paparisteidi[1] Sonico Samedy[2] & Aurélien Pellet[3]

[1] nefeli.paparisteidi@epitech.eu
[2] *sonico.samedy@epitech.eu*
[3] *aurelien.pellet@epitech.eu*

Résumé

Cette étude vise à révéler l'expérience d'apprentissage de ceux qui participent à EPITECH, un établissement d'enseignement supérieur orienté vers le numérique et l'informatique pendant les années universitaires 2018/2019, 2019/2020 et 2020/2021, où l'apprentissage en ligne est devenu une nécessité en raison de la pandémie COVID-19. À travers cette étude, nous souhaitons examiner la performance et l'adaptation des étudiants à l'apprentissage hybride en comparant le revenu de leurs parents. Les résultats de l'étude montrent que les étudiants sont capables de s'adapter à l'apprentissage hybride indépendamment de leurs revenus familiaux. Les résultats de l'étude contribueront à une meilleure compréhension de l'impact de l'e-learning sur l'adaptation et la performance des étudiants dans l'enseignement supérieur.

Mots-clés : Enseignement supérieur, Hybridation numérique, Réussite.

Abstract

This study aims to unveil the learning experience of those who participate in EPITECH, a digital and computer science oriented higher education institution during the academic years 2018/2019, 2019/2020 and 2020/2021, where the e-learning became a need due to COVID-19 pandemic. Through this study we wish to examine students' performance and adaptation to hybrid learning comparing status. In line with research questions and the selected overall approach, the project will adopt a statistical analysis design. The results of the study shows that students were able to adapt to hybrid learning regardless their financial background. The findings of the study will contribute to a better understanding of the impact of the e-learning on students adaptation and performance in Higher Education.

Keywords: Higher Education, Digital Learning, Students' performance

1. Introduction

Le Depuis des décennies, l'enseignement est envahi par une pléthore des outils informatique mais nombreuses institutions n'adhéraient pas toujours à cette approche. Aujourd'hui l'enseignement est plus riche que jamais en termes des richesses informatique et vu la nécessité absolue de la continuité pédagogique celui a adopté un modèle à distance.

En 2019, le terme enseignement à distance est universellement reconnu avec l'arrivée de la pandémie du COVID-19 qui a bouleversé le rythme de l'enseignement supérieur. Les pays ont mis en œuvre des plans d'urgence pour ralentir et limiter la propagation et affronter une éventuelle interruption à long terme de la fréquentation scolaire et universitaire (OECD, 2020). La pandémie a créé une rupture pédagogique sans précédent (Paparisteidi & Lozano Gonzalez, 2021) et l'enseignement numérique à distance a été le seul moyen qui a pu garantir que les étudiants auraient accès à leur éducation.

L'apprentissage en ligne est la pédagogie la plus couramment utilisée pour accéder à des ressources à l'aide d'ordinateurs, de portables, de smartphones et de tablettes. La technologie apporte un avantage supplémentaire dans l'éducation et l'environnement d'enseignement et d'apprentissage. L'apprentissage en ligne présente de nombreux avantages par rapport aux méthodes d'apprentissage

traditionnelles, tels qu'une plus grande accessibilité du matériel d'apprentissage, la communication rapide et la collaboration académique (Alam, et al., 2021).

Les grands défis que représente la mise en place d'un processus éducatif continu dans le cadre de la pandémie de COVID-19 ont stimulé la création d'un nouvel environnement d'apprentissage dans les établissements d'enseignement (Kurbakova, Zlata, & Kurbakov, 2021). Milliers des élèves et les enseignants ont connu des défis et des opportunités grâce à l'enseignement en ligne (Hosszu & Righinis, 2021). Les interactions dans l'école se sont arrêtées, les cours sont passés à une nouvelle modalité, les enseignants et les élèves ont dû s'adapter à ces changements brusques sans avertissement. D'un point de vue sociologique, la structure de la vie quotidienne (Berger & Luckmann, 1966) a été bouleversée.

Dans cette étude on cherche à apporter une première conclusion concernant l'hypothèse suivante ;

Les méthodes de pédagogie active appliquées par EPITECH permettent aux étudiants de s'adapter à un apprentissage numérique hybride en neutralisant l'influence du niveau de revenu de leurs parents.

Pour pouvoir répondre à cette hypothèse, deux populations spécifiques ont été interrogées à l'occasion des trois années de l'étude : d'une part, les étudiants suivant un format à distance et, d'autre part les étudiants suivant une formation en présentiel.

2. Méthodologie

On utilise la base de données des 16406 étudiants de EPITECH. Pour mesurer l'impact du COVID-19 sur la réussite des étudiants. On a divisé notre échantillon de données en deux sous-échantillons. Dans le premier échantillon, on a considéré les étudiants arrivés en 2017 et 2018, avant la pandémie : ils sont 2793. Dans le second sous-échantillon, il y a les étudiants arrivés en 2019 et 2020 et qui ont donc connus la pandémie pendant leur première année d'étude à Epitech. Ils sont au nombre de 2395.

Concernant l'adaptation des étudiants en apprentissage à distance on se base sur un questionnaire délivré aux étudiants pendant l'année scolaire 2020-2021. Les questionnaires étaient distribués par les différentes équipes pédagogiques. Il y a eu 394 réponses, dont toutes ne sont pas exploitables.

Pour vérifier que la performance des étudiants de la première année n'a pas été affectée par le COVID-19, on a utilisé comme outil statistique le test khi-deux d'indépendance de deux caractères et le test statistique de Kolmogorov-Smirnov qu'on analyse ci-dessous :

Le test de Kolmogorov-Smirnov à 2 échantillons est utilisé afin de savoir si deux variables suivent la même distribution. Plus précisément, soient X et Y deux variables aléatoires à valeurs dans un l'ensemble des réels (Corder & Foreman, 2014). Supposons que l'on ait n observation de X et m observations de Y. On peut définir la distribution d'une variable aléatoire comme la façon dont les valeurs observées de cette dernière se répartissent. On suppose que les observations de X sont indépendantes de celles de Y. Ce test nous permet de comparer la distribution de nos deux variables aléatoires. Comme dans tous tests statistiques on pose nos deux hypothèses :

H0 : La distribution de X et la même que celle de Y

H1 : Les deux distributions sont différentes

Nous pouvons calculer grâce à ce test la p-value, plus celle-ci est proche de 1 plus on souhaite conserver notre hypothèse H0. Ou plutôt ne pas le rejeter. Dans l'autre sens, plus la p-value est faible plus on aura tendance à rejeter H0, et à choisir H1.

Test du Chi2 : Soient X et Y deux variables aléatoires qui prennent leurs valeurs dans deux espaces finis, respectivement A et B. On considère que X et Y sont deux variables indépendantes si est seulement si :

$$P[(X=x),(Y=y)]=P(X=x)*P(Y=y) \quad \forall x \in A \text{ et } \forall y \in B$$

De manière plus intuitive, X et Y sont indépendantes si la connaissance de la valeur de X ne nous permet pas de prédire plus précisément la valeur de Y.

Par exemple dans notre étude, on cherche à savoir si le fait d'avoir fait ses études pendant ou avant le covid (X = pendant/avant le covid) est indépendant de la réussite en première année (Y = réussite/échec).

On réalise donc un test statistique du Chi² avec comme hypothèses :

H0 : X et Y sont indépendantes

H1 : X et Y ne sont pas indépendantes.

De la même manière que pour le test de Kolmogorov-Smirnov, on calcul la p-value afin de connaitre la probabilité de nos données observées (ou des données plus extrêmes) sous l'hypothèse H0. Un p-value de 5% par exemple, signifie que les écarts par rapport à notre hypothèse nulle seraient aussi larges ou plus larges que ceux observés dans 5% des cas. On ne peut donc pas conclure que la pandémie à un impact sur la réussite des étudiants de la première année PGE.

3. Résultats

Au cours cet article on va exclusivement aborder le sujet de l'adaptation des étudiants à un apprentissage hybride en testant nos hypothèses.

On commence par présenter dans la Figure-1 l'évolution effectifs des étudiants au sein de EPITECH. On examinera ensuite les différents niveaux d'adaptation des étudiants : la réussite, les problèmes rencontrés et l'impact et l'adaptation au programme d'enseignement hybride.

On les croisera avec d'autres variables que sont le fait d'être entré en première année pendant la pandémie ou non et le salaire des parents des étudiants.

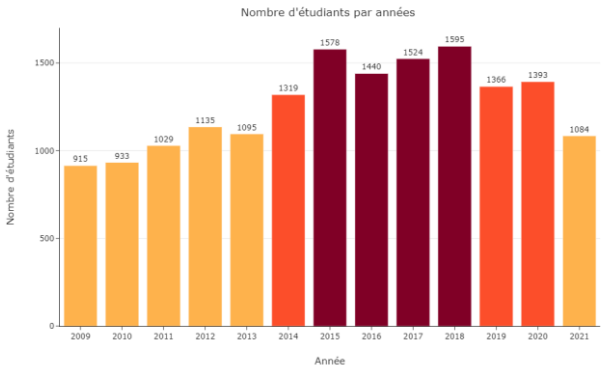


Figure 1 Réussite avant et après covid

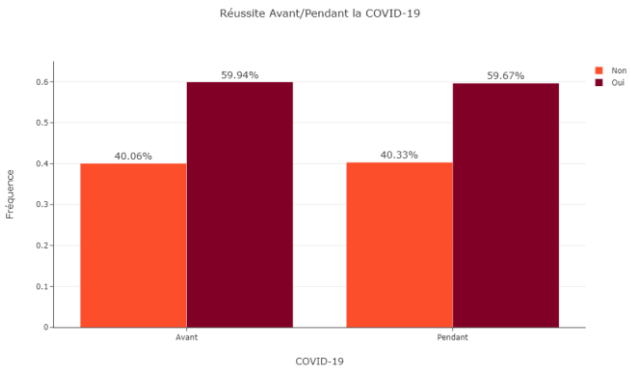


Figure 2 Les effectifs par année

3.1 Réussite des étudiants pendant le COVID

Dans un premier on va considérer une sous-catégorie de l'adaptation : la réussite lors de la première année. On définit la réussite comme étant le passage de l'étudiant en classe supérieure c'est-à-dire que l'étudiant a réussi s'il a validé la première année et si il s'inscrit en deuxième année.

Sur la figure 2 on observe que, pour le groupe d'étudiants avant le COVID-19, le taux de réussite est de 59.94 % contre un taux de réussite de 59.67 % pour le groupe d'étudiants du COVID-19.

Pour vérifier si la performance des étudiants n'a pas été affecté par la crise sanitaire COVID-19, on a réalisé un test khi-deux d'indépendance de deux variables : la période (avant et pendant le COVID) et la réussite. On cherche à savoir si ces deux variables sont indépendantes.

Nos hypothèses sont donc :

H0 : les deux variables sont indépendantes (la réussite n'est pas impactée par le COVID)

H1 : les deux variables sont dépendantes Annulé

On obtient une p-value = 0.865. Ce résultat nous amène à ne pas rejeter l'hypothèse d'indépendance. Les données ne nous permettent pas de conclure que la pandémie à eu un impact significatif sur la réussite en première année.

3.2 Intégration à l'apprentissage hybride et les revenus familiales.

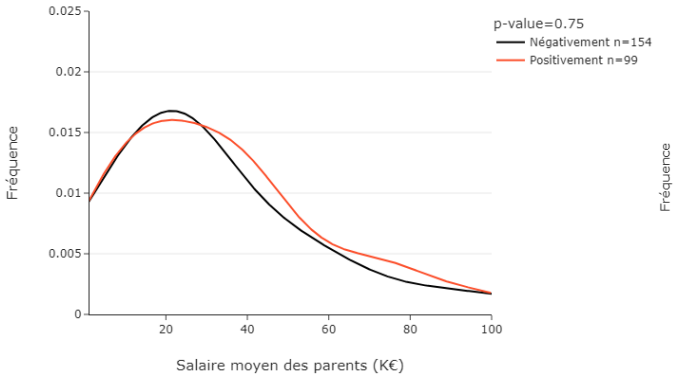
Dans la suite de l'étude on va comparer le salaire des parents dans différentes catégorie afin de savoir si cette variable a un impact sur les capacités d'adaptation à l'enseignement hybride.

Dans la figure 3, on regarde la distribution du salaire des parents selon que l'étudiant ait répondu "positif" ou "négatif" à la question "quel a été l'impact de l'enseignement hybride sur vos relations avec vos collègues''. On réalise un test statistique de Kolmogorov-Smirnov afin de savoir si les deux groupes observés ont des distributions salariales identiques. On pose comme hypothèses :

H0 : La distribution des salaires des étudiants qui ont été affectés positivement est égale à la distribution du salaire des étudiants qui ont été affectés négativement.

H1 : La distributions des salaires des étudiants qui ont été affectés positivement est différente de la distribution du salaire des étudiants qui ont été affectés négativement.

La p-value calculée est 0.75 (75%). On conséquence, l'hypothèse H0 ne peut pas être rejetée. Nos résultants semblent indiquer qu'il n'y a pas de différences significatives de salaire selon l'impact de l'enseignement hybride sur les relations entre les étudiants.



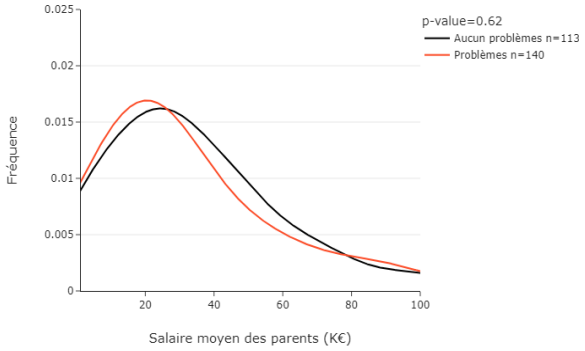


Figure 3 Distribution des salaires selon l'impact de l'enseignement hybride (n=253)

Figure 4 Distribution des salaires selon les problèmes rencontrés (n=253)

3.3 Problème à l'apprentissage hybride et revenus familiales

Ensuite, nous mettons en comparaison les résultats des étudiants qui ne rencontrent aucun problème avec ceux qui rencontrent des problèmes par rapport aux salaires. Dans la figure 4 on observe que la distribution des salaires chez les étudiants ayant rencontrés au moins un problème contre ceux qui ne rencontrent aucun problème est assez similaire. Plus précisément, la courbe des étudiants avec problème touche 20.000 euros de revenus tandis que la courbe des étudiants sans problèmes a atteint son maximum pour 25.000 euros de revenus.

On réalise un test statistique de Kolmogorov Smirnov afin de savoir si les deux groupes observés ont des distributions salariales identiques.

On pose comme hypothèses :

H0 : La distribution des salaires des étudiants qui ont un problème est égale à la distribution du salaire des étudiants sans problèmes.

H1 : La distributions des salaires des étudiants qui ont un problème est différente de la

distribution du salaire des étudiants sans problèmes. Annulé

La p-value calculée est 0.62 (62%). Donc l'hypothèse H0 ne peut pas être rejetée, on ne peut donc pas conclure que les groupe des étudiants ayant rencontrés des problèmes ont des parents avec un salaires significativement différent.

3.4 Capacité à s'intégrer au programme d'apprentissage hybride

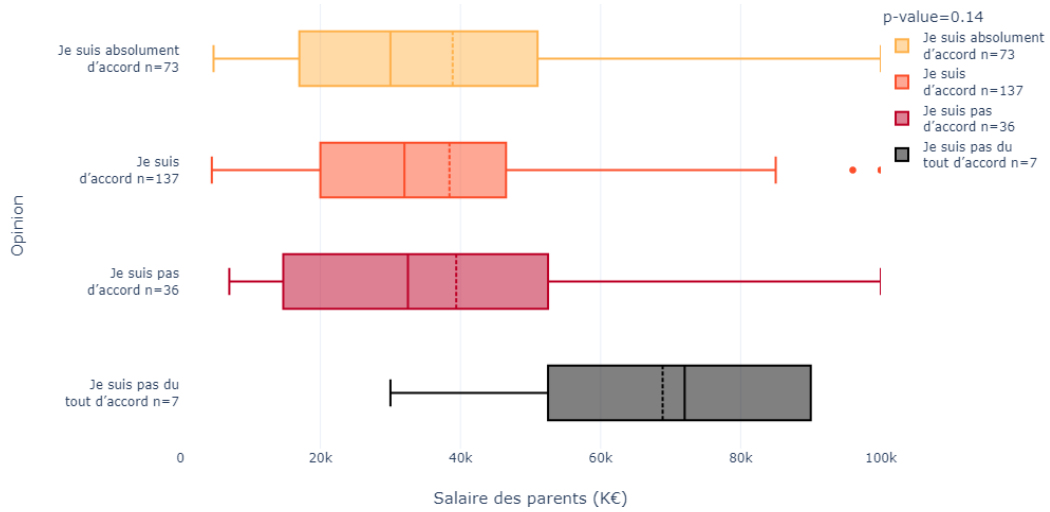


Figure 5 Capacité d'intégration en apprentissage hybride selon les salaires

Dans la figure ci-dessous on affiche des boxplot afin de comparer la distribution des salaires pour chaque catégorie de réponse en fonction de leur adaptation à l'apprentissage hybride. Pour les étudiants qui choisissent comme option le *je suis absolument d'accord*, dont le niveau d'adaptation est assez haut ; on voit que les salaires se concentrent entre 17.000 et 51.000 avec un maximum de 130.000. Par rapport à la catégorie qu'on vient de décrire, la catégorie B, (*je suis d'accord*) concerne les salaires entre 20.000 et 46.500. La catégorie C (je ne suis pas d'accord) a des salaires entre 10.000 et 50.000. Dans ces trois premiers boxplots la médiane et la moyenne sont relativement similaires entre 14.700 et 52.500. Concernant le dernier boxplot, on ne peut pas en tirer de conclusions, l'effectif étant trop faible (moins de 3% des étudiants).

On réalise un test statistique de Kolmogorov-Smirnov avec comme hypothèses :

H0 : La distribution des salaires est la même entre ceux qui s'intègrent et ceux qui ne s'intègrent pas

H1 : il n'y a pas de différence significative de salaire entre ceux qui s'intègrent et ceux qui ne s'intègrent pas

On obtient une p-value de : 0.14 (14%) qui est supérieur au seuil de 5%. En conséquence, on ne peut pas rejeter H0. Les données analysées nous amènent à la conclusion qu'il n'y a pas de différence significative entre ceux qui s'adaptent et ceux qui ne s'adaptent pas à l'enseignement hybride.

À la fin de cette analyse nous avons examiné trois situations séparément pour détecter des éléments de réponse. Premièrement dans les résultats de cette analyse on regarde l'impact de l'enseignement hybride sur les relations entre les étudiants et nous constatons que la variable socio-économique du salaire n'a pas d'impact sur les relations des étudiants avec leurs collègues. Deuxièmement, nous ajoutons la variable des problèmes rencontrés qui nous montre aussi que le salaire ne présente pas des corrélations avec les problèmes rencontrés. Troisièmement, on compare l'estimation, de l'adaptation au programme d'apprentissage hybride avec les salaires des parents qui nous suggèrent également qu'il n'y a pas de différence significative entre les salaires entre ceux qui s'adaptent et

ceux qui ne s'adaptent pas. Les derniers éléments semble converger vers le fait qu'il n'y a pas de différence significative pendant l'adaptation des étudiants à l'enseignement hybride.

4. Conclusion

Pendant cet article on a pu aborder l'hypothèse de notre étude et corréler des données salariales avec le paramètre de l'enseignement hybride. Les différents tests effectués nous amènent à une conclusion : Il ne semble pas y avoir de différences significatives de salaires pour les différents niveaux d'adaptation des étudiants, que ce soit sur les problèmes rencontrés, l'adaptation à l'apprentissage hybride ou l'impact des relations. Cela nous pousse à la conclusion qu'il n'y pas de différence significative selon les revenus familiaux.

Tout nous pousse à valider l'hypothèse que les étudiants ont réussi à s'adapter au format d'apprentissage hybride indépendamment de leurs revenus familiaux.

Bibliographie

Alam, M., Ahmad, N., Naveed, Q., Patel, A., Abohashrh, M., & Khaleel, M. (2021). E-Learning Services to Achieve Sustainable Learning and Academic Performance: An Empirical Study. Sustainability. doi:https://doi.org/10.3390/su13052653

Baron, G.-L. (1994). L'informatique et ses usagers dans l'éducation. Université René Descartes.

Berger, P. L., & Luckmann, T. (1966). The social contruction of reality : A treatise in the sociology of knowlege. Ancor.

Consnefroy, L. (2014). Pédagogie universitaire et pratiques documentaires numériques des étudiants. Dans G. Lameul, & C. Loisy, La pédagogie universitaire à l'heure dunumérique. De boeck.

Corder, G. W., & Foreman, D. I. (2014). Nonparametric Statistics: A Step-by-Step Approach. New Jersey: John Wiley & Sons.

Fouquet, S. (2017). Les parcours et la réussité en Licence, Licence Professionnelle et Master à l'Unviersité. Ministère de l'Éducation nationale, de l'Enseignement supérieur et de la Recherche.

Hosszu, A., & Righinis, C. (2021). Design Issues in e-Learning during the COVID-19. IEEE.

Kurbakova, S. V., Zlata, N., & Kurbakov, A. V. (2021, January 10). Kurbakova, Svetlana and Volkova, Zlata N. and Kurbakov, Alexander V., Developing Students' Cognitive Abilities in E-Learning Environment (January 10, 2021)

Lixi, C., & Seroussi, G. (2021). Hausse sensible des effectifs inscrits dans l'enseignement supérieur en 2020-2021. MESRI-SIES / Systèmes d'information et études statistiques.

Maatuk, A. M., Elberkawi, E. K., Aljawarneh, S., Rashaideh, H., & Alharbi, H. (2021). The COVID-19 pandemic and E-learning: challenges and opportunities from the perspective of students and instructors. Journal of Computing in Higher Education.

Maetz, I. (2017). Les parcours et la réussite en Licence, Licence professionnelle et Master à l'université. Ministère de l'Éducation nationale, de l'Enseignement supérieur et de la Recherche.

Mödritscher, F. (2006). The Impact of an E-Learning Strategy on Pedagogical Aspects. International journal of instructional technology and distance learning.

OECD. (2020, March 23). Education responses to COVID-19: Embracing digital learning and online collaboration.

Paparisteidi, N., & Lozano Gonzalez, E. (2021). ICT in Higher Education: Teachers' experience during COVID-19 pandemic in France, Greece and Mexico. ISRED, 1-7.

Scientific Reform and Visual Data Science: Retiring the EDA/CDA dichotomy

Roger Beecham[1]

[1]*School of Geography, University of Leeds, UK. e-mail: r.j.beecham@leeds.ac.uk*

Abstract

Concerns around the replicability of published scientific findings has prompted much introspection into the way in which scientific knowledge is produced. To address issues of data fishing, searching exhaustively for discriminating patterns in a dataset, picking and then publishing those that are statistically significant, an argument is made that research findings should only be claimed through pre-registered confirmatory data analyses. Pre-registration studies are, though, somewhat inimical to the more informal research environments typical of modern applied data analysis ('Data Science'). In this talk I enumerate some of these challenges and demonstrate, through an analysis of road crash data in the UK, how nascent visualization techniques can be used to navigate and inject statistical rigour into contemporary data analysis environments.

Keywords. Exploratory data analysis; confirmatory data analysis; data visualization; replicability crisis; graphical inference, uncertainty visualization.

1 Scientific Reform and Data Science

A key tenet of Scientific Reform is that research findings are claimed through out-of-sample hypothesis tests, pre-registered in advance (Open Science Collaboration, 2015; Amrhein et al., 2019; Devezer et al., 2021; Szollosi and Donkin, 2021). As an organising framework for this, data analysis is separated into two discrete phases: Exploratory Data Analysis (EDA) and Confirmatory Data Analysis (CDA). EDA makes heavy use of graphical methods in order to discover high-level properties and structure in a dataset and to help formulate plans for further analysis. Once this scoping activity has been completed, a set of hypotheses are recorded (pre-registered) and an entirely new dataset is collected. A formal data analysis, CDA, is then conducted to evaluate whether the new data are consistent with the pre-registered hypotheses. Only at this point can empirical findings can be claimed and submitted for publication.

This $EDA \rightarrow preregistration \rightarrow CDA$ workflow is best suited to mature research settings where the theories and methods are well-developed and where prior studies and

data exist that can be used to judge and evaluate observed effect sizes (McIntosh, 2017). In modern applied data analysis settings, where administrative and passively-collected data are variously combined and repurposed (Singleton and Arribas-Bel, 2021), these certainties are difficult to achieve: the potential analysis space is broad and there is no stand-out approach to formulating research questions, selecting datasets and statistical techniques. Additionally, too strict an adherence to separating EDA (informal pattern-finding) from CDA (statistically-grounded enquiry) may not always lead to strong theory and knowledge development. Attempts to impose pre-specification designs onto research settings that rely on secondary or 'found data' may result in knowledge statements that are overly specific or that do not express sufficient detail or richness in outcomes (Szollosi and Donkin, 2021).

2 Visualization and knowledge development

In a recent discussion in the Harvard Data Science Review, Hullman and Gelman (2021b) make the case for a larger role, or a realignment, of 'exploratory' approaches in knowledge generation. They argue that model development should be intrinsic to EDA. Rather than simply displaying descriptive summaries of observed data, an EDA should support inferential thinking by encouraging analysts to consider, and subsequently model for, the processes that might have generated that structure. A key aspect of exploratory analysis is then enabling detailed comparisons against those reference distributions. If the ambition of a CDA is to accept/reject pre-specified hypotheses, then EDA is concerned with '*the particularities of the discrepancies between model and data*' (Hullman and Gelman, 2021b) — or locating and characterising *where* the data depart from the reference distribution.

Data visualization is instrumental to this sort of activity. Hullman and Gelman (2021b) and others (Hullman and Gelman, 2021a; Heer, 2021; Cook et al., 2021) envision data graphics supporting iterative comparison to models of increasing complexity and sophistication. This may happen either explicitly, for example in graphical inference where plots are compared to others generated under simulated data (Buja et al., 2009; Wickham et al., 2010; Beecham et al., 2017; Morris et al., 2019), or indirectly by the way in which the plot is composed and read.

3 Visualization and spurious discovery

A charge against elevating the status of exploratory analysis in this way is that visual approaches are vulnerable to over-interpretation and false discovery. Data graphics emphasise data patterns, but the complexities around those patterns may be overlooked (Hullman and Gelman, 2021b). It is worth mentioning two counter-arguments here.

First is that Hullman and Gelman (2021a) invoke the idea of graphical inference as model check (Gelman, 2004) – where observed data are compared graphically to reference

data replicated under a model. They signpost to recent work on uncertainty visualization and a growing repertoire of empirically tested techniques for representing these reference distributions, deliberately designed to protect against spurious discovery. These techniques have gained widespread use in scientific communication (Data Journalism) and for which there are software libraries to support easy implementation (cf. Kay, 2021a,b).

Second is more fundamental to Hullman and Gelman (2021b)'s central thesis. Rather than constraining graphical inference as model check to narrow null hypothesis tests, as in graphical line-up tests (Buja et al., 2009), Hullman and Gelman (2021a) locate graphical model checks within a Bayesian framework. Their re-thinking of graphical inference accepts flexibility around the features in the observed data being tested – there may be several irregularly-defined estimators – and that expectation will be contingent on analysts' prior knowledge and experience. This is particularly compelling for data analysis settings that rely on 'found data', as conventional statistical procedures may not transfer well to the observations being claimed and therefore requiring scrutiny.

4 SFDS 2022

Motivating this extended abstract was a real data analysis scenario: a project where pedestrian casualties in STATS19 road crash data, the canonical road safety dataset for the UK, were analysed in order identify and justify, with statistical evidence, investment decisions. It was possible to derive many interesting geographic patterns in crash rates from detailed information on crash context and the vehicles and individuals involved. These were nevertheless patterns which could be reported with limited levels of confidence. Uncertainty around datasets, the selection of context variables and of appropriate statistical techniques and reporting mechanisms, meant it was difficult to make crisp statements around priority areas using the sorts of research designs required by the Scientific Reform movement (Devezer et al., 2021). At SFDS 2022 I will outline some of these difficulties, and demonstrate how the model-based visual techniques advocated by Hullman and Gelman (2021b) can be practically implemented to inject greater rigour into exploratory visual analysis activities that remain widespread, and necessary, to modern data analysis.

References

Amrhein, V., D. Trafimow, and S. Greenland
2019. Inferential statistics as descriptive statistics: There is no replication crisis if we don't expect replication. *The American Statistician*, 73(sup1):262–270.

Beecham, R., J. Dykes, W. Meulemans, A. Slingsby, C. Turkay, and J. Wood

2017. Map line-ups: effects of spatial structure on graphical inference. *IEEE Transactions on Visualization & Computer Graphics*, 23(1):391–400.

Buja, A., D. Cook, H. Hofmann, M. Lawrence, E.-K. Lee, D. F. Swayne, and H. Wickham
2009. Statistical inference for exploratory data analysis and model diagnostics. *Royal Society Philosophical Transactions A*, 367(1906):4361–4383.

Cook, D., N. Reid, and E. Tanaka
2021. The foundation is available for thinking about data visualization inferentially. *Harvard Data Science Review*.

Devezer, B., D. J. Navarro, J. Vandekerckhove, and E. O. Buzbas
2021. The case for formal methodology in scientific reform. *R. Soc. open sci*, 8:200805.

Gelman, A.
2004. Exploratory data analysis for complex models. *Journal of Computational and Graphical Statistics*, 13(4):755–779.

Heer, J.
2021. Exploratory analysis and its malcontents. *Harvard Data Science Review*. https://hdsr.mitpress.mit.edu/pub/vszs87oj.

Hullman, J. and A. Gelman
2021a. Challenges in incorporating exploratory data analysis into statistical workflow. *Harvard Data Science Review*. https://hdsr.mitpress.mit.edu/pub/2ym7zm34.

Hullman, J. and A. Gelman
2021b. Designing for interactive exploratory data analysis requires theories of graphical inference. *Harvard Data Science Review*. https://hdsr.mitpress.mit.edu/pub/w075glo6.

Kay, M.
2021a. *ggdist: Visualizations of Distributions and Uncertainty*. R package version 3.0.1.

Kay, M.
2021b. tidybayes: Tidy data and geoms for Bayesian models. R package version 3.0.0.

McIntosh, R. D.
2017. Exploratory reports: A new article type for Cortex. *Cortex*, 96:A1–A4.

Morris, T. P., I. R. White, and M. J. Crowther
2019. Using simulation studies to evaluate statistical methods. 38(11):2074–2102.

Open Science Collaboration
2015. Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716.

Singleton, A. and D. Arribas-Bel
2021. Geographic Data Science. *Geographical Analysis*, 53:61–75.

Szollosi, A. and C. Donkin
2021. Arrested theory development: The misguided distinction between exploratory and confirmatory research. *Perspectives on Psychological Science*, 16(4):717–724.

Wickham, H., D. Cook, H. Hofmann, and A. Buja
2010. Graphical inference for infovis. *IEEE Transactions on Visualization and Computer Graphics (Proc. InfoVis '10)*, 16(6):973–979.

Comprendre l'échec des étudiants en vue d'un enseignement différencié

Sylvie Viguier Pla [1] & Mouna Kamel [2]

May 9, 2022

[1] *Laboratoire de Modélisation Pluridisciplinaire et Simulations, Université de Perpignan via Domitia, 66000 Perpignan, France, viguier@univ-perp.fr*

[2] *Institut de Recherche en Informatique de Toulouse, Université Paul Sabatier, 31062 Toulouse Cedex, France, kamel@univ-perp.fr*

Résumé. Le niveau de scolarisation étant en constante progression depuis les années 1900, l'accès aux études est devenu plus large et leur durée plus longue, notamment par le biais de l'Université qui offre l'opportunité à tous les étudiants d'accéder à au moins une formation. Cette évolution est accompagnée, notamment au cours des dernières décennies, d'une importante augmentation des échecs, des abandons et des réorientations, plus spécifiquement chez les étudiants de 1ère année du premier cycle universitaire. Différentes études ont identifié les éléments pouvant être à l'origine de ces taux d'échec, ceci à l'échelle d'une université, d'un pays, voire au niveau international. L'étude que nous présentons aujourd'hui s'inscrit dans cette démarche, à la différence qu'elle vise à promouvoir l'enseignement différencié en se focalisant sur l'identification des facteurs d'échec pour une formation précise, en un lieu donné. Nous avons analysé la formation STID d'IUT de Carcassonne. Les critères identifiés comme facteurs sont le type de bac, la CSP du parent 1, le sexe et le fait d'être néo-bachelier ou pas. L'échec est aussi lié à la motivation. La méthodologie mise en œuvre est reproductible pour toute formation.

Mots-clés. enseignement supérieur, abandon, facteurs d'échec

Abstract. As the level of education has constantly increased since the 1900s, more and more people perform studies which are longer and longer, in particular thanks to the University, which offers to all students the opportunity to access at least one training. This development has resulted, particularly in recent decades, in a significant increase in failures, dropouts and reorientations, more specifically among first-year undergraduate students. Various studies have attempted to identify the causes of these failure rates, at the level of a university, a country, or even internationally. The study that we present today takes part of this problem, except that it aims to promote differentiated teaching by focusing on the identification of failure factors for a specific training, in a given place. We analyzed the training STID-IUT located in Carcassonne. The main significant failure factors are the baccalauréat section, the social category of a parent, the gender, and the number of years since the student has its baccalauréat. The failure is also linked with the motivation of the student. The implemented methodology is reproducible for any training.

Keywords. university education, dropout, failure factors

1 Introduction

Depuis les années 1900, une progression constante du niveau de scolarisation est observée, comme illustré dans la Table 1 .

Année	1900	1945	1960	1994	Aujourd'hui
Pourcentage des 17-19 ans ayant le bac	1%	4%	10%	59%	plus de 63%

Table 1: Niveau de scolarisation de 1900 à nos jours selon Millet (2012)

Cela a permis et permet encore aujourd'hui un accès plus large et plus long aux études, notamment par le biais de l'Université qui, de par la différenciation et la hiérarchisation des filières qu'elle propose (filières moins sélectives que d'autres), offre l'opportunité à tous les étudiants d'accéder à au moins une formation. Cette évolution est accompagnée, notamment au cours des dernières décennies, d'une importante augmentation des échecs, des abandons et des réorientations, plus spécifiquement chez les étudiants de 1ère année du premier cycle universitaire. Ces échecs/abandons/réorientations concerneraient 1/3 de ces étudiants au niveau de l'OCDE selon Romainville (2000), tandis qu'en France, seuls 48% des étudiants passent en 2ème année sans redoubler, alors que 30% redoublent, 16% se réorientent et 6% abandonnent définitivement leurs études selon Morlaix et Perret (2018). Une étude du Ministère de 2007 montre que ce sont les étudiants d'origine populaire et détenteurs d'un baccalauréat professionnel ou technologique qui le plus abandonnent leurs études au cours du premier cycle universitaire ou qui sortent de ce cycle universitaire sans diplôme.

Plusieurs éléments peuvent être à l'origine de ces taux d'échec. Nombre d'auteurs pointent le phénomène d'intégration lié à la transition études secondaires/études supérieures, où les caractéristiques cognitives (stratégies d'apprentissage et maîtrise des préalables aux apprentissages, notamment la langue d'enseignement) et le passé scolaire de l'étudiant sont déterminants, même si Morlaix et Suchaut (2012) estiment que ces variables ne sont plus discriminantes à l'entrée à l'université, le tri ayant été fait avant. Un autre élément également largement pointé par les sociologues relève du manque de projet personnel et professionnel pour certains étudiants, qui souvent sont soucieux de trouver une voie professionnelle, et ne sont pas forcément enclins aux études prolongées. Une étude quelque peu contradictoire montre cependant que ce critère ne peut être retenu dans le cadre des IUT par exemple où les taux d'échecs/abandons/réorientations sont faibles alors que les taux d'absence de projet professionnel sont élevés selon Bodin et Millet (2011). D'autres facteurs peuvent également expliquer ces échecs, comme les programmes d'étude, les pratiques pédagogiques, les dispositifs d'accompagnement, les conditions socio-démographiques.

En complément de ces approches qui analysent des comportements d'étudiants indépendamment du type de formation suivie, l'approche que nous développons ici vise à identifier les facteurs d'échec ou d'abandon des étudiants pour une formation donnée, en un lieu donné, en prenant en compte des critères administratifs, sociologiques, pédagogiques, et environnementaux. Notre objectif, à travers cette étude, est de constituer des profils étudiants en fonction de leur faiblesses, afin de leur proposer des outils pédagogiques adaptés à leurs capacités d'apprentissage, et ce dès le début de leur formation. L'idée à terme est donc de développer l'enseignement différencié pour favoriser la réussite des étudiants en 1ère année universitaire.

Les principes méthodologiques que nous avons mis en œuvre sont décrits dans la section 2, la section 3 étant consacrée à la description d'un cas d'usage, soit la 1ère année universitaire en IUT – spécialité STID de l'Université de Perpignan Via Domitia. En effet, le taux d'échec/abandon/réorientation y est relativement élevé (autour de 38%), et l'identification des causes de ces échecs est nécessaire pour mettre en place des mesures pour sa réduction. La section 4 présente un court bilan de cette étude et fournit les perspectives envisagées en prolongement de ces travaux.

2 Principes méthodologiques

Nous rappelons que notre but est de caractériser les facteurs d'échec/abandon par profil d'étudiant. Les étapes de la méthodologie mise en œuvre sont :

- collecter, nettoyer, recoder, agréger un maximum d'informations d'ordre administratif, pédagogique, sociologique, voire environnemental, sur un ensemble d'anciens étudiants ayant suivi une même première année universitaire (même programme pédagogique);
- rechercher les variables environnementales et structurelles liées à l'échec, à l'aide de tris croisés et de tests du khi-deux;
- identifier les modules les plus discriminants par rapport au résultat échec/réussite à l'aide de méthodes de régression logistique;
- Utiliser ces modèles prédictifs pour anticiper d'éventuels échecs et mettre en œuvre un enseignement différencié.

3 Cas d'usage : IUT-STID Carcassonne

3.1 Contexte

L'IUT de Perpignan est composé de 8 départements, dont 4 sur le site de Perpignan, 2 sur le site de Narbonne et 2 sur le site de Carcassonne. Le département STID, objet de cette analyse, se situe à Carcassonne et souffre d'un taux d'échec/abandon/réorientation important (voir Table 2).

	Inscrits	Abandons/Réorientations	Redoublements
Effectifs	230	89	15
Pourcentages		38%	6%

Table 2: Echecs en DUT STID Carcassonne 1ère année sur la période 2013-2021

Cette situation s'explique par la difficulté de recruter : (1) la statistique et l'informatique décisionnelle sont des disciplines méconnues, (2) la ville de Carcassonne ne bénéficie pas de l'attrait des grandes villes universitaires alentour comme Toulouse, Perpignan ou Montpellier. De fait, dans ce département, le taux de pression est faible, beaucoup d'étudiants s'inscrivent alors que ce choix n'est pas dans leurs premiers vœux. Le défi est donc d'identifier les causes d'échecs/abandons, défi accru par l'avènement du BUT (Bachelor Universitaire de Technologie) et le recrutement plus massif des bacs technologiques préconisé par le Ministère.

3.2 Description des données

Les données que nous avons récupérées et analysées sont issues de deux applications universitaires :

- APOGEE (Application Pour l'Organisation et la Gestion des Enseignements et des Etudiants). Il s'agit d'un progiciel destiné à la gestion des inscriptions et des dossiers des étudiants dans les universités françaises. APOGEE permet de récupérer des données d'ordre administratif.
- ScoDoc. Il s'agit d'un logiciel interfaçable avec APOGEE et destiné au suivi de la scolarité des étudiants. Les données sont d'ordre pédagogique.

Le croisement de ces données permet de connaître pour chaque étudiant son nom, prénom, nationalité, sexe, date et lieu de naissance, situation familiale, adresse, parcours universitaire (1ère inscription à l'université, 1ère inscription dans l'établissement, année d'inscription en STID 1ère année), catégories socio-professionnelles des parents, dernier établissement fréquenté (nom, lieu et année), dernier diplôme obtenu (type, lieu et année), boursier, bac (type, année, mention), nombre d'absences, validation en STID (semestre 1 dit S1, semestre 2 dit S2, année), les moyennes obtenues (dans chaque module et chaque Unité d'Enseignement). Les données ont été anonymisées.

Cette étude porte sur la période 2013 à 2021, correspondant à l'application du dernier Programme Pédagogique National.

La figure 1 présente les effectifs par année de première inscription, entre 2013 et 2020. Chaque année, l'effectif de bac S est le plus élevé. Les bacs technologiques apparaissent à partir de 2015, et augmentent légèrement en 2019, principalement grâce au recrutement des TI2D, mais leur effectif total ne dépasse pas 10 étudiants par an. La raison vient du manque de candidats, malgré un encouragement ministériel à recruter ces bacs dans les filières d'IUT.

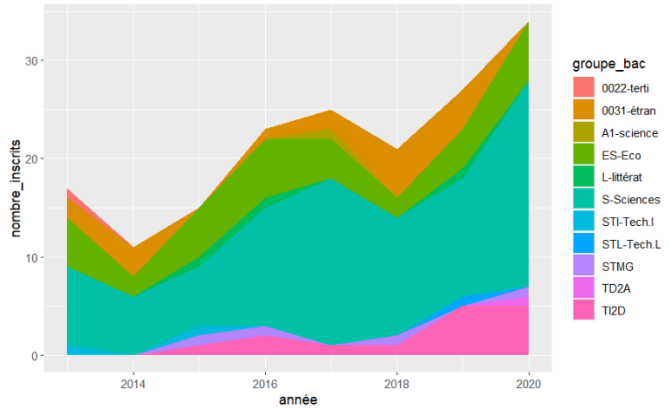


Figure 1: Evolution des nombres de primo-inscrits par bac

.

3.3 Résultats

3.3.1 Tris croisés et tests du khi-deux

Nous avons cherché les facteurs d'échec parmi les caractéristiques suivantes : l'âge, le sexe, la CSP du parent 1, la CSP du parent 2, le bac, le nombre d'années écoulées depuis l'obtention du bac, le fait d'être boursier, et la mention au bac. Une première analyse a montré que l'âge, la CSP du parent 2, le fait d'être boursier et la mention au bac ne sont pas significativement liés à l'échec. Les résultats des tests avec les autres variables sont résumés dans le tableau 3.

	sexe	CSP parent 1	BAC	Nb années depuis bac
Echec S1	ns (p-val=0.19)	s (p-val=0.09)	s (p-val=2.10^{-5})	ns (p-val=0.14)
Echec S2	s (p-val=0.04)	ns (p-val=0.18)	s (p-val=3.10^{-4})	s (p-val=0.07)

Table 3: Tests du khi-deux pour les croisements de l'échec avec divers critères (ns=non significatif, s=significatif)

Au S1, le taux d'échec quand le parent 1 est inactif est le plus élevé (5/8). Remarquons que pour cette catégorie le nombre d'inscrits (8) est faible, nous ne pouvons pas en faire une généralité.

L'échec au S2 n'est plus lié aux CSP des parents, mais nous constatons un taux d'échec plus élevé chez les garçons (ils représentent 90% des échecs alors qu'ils sont 81% des inscrits), et chez les néo-bacheliers (ils représentent 70% des échecs alors qu'ils sont 57% des inscrits).

Sur les deux semestres, la variable la plus liée à l'échec est le type de bac. Détaillons ce point. Sur la figure 2, il apparaît que les différents bacs sont représentés avec des proportions équivalentes chez les néobacheliers comme chez les autres entrants. Nous constatons un fort taux d'échec chez les bacs technologiques. Les bacs étrangers non néobacheliers (donc en réorientation par rapport à un choix antérieur de formation) subissent plus d'échecs que ces mêmes bacs néobacheliers, mais restent à des taux d'échec parmi les plus bas, équivalents à ceux des bacs S. En revanche, les bacs TI2D, S et ES non

néobacheliers accusent un taux d'échec inférieur à ces mêmes bacs respectifs néobacheliers. Les autres bacs sont trop peu nombreux pour que les résultats soient interprétables.

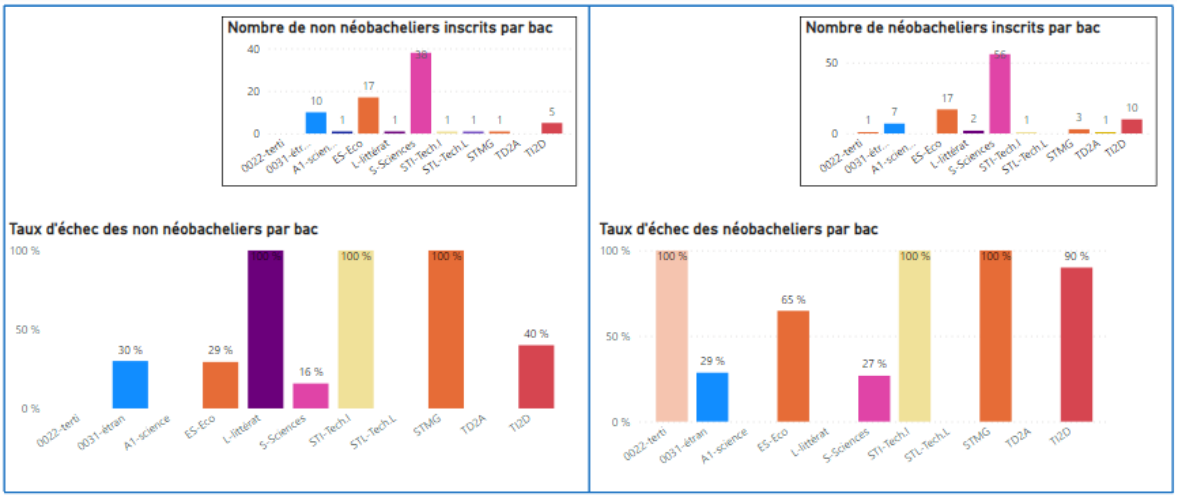


Figure 2: Comparaison des taux d'échec des néobacheliers vs. non-néobacheliers par bac

.

3.3.2 Recherche d'indicateurs avec la régression logistique

L'échec est par ailleurs lié aux performances dans certains modules. Dans la table 4, nous présentons le résultat, pour chaque semestre, de la régression logistique avec sélection stepwise pour détecter, parmi les modules, ceux qui sont le plus liés à cet échec.

```
Echec au S1
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  46.8725    12.5099   3.747 0.000179 ***
M1103        -0.5258     0.3276  -1.605 0.108525
M1201        -0.4229     0.2030  -2.083 0.037242 *
M1202        -0.4078     0.2219  -1.838 0.066023 .
M1203        -0.6955     0.2851  -2.440 0.014693 *
M1204        -0.6488     0.2956  -2.195 0.028159 *
M1302        -0.5648     0.2906  -1.944 0.051942 .
M1303        -0.8327     0.3002  -2.774 0.005540 **
M1403        -0.5268     0.2342  -2.249 0.024500 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Echec au S2
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  66.9380    23.6538   2.830  0.00466 **
M2101        -0.9139     0.3923  -2.330  0.01982 *
M2102        -0.7371     0.3313  -2.225  0.02607 *
M2103        -0.7600     0.3316  -2.292  0.02190 *
M2104        -0.9519     0.3719  -2.560  0.01047 *
M2201        -1.1363     0.4533  -2.506  0.01220 *
M2301        -0.9443     0.3745  -2.521  0.01169 *
M2304        -0.5233     0.2298  -2.277  0.02277 *
M2401        -0.5854     0.2931  -1.997  0.04580 *
M2402        -0.5205     0.2980  -1.747  0.08066 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table 4: Résultat des régressions logistiques avec sélection stepwise expliquant l'échec en S1 et en S2

Les modules liés aux échecs sont, dans l'ordre décroissant d'importance, le M1303 ("Outils de pilotage 1"), le M1203 ("Etudes statistiques et enquêtes"), le M1403 ("Anglais de spécialité"), le M1204 ("Mathématiques pour les probabilités et la statistique") et le M1201 ("Statistique descriptive 2") pour le S1 ; le M2104 ("Mathématiques pour l'analyse des données"), le M2301 ("Économie générale et management des organisations"), le M2201 ("Développement logiciel et technologies web") et le M2101 ("Initiation à la statistique inférentielle") pour le S2.

Nous constatons que les modules réputés les plus académiques, comme les mathématiques, les probabilités ou l'informatique sont des indicateurs secondaires de l'échec, surtout au S1. Ce sont les modules de mise en situation pratique (enquête et outils de pilotage) qui discriminent le plus ceux qui vont réussir

de ceux qui vont échouer. Remarquons aussi que le M1103 ("Projet Personnel et Professionnel") apparait comme un indicateur, mais faiblement significatif. Ceci est assez surprenant, et nous dit que le niveau de l'étudiant n'est pas le seul facteur d'échec : la façon dont il se projette dans le métier auquel il se prépare est révélateur de la façon dont il s'investit dans les applications. Au S2, il apparaît que les modules scientifiques en périphérie du métier de statisticien sont délaissés par les étudiants en difficulté.

Si nous analysons par type de bac, les causes de l'échec sont plus diffuses pour les bacs généraux, au S1. Bien que les modules soient moins significatifs, nous constatons que le PPP (M1103) est toujours un indicateur d'échec, avec les modules M1201, M1204, plus le M1303. Pour les autres groupes de bacs, les modèles ne donnent pas de module significatif par manque de données.

Regardant plus particulièrement les unités d'enseignement du cœur de métier (UE11 et UE21 pour la statistique, UE12 et UE22 pour l'informatique, cf. table 5), nous retrouvons la grande difficulté des bacs technologiques : plus de la moitié des bacs TI2D, et tous les STMG, ont échoué. Les bacs ES sont aussi en difficulté sur ces UE, ils sont plus d'un tiers à échouer.

BAC	N	UE11	UE12	UE21	UE22
S-Sciences	94	11.7	12.8	19.2	20.2
ES-Eco	34	32.4	38.2	44.1	41.2
0031-etr	17	29.4	29.4	29.4	29.4
TI2D	15	53.3	60.0	66.7	60.0
STMG	4	100.0	100.0	100.0	100.0

Table 5: Taux d'ajournés par Unité d'Enseignement (UE)

4 Conclusion et Perspectives

Le département STID de Carcassonne est un département qui a vu, en 8 ans, relativement peu de primo-inscrits, dont une très forte majorité de garçons. De plus, les bacs sont relativement diversifiés (tertiaire, STI, STMG, STL, STD2A, A1, L). De telles données nous amènent à apporter des conclusions seulement en termes de tendances. Les échecs sont plus fréquents parmi les bacs technologiques, et chez les garçons. Ces échecs sont détectables précocément par un manque d'investissement dans les modules liés à l'enquête et aux outils de pilotage. Nous pouvons également cibler les bacs technologiques et ES pour un renforcement adéquat dès leur entrée.

Le prolongement de ces travaux consistera, dans un premier temps, à confronter les tendances que nous avons observées lors de l'analyse du jeu de données du département d'IUT – STID Carcassonne, aux jeux de données fournis par les autres départements STID existant en France. L'étape suivante aura pour objet de développer un outil décisionnel par apprentissage automatique qui permettra à toute formation, de cibler les causes d'échecs/abandons de ses étudiants, sur la base de jeux de données équivalents à ceux présentés dans ce travail.

Bibliographie

Bodin, R. and Millet, M. (2011), L'université, un espace de régulation. L'≪ abandon ≫ dans les 1ers cycles à l'aune de la socialisation universitaire. , *Sociologie, 2011/3 (Vol. 2), p. 225-242. DOI : 10.3917/socio.023.0225. URL : https://www.cairn.info/revue-sociologie-2011-3-page-225.htm*

Morlaix, S. and Suchaut, B. (2012), Les déterminants sociaux, scolaires et cognitifs de la réussite en première année universitaire. *Revue Française de Pédagogie (ISSN : 0556-7807, ISSN électronique : 2105-2913)*

Morlaix S. and Perret, C. (2018), L'évaluation du Plan Réussite en Licence : quelles actions pour quels effets ? : Analyse sur les résultats des étudiants en première année universitaire. *Recherches en éducation, Université de Nantes, 2013, pp.137-150. ⟨halshs-00780580⟩*

Romainville, M. (2000), L'échec dans l'université de masse. *Paris, L'Harmattan.*

TRANSPOSITION DIDACTIQUE ET ENSEIGNEMENT DE LA STATISTIQUE

Antoine Rolland
[1] IUT Lumière Lyon 2, Univeristé Lumière Lyon 2, Université de Lyon
antoine.rolland@univ-lyon2.fr

Résumé.
La plupart des travaux sur la didactique de la statistique dans l'enseignement supérieur se focalisent sur l'enjeu de présenter la statistique à des étudiants a priori réfractaires. Au contraire, nous interrogeons les choix pédagogiques effectués par les enseignants dans les filières de formation en statistique. A travers l'exemple de la régression linéaire, et en mobilisant le cadre de la transposition didactique, nous analysons ces choix par l'étude de documents de cours dans six formations de différents niveaux. Cette étude met en évidence d'une part l'existence d'un « savoir savant » partagé, et d'une autre part un « savoir à enseigner » témoin d'une tension entre théorie généraliste et application pratique, tension propre à la science statistique.

Mots-clés. transposition didactique, statistique, enseignement supérieur, régression linéaire

Abstract. – Most of the didactic works about statistics in higher education focuses on the issue of introducing statistics to supposedly refractory students. On the contrary, we focus in this paper on statistics training as a matter of interest, by questioning the educational choices made by teachers in these specific field. Through the example of linear regression, we analyze these choices by studying course documents in six training courses at different levels. We mobilize the formal framework of didactical transposition. On the one hand, this study highlights the existence of a common body knowledge on the subject, and on the other hand it shows a "knowledge to be taught" in tension between generalist theory and practical application, which tension is specific to statistical science.

Keywords. didactical transposition, statistics, higher education, linear regression

1. Introduction

1.1 Didactique et statistique

La didactique de la statistique représente un champ marginal (en quantité) mais cependant actif de la didactique mathématique (Rolland 2020). Il existe relativement peu de travaux d'analyse de l'enseignement de la statistique concernant directement l'enseignement supérieur, et la plupart des travaux s'intéressant à la didactique de la statistique étudient la question des blocages, des difficultés et des préjugés envers la statistique d'étudiants non statisticiens (voir Régnier 2012 et Hahn 2015).
En particulier, la question de l'enseignement de la statistique de manière spécifique pour des futurs statisticiens n'est que très peu abordée dans la littérature. Est-ce parce que les étudiants étant supposés être attirés par la matière, il n'y aurait pas à s'interroger sur la manière d'enseigner la statistique ? Nous pensons au contraire qu'une approche didactique spécifique mériterait d'être analysée et développée.

Nous présentons ici un premier travail d'exploration, consistant en une analyse comparée de situations d'enseignements d'un même concept statistique dans des formations de différents niveaux (Rolland 2022). Cette première analyse nous permettra de dégager les invariants et les différences

dans la manière d'aborder le sujet considéré. Nous nous appuierons sur le concept de transposition didactique pour expliciter les choix conscients ou inconscients fait par chaque enseignant, et ainsi essayer de dégager des pistes de réflexions sur la manière dont inconsciemment ou non chaque enseignant adapte son enseignement au public spécifique auquel il s'adresse.
Pour ce travail, nous avons retenu la notion de régression, ou modélisation linéaire. Il nous semble que cette notion peut être un bon support à une première étude car le champ de la modélisation linéaire couvre une étendue de savoirs et savoir-faire allant de notions relativement simples pour le modèle de base à de nombreux développements complexes dans le cadre d'un modèle linéaire généralisé. En outre la modélisation linéaire est au fondement de la démarche de la modélisation statistique, et se trouve donc forcément enseignée dans une formation de statistique de niveau Bac+2 à niveau Bac+5.

1.2 La transposition didactique

Le concept de transposition didactique, développé par Chevallard (1985), permet de mettre en lumière l'existence simultanée de plusieurs niveaux de savoir correspondant à un domaine donné : d'une part le savoir savant, et d'autre part le savoir à enseigner et, ce qui lui est proche, le savoir enseigné. Par exemple, le savoir savant « l'opérateur additif est commutatif dans le corps des réels » devient dans le programme de mathématique de la classe de sixième « Connaître des propriétés de l'addition […] et notamment : 12 + 199 = 199 + 12 ». Ce savoir à enseigner deviendra certainement dans une classe un savoir enseigné sous la forme « on peut changer le sens les termes d'une addition, cela revient toujours au même ».
Dans le cadre théorique défini par Chevallard, ce processus de transposition est le fait (en grande partie) de la noosphère, nom désignant un regroupement informel d'acteurs, pour la plupart institutionnels, qui vont délimiter le savoir savant, et figer et légitimer le passage du savoir savant au savoir enseigné. Dans le cadre universitaire, une noosphère formelle n'existe généralement pas, ou simplement à l'état de trace. Les programmes, quand ils existent au-delà de l'intitulé du cours, sont écrits par ceux-là même qui vont les mettre en œuvre et se limitent généralement à une liste de quelques notions présentes sur la plaquette de la formation à destination des étudiants. Aucune préconisation pédagogique précise n'est liée au contenu des cours. Il n'y a pas, sauf rares cas comme le DUT/BUT par exemple, d'instance nationale pilotant et tranchant le contenu de ces cours. Il n'existe pas de manuels scolaires mais des ouvrages de référence (du savoir savant), et des ouvrages destinés aux étudiants (cours et exercice) mais non rattachés à une formation ou un format de cours particulier.

1.3 Axes d'analyse

En mobilisant le cadre de la transposition didactique, nous formulons deux axes d'analyse de l'enseignement de la statistique (et en particulier de la régression linéaire) dans des formations supérieures spécifiquement identifiée comme des cursus en statistique.

- Le premier axe a trait à la cohérence : l'absence de noosphère formelle ne devrait pas nuire à l'homogénéité des contenus entre formation. Nous espérons, à travers l'étude des différents cours, mettre en lumière ce socle commun de notions consensuelles dans le passage du savoir savant au savoir enseigné.
- Le deuxième axe d'analyse porte sur la tension entre apport théorique formel et application pratique concrète, tension qui parcourt la communauté des statisticiens. Notre intuition est que cette tension se retrouve au sein des formations et est influencée par le niveau de la formation, et son inscription dans un cursus professionnalisant ou non.

2. Méthodologie

Nous avons étudié de manière comparative les supports de cours de 5 enseignants dans 6 formations différentes : DUT STID, L2 MASS, L3 MISASH, M1 (INSA), M1 math/stat et M2 pro math/stat.

L'étude des supports de cours permet de mettre en lumière le « socle commun » de notions abordées de manière incontournable dans le cadre d'un cours sur la régression linéaire. Cette étude permet également de remarquer des différences de traitement de ces notions, ainsi que la présentation ou non de notions complémentaires. Nous avons choisi de focaliser notre analyse sur six thèmes particuliers qui sont apparus à la lecture des documents : l'articulation entre modèle de régression simple et modèle à plusieurs variables explicatives ; estimateur des moindres carrés ; présence de démonstrations des résultats principaux ; étude des cas gaussiens et non gaussiens ; introduction du coefficient de détermination (via le coefficient de régression ou le tableau de l'anova) ; démarche de modélisation et choix de modèle.
Deux d'entre eux se retrouvent dans toutes les formations de manière quasi identique, les quatre autres montrant des divergences. Ces points de divergence permettent de dessiner les choix effectués par chaque enseignant dans la manière de présenter les notions liées à la régression linéaire, et partant d'esquisser un processus de transposition didactique d'un savoir savant partagé à un savoir à enseigner dépendant du contexte d'enseignement.

3. Résultats

Nous avons mené l'analyse comparée des six supports de cours différents suivant les six axes différents. Ces analyses permettent de dégager une vue d'ensemble des choix didactiques effectués par les enseignants à propos de la modélisation linéaire. Un savoir à enseigner a été identifié par déduction, tant dans sa partie théorique que pratique. Il contient les estimateurs des moindres carrés et les tests possibles sur ces estimateurs, la définition du coefficient de détermination, ainsi que l'analyse des conditions d'application du modèle et le choix du meilleur modèle possible.
Ce savoir à enseigner nous semble cohérent avec les ouvrages spécialisés consultés. L'absence de noosphère externe formelle n'empêche donc pas l'apparition d'un socle de savoir à enseigner partagé, confirmant ainsi l'intuition d'observer également dans le domaine de la statistique le constat évoqué par Bosch *et al.* (2021) d'une faible liberté réelle laissée aux enseignants dans la définition du savoir à enseigner. La table ci-dessous résume les différences d'approche entre formation.

Axe d'étude	Approche orientée théorie	Approche orientée pratique
Modèle simple et modèle multiple	Aucune formation	Toutes les formations
Estimateur des moindres carrés	L3	DUT, L2, M1, M1 INSA, M2
Démonstrations	L2, M1, M2	DUT, L3, M1 INSA
Normalité des résidus	L2, M1, M2	DUT, L3, M1 INSA
Coefficient de détermination	L2, M1 – et en partie L3 et M2	DUT – et en partie L3 et M2
Démarche de modélisation	Aucune formation	Toutes les formations

Notre problématique initiale est que la tension entre théorie et pratique au sein de la discipline statistique se retrouve au sein du panel de formations, et est influencée par le niveau de la formation et son inscription dans un cursus professionnalisant ou non. L'analyse des corpus de cours relatifs à la régression linéaire (table 2) permet de constater que cette tension entre présentations théorique et pratique existe bien, mais que les oppositions ne sont pas univoques :

- Toutes les formations considèrent, au moins pour trois axes, une approche orientée pratique : c'est donc le choix d'un équilibre spécifique entre approche théorique et approche pratique qui distingue les formations entre elles.
- L'opposition n'est pas spécifiquement une opposition de niveau, la L2 se comportant comme le M1 et le M2, et le DUT comme le M1 INSA.

- L’opposition n’est pas non plus uniquement entre filière généraliste et technologique, la L3 se retrouvant souvent avec le DUT et le M1 INSA.

Notre analyse des documents ne montre donc pas une opposition franche entre théorie mathématique et pratique empirique. L’analyse des entretiens permet alors de proposer une grille de lecture complémentaire par la perception qu’a l’enseignant des enjeux du cours donné au sein de la filière. L’opposition (relative) se situe alors entre une approche centrée sur la maîtrise de la pratique dans un environnement donné (DUT, L3, M1 INSA) et une approche insistant sur la généralisation possible des techniques utilisées (L2, M1, M2). On attend d’abord de l’étudiant une capacité d’application dans des cas-types dans le cas des DUT, L3, M1 INSA. On attend une compréhension des mécanismes mathématiques à l’œuvre, permettant une généralisation, dans le cas des L2, M1, M2. La tension entre application et théorie générale dans la transposition didactique de la part des enseignements étudiés intègre pleinement la place spécifique de la filière dans le parcours de formation du futur statisticien.
L’absence d’une noosphère formelle dans l’enseignement supérieur ne signifie pas que l’enseignant a toute liberté pour décider du contenu de son cours .Il lui reste cependant de la place pour effectuer des choix de transposition du savoir savant au savoir à enseigner. Cette place est plus ou moins investie par les enseignants, certains faisant des choix pédagogiques explicites et personnels, d’autres se reposant sur des « standards de fait » à travers la reprise de propositions de cours existants par ailleurs. Le caractère collectif de la transposition didactique est peu marqué : au mieux, il reste implicite par les échanges de documents de cours, mais le plus souvent l’enseignant ne peut compter que sur sa propre compréhension des enjeux du cours au sein de la formation pour décider de ses choix pédagogiques. Il en ressort, pour le cadre de la régression linéaire, une tension entre présentation de méthodes pratiques spécifiques et effort de généralisation théorique.
Il est à noter que contrairement à de nombreux domaines des mathématiques pouvant être étudiés dans une perspective purement abstraite, cette tension entre théorisation et application est constitutive de la statistique en tant que discipline savante. Il est donc normal que cette tension apparaisse également dans les cours de statistique destinés à des statisticiens.

4. Perspectives

Cette première approche analytique de la transposition didactique dans l’enseignement de la statistique dans des filières spécialisées nécessite d’être poursuivie pour mettre en évidence le rôle personnel joué par l’enseignant dans cette transposition. En effet, l’analyse effectuée dans cet article est limitée par le fait que nous avons étudié des couples formation/enseignant, sans pouvoir isoler l’effet formation de l’effet enseignant au sein du dispositif. Une étude conduite sur des cours identiques dans des formations comparables devrait permettre d’affiner notre compréhension des enjeux de la transposition didactique dans les cours de statistique pour statisticiens.

Bibliographie

Bosch, M., T. Hausberger, R. Hochmuth, M. Kondratieva et C. Winsløw (2021), « External Didactic Transposition in Undergraduate Mathematics », *International Journal of Research in Undergraduate Mathematics Education*, Springer, 7 (1), pp.140-162.
Hahn, C. (2015), « La recherche internationale en éducation statistique : état des lieux et questions vives », *Statistique et enseignement*, vol. 6, no 2, pp. 25–39.
Régnier, J.-C. (2012). « Enseignement et apprentissage de la statistique : entre un art pédagogique et une didactique scientifique », *Statistique et enseignement*, vol. 3, n° 1, pp. 19–36.
Rolland, A. (2020). « 2009-2019 : dix ans de publications sur l’enseignement de la statistique en France », *Statistique et Société*, vol. 8, n°1, pp 55–71.

Application : santé & société

Application des réseaux de neurones aux données longitudinales de santé : cas du suivi de la fonction rénale aux patients transplantés rénaux

C. Benoist[1] & M. Labriffe[1,2]& A. Åsberg[3] & P. Marquet[1,2] & J.-B. Woillard[1,2]

[1] *Service de pharmacologie, Toxicologie et pharmacovigilance, CHU de Limoges,France*
[2] *Pharmacologie & Transplantation, INSERM 1248, Université de Limoges, Limoges, France*
[3] *Departement of Pharmacy, Faculty of Mathematics and Natural Sciences, Oslo, Norway*
Corresponding author : Clément Benoist, `Clement.BENOIST@chu-limoges.fr`

Résumé

Chez les patients transplantés rénaux, le fonctionnement du greffon est régulièrement suivi par l'estimation du débit de filtration glomérulaire (DFG). Plus cette valeur est basse, plus elle indique un défaut de fonctionnement du greffon. Savoir prédire le DFG à horizon de 6 mois ou 1 an permettrait d'anticiper une détérioration du DFG afin de proposer une prise en charge précoce adaptée et éviter un retour en dialyse. La surestimation du DFG aurait pour conséquence l'absence d'alerte du clinicien et d'un suivi plus rapproché pouvant entraîner une augmentation du risque de perte du greffon à moyen terme. Les données de suivi post-transplantation utilisées étaient sous forme de données longitudinales ou de séries temporelles courtes (3819 patients). Nous avons modélisé le DFG à horizon de 12 mois pour chaque patient grâce à un Long Short Term Memory (LSTM). La médiane de l'erreur absolue était 5.2mL/min/1.73m^2, compatible avec une utilisation en clinique de l'algorithme. Le 99^e percentile de l'erreur était 25.4mL/min/1.73m^2.

Mots-clé : réseau de neurones, LSTM, santé

Abstract

For kidney transplant recipients, the graft function is regularly monitored by the estimation of glomerular filtration rate (GFR). The lower this value, the more it indicates a dysfunction of the transplanted kidney. Knowing how to predict the GFR at 6 months or 1 year would anticipate a worsening of the GFR, propose appropriate early therapeutics and avoid a return to dialysis. Over-estimation of GFR may lead to the lack of alert of the clinician, which may result in a medium-term graft loss. The objective of this work was to develop a loss function adapted to this problem. The post-transplantation follow-up data were longitudinal (or a dataset of short time series). We modeled the future GFR at 12 months by using for each patient a Long Short Term Memory (LSTM). A loss function adapted to the medical problem was developed: it penalizes more the errors associated to lower GFR values. The median absolute error was 5.2mL/min/1.73m^2, compatible with a clinical use of the algorithm. The 99$^{\text{th}}$ percentile was 25.4mL/min/1.73m^2.

Keywords: neural network, LSTM, health

1 Introduction

Habituellement, chez un patient transplanté rénal, le risque de retour en dialyse est appréhendé par des modèles de survie tels que les modèles de Cox [Lin *et al.*, 2008]. Néanmoins, ces modèles ne prennent pas en compte les dernières valeurs du paramètres le plus utilisé dans le suivi du patient greffé : le débit de filtration glomérulaire (DFG). Cet indicateur de la fonction du greffon est régulièrement estimé chez les patients transplantés, en utilisant une mesure de la créatinine dans le sang. La prédiction dynamique du DFG permettrait de prévenir un retour en dialyse à moyen terme. Un DFG inférieur à 30mL/min/1.73m^2 correspond à une insuffisance rénale sévère et une valeur inférieure à 15mL/min/1.73m^2 signe généralement la nécessité d'un retour en dialyse.

Dans Khazaee *et al.* [2019], les prédictions de DFG sont obtenues par des réseaux de neurones. Ils ont néanmoins employé des fonctions de perte classiques, notre travail avait pour objectif de définir une fonction de perte adaptée à la problématique médicale. En médecine, les conséquences d'une erreur de prédiction du DFG sur la survie du greffon ne dépendent pas uniquement du carré de l'erreur. Surestimer le DFG est plus grave que de le sous-estimer. En effet, ceci aura pour conséquence l'absence d'alerte du clinicien et d'un suivi plus rapproché pouvant entraîner une augmentation du risque de perte du greffon à moyen terme. Pour des valeurs assez basses, il est primordial de ne pas manquer le diagnostic d'une insuffisance rénale sévère (DFG inférieur à 30mL/min/1.73m^2). Cela nous incite à développer des fonctions de perte adaptée à ces problèmes. La fonction de perte associée à l'équation (1) permet de répondre à cette spécificité médicale.

2 Données

Quatre jeux de données pseudonymisés étaient disponibles pour l'étude : des données longitudinales du suivi après la transplantation (traitement immunosuppresseur utilisé, créatininémie, protéinurie...), les données relatives au receveur (âge, groupe sanguin, nombre d'incompatibilités HLA, statut sérologique par rapport à diverses infections), des données relatives à la greffe (temps d'ischémie froide, rang de la greffe...) et un jeu de données indiquant l'âge du donneur. Ces données suivaient 3819 patients avec une médiane de 6 visites par patient, un premier quartile de 3 visites par patient et un troisième quartile de 9 visites par patient. Sauf exception, les données post-transplantation étaient temporellement régulières : il y a des valeurs à 3 mois et des valeurs aux dates anniversaires de la transplantation.

3 Modélisation

La fonction de perte choisie devait particulièrement pénaliser la surestimation du DFG. Par ailleurs, nous avons choisi d'augmenter le poids associé aux faibles valeurs de DFG en utilisant la fonction de perte suivante :

$$L(\hat{y}, y) = \frac{1}{y}(\hat{y} - y)^2. \tag{1}$$

Les données sont séparées entre une base d'entraînement (67% des patients) et une base de test (33% des patients), les deux bases n'avaient aucun patient en commun.

Après exploration graphique et statistique des données, une première idée est de retenir 12 variables dont 7 longitudinales : le délai depuis la transplantation, les DFG précédents, la présence d'un traitement par inhibiteur de l'inosine monophasphate déshydrogénase, le présence d'un traitement inhibiteur de la mTOR (évérolimus ou sirolimus), la présence d'un traitement par corticoïdes, la proténiurie ; et 5 non longitudinales : le statut du donneur (vivant ou décédé), l'âge du donneur, l'âge du receveur à la greffe, le temps d'ischémie froide, la présence d'anticorps anti-HLA de classe II.

Nous avons utilisé une approche de réseau de neurones récurrents par utilisation des Long Short Term Memory (LSTM) [Hochreiter and Schmidhuber, 1997]. Rappelons les formules pour une cellule de LSTM qui se fonde sur la récurrence :

$$\begin{aligned}
\vec{\imath}_t &= \sigma(\mathbf{W}_i^{(i)}\vec{x}_t + \vec{b}_i^{(i)} + \mathbf{W}_h^{(i)}\vec{h}_{t-1}),\\
\vec{f}_t &= \sigma(\mathbf{W}_i^{(f)}\vec{x}_t + \vec{b}_i^{(f)} + \mathbf{W}_h^{(f)}\vec{h}_{t-1}),\\
\vec{g}_t &= \tanh(\mathbf{W}_i^{(g)}\vec{x}_t + \vec{b}_i^{(g)} + \mathbf{W}_h^{(g)}\vec{h}_{t-1}),\\
o_t &= \sigma(\mathbf{W}_i^{(o)}\vec{x}_t + \vec{b}_i^{(o)} + \mathbf{W}_h^{(o)}\vec{h}_{t-1}),\\
\vec{c}_t &= \vec{f}_t \odot \vec{c}_{t-1} + \vec{\imath}_t \odot \vec{g}_t\\
\vec{h}_t &= \vec{o}_t \odot \tanh(\vec{c}_t)
\end{aligned}$$

où σ représente la fonction sigmoïde $\sigma : \mathbb{R} \to \mathbb{R}; u \mapsto \frac{1}{1+\mathrm{e}^{-u}}$, $\vec{h}_t$ représente les états cachés au temps t ($\vec{h}_0$ représente les états cachés initiaux), $\vec{c}_t$ est l'état de la cellule au temps t, $\vec{x}_t$ est l'entrée au temps t correspondant au DFG, $\vec{\imath}_t$, $\vec{f}_t$, $\vec{g}_t$, $\vec{o}_t$ désignent respectivement les portes d'entrée, d'oubli, de cellule et de sortie. Le symbole $\odot$ représente le produit de Hadamard (multiplication élément par élément). Pour renforcer la profondeur du réseau de neurones, deux niveaux de LSTM ont été utilisés. Le nombre d'états cachés du 1er niveau, celui du 2e niveau, le dropout, le learning rate ont été échantillonnés au moyen d'une grille aléatoire parmi 500 jeux de paramètres. Le nombre optimal d'états cachés des LSTM du 1er et du 2e niveau était de 18 et 13 respectivement. À la sortie du deuxième LSTM, une couche pleinement connectée avec une sortie sans fonction d'activation permettait d'obtenir directement le DFG. Pour des raisons d'applicabilité, 4 observations

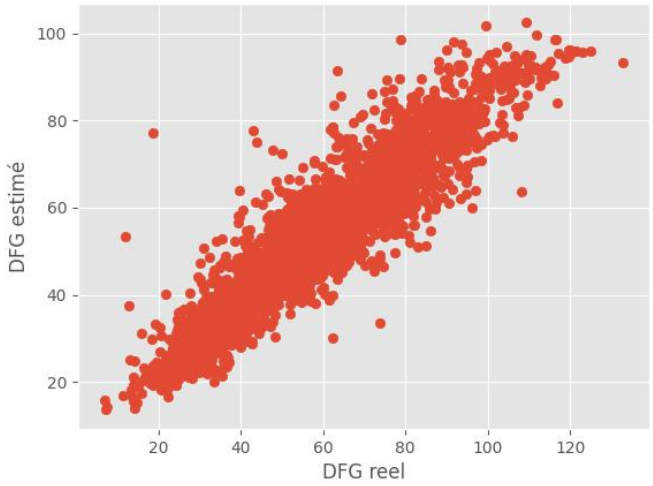


FIGURE 1 : DFG prédit par le réseau de neurones en fonction du DFG réel

précédentes ont été utilisées pour la prédiction du DFG à horizon de 12 mois (la 1re prédiction survenant à 48 mois après la greffe par utilisation de la valeur observée à 3, 12, 24, 36 mois). L'optimisation a été effectuée sur des batchs de 512 éléments par un optimiseur d'Adam avec un learning rate optimisé de $3.3 \cdot 10^{-4}$.

Rappelons que le DFG est une mesure de la qualité du fonctionnement des reins. Plus le DFG est élevé, mieux les reins fonctionnement. Un DFG bas signe une insuffisance rénale.

Notons l'erreur relative : $\hat{\varepsilon} = \hat{y} - y$. Cette erreur n'était pas centrée sur 0. En effet, sa médiane était de $-2.35\text{mL/min/1.73m}^2$. Dans ce contexte médical, il est préférable de sous-estimer la valeur de DFG plutôt que de la surestimer car la sous-estimation conduirait à une surveillance renforcée du patient sans conséquence sur son état de santé. L'erreur quadratique moyenne (MSE) était de 76.2. Notons que les DFG sont toujours strictement positifs. Les figures 1 et 2 montrent respectivement le DFG estimé et l'erreur de DFG en fonction du DFG réel.

Dans le domaine médical, nous ne nous intéressons pas seulement à des valeurs centrales de l'erreur, mais aussi aux erreurs extrêmes, car elles peuvent avoir de lourdes conséquences cliniques. Ainsi, le 99e percentile de l'erreur était de $25.4\text{mL/min/1.73m}^2$. Cette erreur sur le DFG peut paraître élevée, mais, d'après la figure 2, ces valeurs correspondent majoritairement à une sous-estimation du DFG (résidus du DFG négatifs ou erreurs relatives négatives) qui engendre l'alerte à tort du clinicien qui proposera vraisemblablement au patient à un suivi renforcé, mais cette sous-estimation n'aura pas de conséquence sur l'état clinique du patient.

La figure 2 montrait 3 valeurs surestimées pour une valeur réelle inférieure à

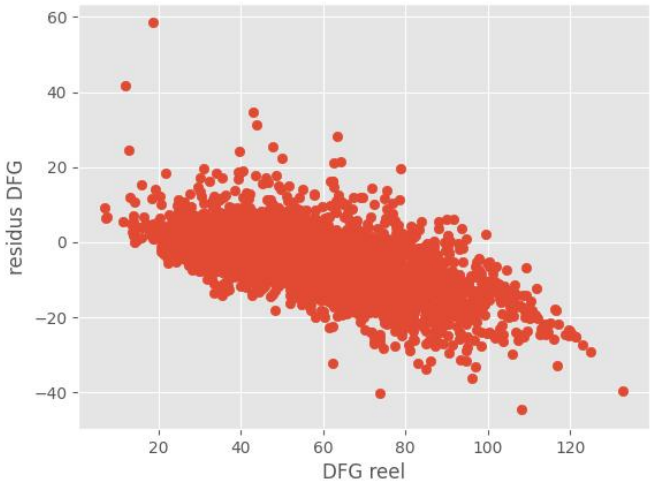


FIGURE 2 : Erreur induite par le réseau de neurones en fonction du DFG réel

20mL/min/1.73m^2 (associées à trois transplantations différentes) correspondant au pire scénario, car il conduirait à une absence de prise en charge renforcée et, possiblement, à une dégradation de la santé du patient. Notons que sur 119 valeurs inférieures à 30mL/min/1.73m^2 (borne supérieure de l'insuffisance rénale sévère selon les recommandations internationales [KDIGO, 2009]), 79 ont été prédites 12 mois à l'avance (soit 69% des cas) correspondant à un résultat relativement bon.

4 Discussion

Le principal apport de la méthode est la fonction de perte développée. Cette fonction de perte était adaptée à la problématique médicale de prédiction du DFG.

Khazaee *et al.* [2019] ont étudié cette problématique en utilisant des fonctions de perte classiques (mean squared error : MSE : mean absolute error : MAE) et en utilisant des couches complètement connectées. Par ailleurs, ils ont réalisé des prédictions à 12 mois avec des tailles de fenêtre testées entre 4 et 10. Leurs meilleures performances étaient obtenues pour une fenêtre de 10 valeurs consécutives (MSE : 142), le RMS associée est 11.9mL/min/1.73m^2. L'utilisation des LSTM dans ce travail permet d'obtenir une performance meilleure (MSE : 76.2). Le RMS associé est 8.7mL/min/1.73m^2. Il est à noter que les intervalles de temps entre les mesures étaient plus courts dans l'étude de Khazaee *et al.* [2019] (6 mois au maximum vs un an pour notre étude). Mon algorithme permet par conséquent une bonne évolution dans l'algorithme de prédiction du DFG. Cette évolution permet vraisemblablement d'anticiper une tendance à l'insuffisance rénale sévère et de la

prévenir grâce à une prise en charge et un suivi adaptés.

Notre étude a un intérêt méthodologique notable, mais est perfectible sur le plan clinique car l'algorithme suppose pour le clinicien d'attendre 3 ans avant d'avoir accès à une prédiction du DFG, ce qui limite son utilisation en clinique.

Bibliographie

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

KDIGO (2009). KDIGO clinical practice guideline for the care of kidney transplant recipients. *American journal of transplantation : official journal of the American Society of Transplantation and the American Society of Transplant Surgeons*, 9, S1–S155.

Khazaee, P. R., Bagherzadeh, J., Niazkhani, Z., & Pirnejad, H. (2019). Predicting the function of transplanted kidney in long-term care processes : application of a hybrid model. *Journal of Biomedical Informatics*, 91, 103116.

Lin, R. S., Horn, S. D., Hurdle, J. F., & Goldfarb-Rumyantzev, A. S. (2008). Single and multiple time-point prediction models in kidney transplant outcomes. *Journal of biomedical informatics*, 41(6), 944-952.

DATA VISUALISATION OF ADVERSE EVENTS IN RANDOMISED CLINICAL TRIALS

Dr Victoria Cornelius[1], Dr Suzie Cro[1], Dr Leila Janani[1] & Dr Rachel Phillips[2]

[1]Imperial Clinical Trials Unit, School of Public Health, Imperial College London

[2]Pragmatic Clinical Trials Unit, Queen Mary University of London

Résumé:

Victoria Cornelius is a Reader in Medical Statistics and Clinical Trials at Imperial College London and Director Designate at Imperial Clinical Trials Unit (ICTU). She leads the Child and Digital Health portfolios and established the ICTU Trial Methodology Group. The Group undertakes research to improve the analysis in clinical trials, including the analysis of adverse events. She also has undertaken research into the design of trials for small populations and rare disease. Her work has primarily been applied in the evaluation of drug interventions and across a range of areas including severe pediatric asthma & allergy, mental health, and critical care. She has been involved and led national collaborative statistical projects for the analysis of laboratory studies and early phase clinical trials research to support methods translation.

Keywords: Adverse events, Harms, Randomised Controlled Trials, Data Visualisation

Background

Methods to analyse efficacy outcomes in randomised controlled trials (RCTs) are well established. As the process of randomisation provides protection against confounding, unadjusted hypothesis tests are frequently used to compare outcomes between arms. [Phillips et al 2019] However, these simple approaches cannot take account of important and frequently occurring features such as repeated measurements and clustering or the benefit for adjustment. With substantial improvements in statistical software it's now trivial to undertake advanced statistical modelling and present this data well. Despite this progress, the analysis and presentation of adverse events (AEs) in trial publications has seen very little progress.

AE data is particularly difficult to analyse due to its multi-faceted nature. One of these features is the large number of AE outcomes that get recorded in a trial. This can be due to the fact that AEs are collected from several sources including spontaneous reports from participants, clinical observations, as well as clinical and biological screening. How to analyse, summarise and present this large volume of data is problematic. In 2016 a collaboration of industry experts and journal editors undertook a consensus and published recommendations to improve AE reporting in trial publications. [Lineberry et al] The recommendations included: containing a summary of clinically important events; reporting important minimum numeric information; and a proposal to encourage the use of graphical displays to visualise AE data.

There is a lack of guidance on what and how to visually display complex AE data. We previously undertook a methodology review to identify statistical methods specifically developed to analyze AE data. [Phillips et al 2020] This current paper examines two visual analysis methods identified to be suitable for any adverse events collected during a trial, and one new approach proposed by the

authors for pre-specified harm outcomes that is particularly valuable for multi-arm studies We explore their value using data from a COVID-19 treatment trial and COVID-19 vaccination trial.

Methods

We first examined two plots identified from a previous methodological review [Phillips et al 2020]. These were the **Volcano plot** [Zink et.al. 2013] and **Dot plot** [Amit et.al. 2008]. Both are suitable for any adverse events collected during the clinical trial (emerging events or pre-specified) and are suitable to display binary, count and time-to-event data. We reanalyzed data reported in two trials. The first was a Covid-19 treatment trial, which was a multi-centre, placebo-controlled trial of remdesivir in adults admitted to hospital with severe COVID-19 infections in Wuhan, China. We extracted aggregated data from the published report to produce the visualisations and compared the interpretation to the original trial report.

We also developed a new approach to visualize data for a COVID-19 vaccine trial we were working on that had a number of pre-specified 'reactogenicity' harm events of interest and multiple arms that we wanted to compare. A visulisation was required to enable us to compare a 'fingerprint' of these harms. In order to facilitate easier between arm comparison and display a summary measure of burden we based our visulisation on a radial graph. The trial compared 7 different vaccines as a 3rd booster vaccine (following 2 original doses) to a control arm. Three of the vaccines were at full and half dose. [Munro et al. 2021] Reactogenicity events are important to understand and compare between different vaccine options as they have economic burden in terms of days off work and school. This information is important for governments to use to inform their policy decisions for population level vaccination deployment.

We developed statistical software in Stata to implement the volcano and dot plot [AEVOLCANO and AEDOT], and the radial graph can be constructed in R using ggradar package.

Results

The **volcano plot** [Zink et.al. 2013] displays the between-arm statistic on the x-axis against the log transformed p-value from a between-arm statistical test of the researcher's choice on the y-axis. An example can be seen in Figure 1. In this plot the absolute risk difference is plotted on the x-axis and the log transformed p-value from the Fisher's exact test is plotted on the y-axis. The summary statistic plotted can be changed to a relative measure such as the risk ratio, odds ratio, incidence rate ratio etc. The corresponding x-axis would then display the log transformed p-value based on the chosen statistical test. To adjust for issues of multiple testing Zink et al. suggest incorporation of a multiplicity adjustment such as the false discovery rate.

Compared to the standard practice of a frequency table, the volcano plot provides a way to immediately identify outlying events. In the absence of any disproportionality in harm between arms we would expect to see a symmetrical curve, with a random scatter of events. The strength of the volcano plot is its provision of an overarching and immediate representation of the AE profile that allows the user to draw attention to a particular signal of interest, but it does not allow a detailed review of all AEs. For example, we can see in Figure 1 fairly symmetrical scatter for most AEs, with only two AEs contributing to the asymmetry. These events have been selected to be labelled in the plot to draw attention to them. Labelling shows that the clear outlier is increased aspartate aminotransferase in the placebo arm (9 (12%) vs 7 (5%)). The greatest harm signal in the remdesivir arm is for rash (2 (3%) vs 11 (7%)).

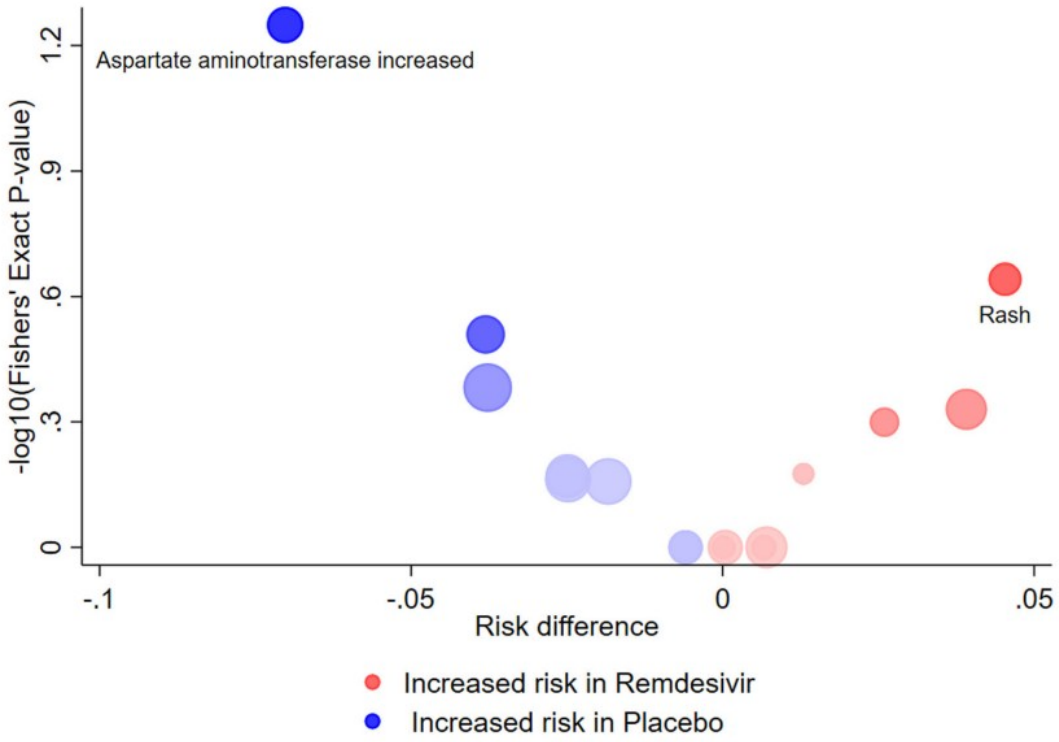


Figure 1: Example of a Volcano Plot for two treatment arms. Colour is used to indicate direction of treatment effect with red indicating greater risk in the Remdesivir group and blue indicating greater risk in the placebo group. Colour saturation corresponds to strength of p-value (level of evidence). The size of the bubble is proportional to the total number of events for both treatment groups combined.

A second valuable plot to display a summary across multiple outcomes is the Dot plot [Amit et.al. 2008]. Unlike the volcano plot, in which only one summary statistic can be displayed at one time, the Dot plot displays both the relative and absolute measure simultaneously. This is particularly important when the event rate is low. The same data in Figure 1 are shown in Figure 2. In this plot the relative risk (RR) with accompanying 95% CIs has been plotted on the central panel and AEs are ordered from the bottom to the top by increasing RR. We can see that the point estimates of the RRs are evenly distributed either side of the vertical line of no difference (RR = 1), and that there is great uncertainty in these estimates. The point estimate furthest away from one, communicates an elevated risk of increased aspartate aminotransferase in the placebo arm, and the upper 95% CI excludes the value of no difference. The corresponding absolute risks can be read from the left-hand side of the graph with the number of adverse events by arm included in the right-hand side of the plot. This detailed information on absolute values can be especially useful in situations such as when the event rate is low. In contrast to the volcano plot, the dot plot requires more assimilation but contains incorporates more information that is useful and we would expect to see in table.

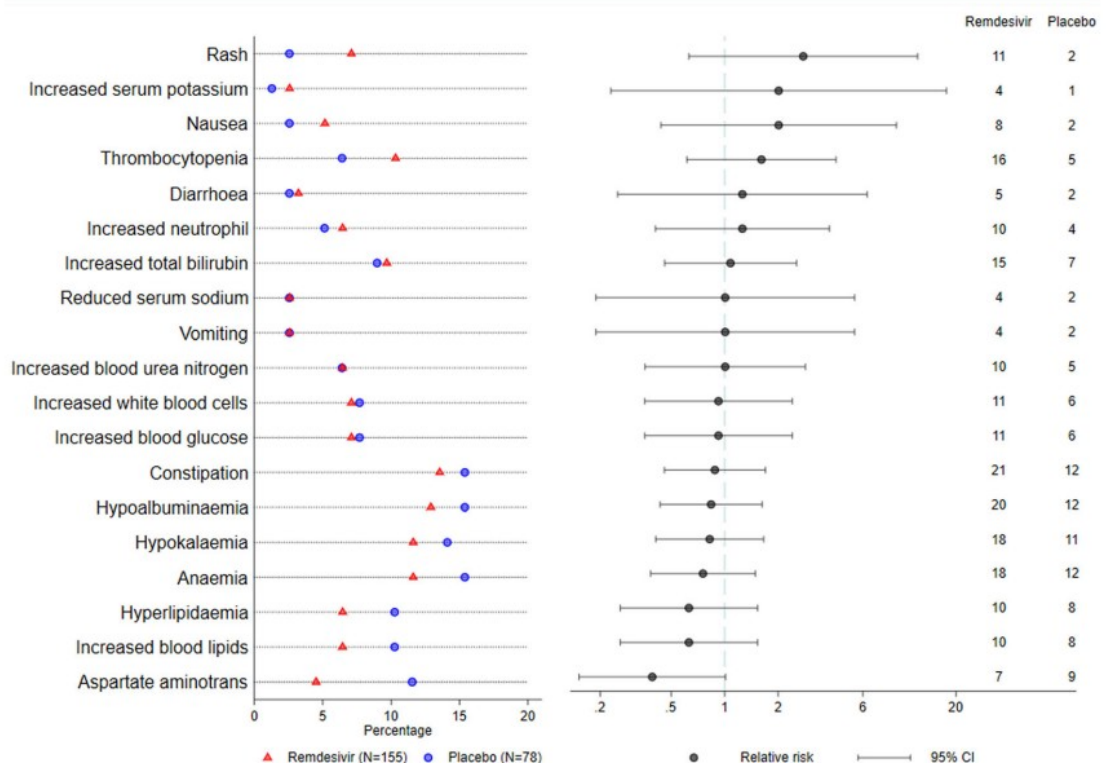


Figure 2. Example of a Dot Plot for two treatment arms. The left-hand side displays the absolute percentage value by treatment group. The right-hand side displays the relative risk and corresponding 95% CI on the log scale

While the dot and volcano plot are helpful to convey between arm imbalances, they do not provide a good summary measure of comparative overall burden. With some trials, such as trials of vaccines, we know the harm outcomes that we would like to compare in advance. When this is the case this enables us to use alternative approaches to the analysis. While rare events cannot be well detected within trials due to the relatively small sample size, we can reliably determine and compare reactogenicity events from vaccination. The standard approach to examining this burden is through a bar chart with stacks by severity. In order to interpret the information across multiple treatment arms in a stacked bar chart the reader has to visually compare several aspects (e.g. height of each section across rows and columns) and there is no direct between arm comparison.

In Figure 3 we present data from the COV-BOOST trial based on the principle of a radial graph. The figure compares different randomized vaccination arms for two different cohort of participants: participants who received two doses of Pfizer (BNT/BNT) prior to their 3rd randomized vaccine arm (right hand side); Participants who received two doses of AstraZeneca (ChAd/ChAd) prior to their 3rd randomized vaccine arm (left hand side). While information on the day that the event occurred is lost compared to a stacked bar chart, the maximum burden over 7 days is shown and the between arm comparison inference can be instantly observed. In Figure 3 it can be seen that the cohort who originally received AstraZeneca (ChAd/ChAd), for all booster vaccine arms, have few events and the rates are similar between arms as the coloured lines sit on top of one another. In contrast, participants who originally received two doses of Pfizer (BNT/BNT) then the third dose with AstraZeneca (ChAd) had higher reactogenicity events. Events of malaise, headache, feverish, fatigue and chill were higher.

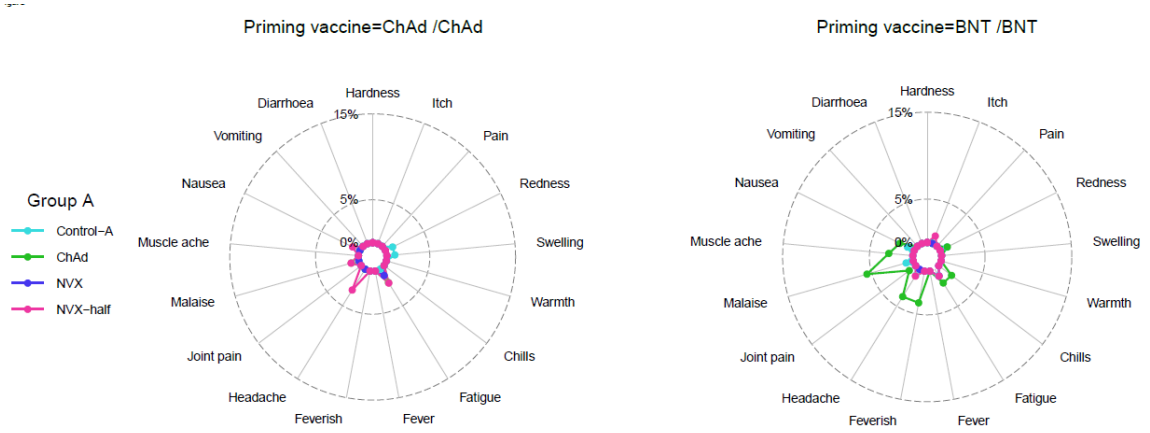


Figure 3: A fingerprint profile of reactogenicity events to facilitate between-arm comparison of the burden to participants. Each spoke is a pre-specified reactogenicity event, and the scale is the proportion of participants with a moderate or severe event. The area defined by each coloured circle is a summary measure of burden for each treatment arm.

Statistical software in Stata has developed to implement the volcano and dot plot [AEVOLCANO and AEDOT], the radial graph can be constructed in R using ggradar package

The volcano plot can be produced in Stata using the command aevolcano, which can be downloaded for use as follows, [AEVOLCANO]

```
ssc install aevolcano
```

The Dot plot can be produced using the command aedot [AEDOT]

```
ssc install aedot
```

Discussion

Assessing the benefit-harm balance is a critical part of a randomized controlled trial as this informs prescribing decisions and the design of future studies. Despite this, harm outcomes in trial publications are poorly presented. Poor presentation can obscure an informative assessment of the true harm profile. Graphical approaches offer a valuable way to summarize large amounts of emerging and pre-specified harm data from RCTs. However, the use of visualizations in journal articles to date has been limited. We have demonstrated the value of three approaches that are easy to implement with standard statistical software. We endorse the routine adoption of visualizations in trial publications as a powerful way to assess and communicate potential harms.

References

Amit O, Heiberger RM and Lane PW. Graphical approaches to the analysis of safety data from clinical trials. *Pharmaceutical Statistics* 2008; 7: 20-35.

Arksey H, O'Malley L. Scoping studies: towards a methodological framework. International Journal of Social Research Methodology. 2005;8(1):19-32.

Cornelius V, Cro S, Phillips R. Advantages of visualisations to evaluate and communicate adverse event information in randomised controlled trials. Trials. 2020; Vol: 21, ISSN: 1745-6215

Lineberry N, Berlin J, Mansi B, C et al. Recommendations to improve adverse event reporting in clinical trial publications: a joint pharmaceutical industry/journal editor perspective. BMJ. 2016; 355.

Munro APS, Janani L, Cornelius V, et al. Safety and immunogenicity of seven COVID-19 vaccines as a third dose (booster) following two doses of ChAdOx1 nCov-19 or BNT162b2 in the UK (COV-BOOST): a blinded, multicentre, randomised, controlled, phase 2 trial, The Lancet. 2021; Vol: 398,

Pitrou I, Boutron I, Ahman N, Ravaud P. Reporting of safety results in published reports of randomized controlled trials. Arch Intern Med 2009; 129 (19): 1756-61.

Phillips R, Hazell L, Suazet O, Cornelius V. Analysis and reporting of adverse events in randomised controlled trials: a review. BMJ Open. 2019.

Phillips R, Cornelius V, Sauzet O. Statistical methods for the analysis of adverse event data in randomised controlled trials: a scoping review and taxonomy, BMC Medical Research Methodology/ 2020; Vol: 20, ISSN: 1471-2288

AEVOLCANO. Phillips R, Cro S.: Stata module to produce volcano plot for adverse event data. https://ideas.repec.org/c/boc/bocode/s458736.html

AEDOT. Phillips R, Cro S.: Stata module to produce dot plot for adverse event data. https://ideas.repec.org/c/boc/bocode/s458735.html

Protocol: An overview of statistical methods developed to analyse adverse events in clinical trials: protocol for a methodological review: https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=97442

Zink RC, Wolfinger RD and Mann G. Summarizing the incidence of adverse events using volcano plots and time intervals. *Clinical Trials* 2013; 10: 398-406.

IMPACT DE LA VARIABILITÉ DE LA PRESSION ARTÉRIELLE SUR LE RISQUE D'AVC

Léonie Courcoul [1,a] & Antoine Barbieri [2,a] & Christophe Tzourio [3,a] & Hélène Jacqmin-Gadda [4,a]

[a] *Univ. Bordeaux, INSERM, BPH, U1219, F-33000 Bordeaux, France*
[1] *leonie.courcoul@u-bordeaux.fr*
[2] *antoine.barbieri@u-bordeaux.fr*
[3] *christophe.tzourio@u-bordeaux.fr*
[4] *helene.jacqmin-gadda@u-bordeaux.fr*

Résumé. Étant donné l'incidence des accidents vasculaires cérébraux (AVC) et leurs mauvais pronostics, leur prévention est un enjeu majeur de santé publique. Il est maintenant bien démontré qu'un niveau élevé de pression artérielle est un facteur de risque d'AVC, mais un nombre croissant d'études suggère que la variabilité de la pression artérielle pourrait également être un facteur de risque indépendant d'AVC. Cependant, ces études souffrent souvent de faiblesses méthodologiques importantes. L'objectif de ce travail était de développer un modèle conjoint à variance hétérogène pour les mesures répétées d'un marqueur longitudinal et le risque d'événements compétitifs afin d'étudier l'association entre la variabilité de la pression artérielle et le risque d'AVC en tenant compte du risque compétitif de décès. Ce modèle conjoint combine un modèle mixte incluant un effet aléatoire spécifique au sujet pour la variance résiduelle et un modèle à risque proportionnel cause-spécifique pour les risques compétitifs. Les risques peuvent dépendre simultanément de la variance résiduelle spécifique au sujet et de la valeur et de la pente courantes du marqueur. Le modèle a été estimé sur les données de l'essai clinique PROGRESS pour la prévention de la récurrence des accidents vasculaires cérébraux qui inclut 6105 sujets suivis pendant 5 ans avec 12 temps de mesure de la pression artérielle. Nous avons constaté que le risque de récidive d'AVC augmentait avec la valeur courante de la pression artérielle mais n'était pas associé à la variabilité intra-individuelle ou à la pente de la pression artérielle. Les capacités prédictives des modèles avec différentes structures de dépendance ont été comparées en utilisant la vraisemblance conditionnelle du temps d'événement schant le marqueur, le Brier Score et l'aire sous la courbe ROC en utilisant des estimateurs tenant compte de la censure à droite et des risques compétitifs.

Mots-clés. Modèle conjoint, variabilité intra-individuelle, pression artérielle, accident vasculaire cérébral.

Abstract. Given the incidence of stroke and and its poor prognosis, its prevention is a major public health issue. It is now well demonstrated that a high level of blood pressure is a risk factor for stroke but an increasing number of studies suggests that blood pressure variability may also be an independent risk factor for stroke. However, these

studies often suffer from significant methodological weaknesses. The objective of this work was to develop a joint model with heterogeneous variance for the repeated measures of a longitudinal marker and the risk of competing events to study the association between blood pressure variability and the risk of stroke. This joint model combines a mixed model including a subject-specific random effect for the residual variance and a proportional cause-specific hazard model for the competing risks. The risks of events may depend simultaneously on the subject-specific residual variance and the current value and slope of the marker. The model was estimated on the data of the PROGRESS clinical trial for the prevention of the recurrence of stroke that includes 6105 subjects followed over 5 years with 12 measurement times for blood pressure. We found that the risk of stroke recurrence increased with the current value of blood pressure but was not associated with intra-individual variability or slope of blood pressure. The predictive abilities of models with different dependence structures were compared using the conditional likelihood of the time-to-event given the marker, the Brier Score and the area under the ROC curve using estimators accounting for right-censoring and competing risks.

Keywords. Joint model, intra-individual variability, blood pressure, stroke.

1 Introduction

Chaque année, en France, près de 140000 personnes sont victimes d'un accident vasculaire cérébral (AVC) [1]. L'AVC représente la première cause d'handicap physique acquis de l'adulte, la deuxième cause de démence après la maladie d'Alzheimer, et la deuxième cause de mortalité (première chez les femmes). Par ailleurs, les soins médicaux et la prise en charge des handicaps suite à un AVC engendrent d'importantes dépenses sociétales, humaines et financières [2]. Enfin, les risques de récidive ne sont pas négligeables puisque près d'un sixième des patients font une récidive dans les 5 années suivant le premier AVC. Compte tenu de la fréquence des AVC et de leurs conséquences dramatiques au niveau individuel et sociétal, l'identification de leurs facteurs de risque est essentielle, afin de mettre en place des programmes de prévention lorsque ces facteurs sont modifiables. Dans ce contexte, l'hypertension artérielle est déjà connue pour être un facteur de risque majeur d'AVC [3]. Ce facteur est intéressant à étudier car sa prévalence est élevée et augmente fortement avec l'âge et il existe aujourd'hui des traitements hypotenseurs efficaces. Cependant, jusqu'à présent, la majorité des études portant sur l'hypertension comme facteur de risque vasculaire-cérébral s'est focalisée sur le rôle de la valeur de la pression artérielle alors que la variabilité de la pression artérielle augmente avec l'âge et certains travaux suggèrent que la variabilité serait un facteur de risque d'accident vasculaire cérébral [4,5] indépendant de la pression artérielle moyenne.

Ces études utilisent l'écart-type empirique ou le coefficient de variation de la pression artérielle comme variable explicative d'un modèle de Cox pour le risque de l'événement et souffrent de faiblesses méthodologiques importantes [6,7]. En effet, la première stratégie consiste à calculer l'écart-type empirique de la pression artérielle sur l'ensemble des mesures disponibles. Cela induit un conditionnement sur le futur et un risque de biais car des mesures postérieures au temps courant (et parfois au temps d'événement) sont utilisées pour prédire l'événement au temps courant. La deuxième stratégie consiste à calculer l'écart-type de la pression artérielle sur les mesures collectées sur une période initiale de l'étude en ne gardant que les individus qui n'ont pas présenté l'événement sur cette période afin de prédire le risque d'AVC au-delà de cette période. Cela peut donc induire un biais de sélection et une perte de puissance. Pour éviter ces biais, l'écart-type de la pression artérielle peut être considérée comme une variable dépendante du temps et calculé en utilisant seulement les mesures passées. Néanmoins cette approche néglige l'erreur de mesure sur l'écart-type qui est variable d'un individu à l'autre si le nombre de mesures diffère entre les individus et nécessite l'imputation de l'écart-type à tous les temps d'événement puisqu'il n'est pas mesuré en continu. Ces limites peuvent introduire des biais [8]. Enfin la pression artérielle, de même que son écart-type, sont des variables endogènes, or, le modèle de Cox n'est pas adapté à ce type de variables [9].

En 2019, Barrett et al [10] ont proposé un modèle conjoint combinant un modèle mixte incluant un effet aléatoire spécifique au sujet pour la variance résiduelle et un modèle à risque proportionnel pour le risque d'événement afin d'étudier l'association entre la variabilité de la pression artérielle et de l'AVC. Néanmoins, une des limites de ce modèle est que le risque ne dépend que des effets aléatoires et non de la valeur courante du marqueur.

L'objectif de ce travail était donc d'étendre ce modèle conjoint afin de permettre l'étude d'événements compétitifs, une structure de dépendance plus flexible entre l'événement et la trajectoire du marqueur, ainsi qu'une fonction de risque de base plus souple. Afin d'étudier rigoureusement l'effet de la variabilité de la pression artérielle sur le risque d'AVC, ce modèle est estimé sur les données de l'essai clinique PROGRESS [11] qui inclut 6105 patients suivis pendant 5 ans avec de fréquentes mesures de la pression artérielle.

2 Le modèle

Considérons un échantillon de N individus, et pour chaque individu $i (i = 1, ..., N)$ de cet échantillon le vecteur des n_i mesures répétées $Y_i = (Y_{i1}, ..., Y_{ij}, ..., Y_{in_i})$ où Y_{ij} est la valeur du marqueur de l'individu i mesuré au temps t_{ij}. Supposons qu'il existe deux événements compétitifs et une censure à droite. On note T^*_{i1} et T^*_{i2} les vrais temps pour les deux types d'événements du sujet i et C_i le temps de censure. Soit T_i le temps d'événement observé,

alors $T_i = \min(T_{i1}^*, T_{i2}^*, C_i)$. Soit $\delta_i \in \{0, 1, 2\}$ l'indicateur d'événement, avec $\delta_i = k$ si l'événement compétitif k est observé et $\delta_i = 0$ en cas de censure.

Le modèle conjoint à effets aléatoires partagés est défini par un sous-modèle linéaire mixte :

$$Y_{ij} = Y_i(t_{ij}) = \widetilde{Y}_i(t_{ij}) + \epsilon_{ij} = X_{ij}^\top \beta + Z_{ij}^\top b_i + \epsilon_{ij} \tag{1}$$

avec $X_{ij}^\top$ et $Z_{ij}^\top$ deux vecteurs de covariables au temps t_{ij}. Ils sont respectivement associés au vecteur des effets fixes β et au vecteur des effets aléatoires b_i. Le vecteur des effets aléatoires b_i suit une loi normale multivariée centrée et de matrice de variance-covariance Σ. Comme proposé par Barrett et al en 2019 [10], un effet aléatoire spécifique au sujet sur la variance de l'erreur de mesure est inclu. Ainsi, les erreurs de mesure ϵ_{ij} suivent une loi normale de variance σ_i^2 avec $\log(\sigma_i)$ qui suit une loi normale d'espérance μ_σ et de variance τ_σ^2.

On définit le sous-modèle de survie pour l'événement $k \in \{1, 2\}$ par :

$$\lambda_{ik}(t) = \lambda_{0k}(t) \exp\left(W_i^\top \gamma_k + \alpha_{1k}\tilde{y}_i(t) + \alpha_{2k}\tilde{y}_i'(t) + \alpha_{\sigma k}\sigma_i\right) \tag{2}$$

avec $\lambda_{0k}(t)$ la fonction de risque de base, W_i un vecteur de covariables avec les coefficients de régression correspondants γ, et α_{1k}, α_{2k} et $\alpha_{\sigma k}$ les coefficients de régression associés à l'effet de la valeur courante, de la pente courante et de la variabilité du marqueur sur le risque de l'événement k.

Le risque de base peut être de type exponentiel ou Weibull, ou, pour plus de flexibilité, il peut être défini sur une base de B-splines par :

$$\log(\lambda_{0k}(t)) = \exp\left(\sum_{q=1}^{Q} \eta_{qk} B_q(t, \nu_k)\right) \tag{3}$$

où $B_q(t, \nu_k)$ dénote la q-ème fonction de base de B-splines avec le vecteur de noeuds ν_k.

3 Méthode d'estimation

L'estimation des paramètres se fait sous le paradigme fréquentiste par maximisation de la vraisemblance. La troncature à gauche peut-être prise en compte. Pour cela on définit T_{0i} le temps d'entrée dans l'étude de l'individu i. On définit, pour $k \in 1, 2$:

$$\Lambda_k(t|b_i, \sigma_i) = \int_0^t \lambda_k(u|b_i, \sigma_i) du \tag{4}$$

La contribution de l'individu i à la vraisemblance marginale du modèle avec variance résiduelle hétérogène, s'écrit :

$$L_i(Y_i, T_i, \delta_i) = \frac{\int f(Y_i|b_i, \sigma_i) \exp(-\Lambda_1(T_i|b_i, \sigma_i) - \Lambda_2(T_i|b_i, \sigma_i)) \prod_{k=1}^{2} \lambda_k(T_i|b_i, \sigma_i)^{1_{\delta_i=k}} f(b_i) f(\sigma_i) db_i d\sigma_i}{\int \exp(-\Lambda_1(T_{0i|b_i,\sigma_i})) - \Lambda_2(T_{0i}|b_i, \sigma_i)) db_i d\sigma_i} \tag{5}$$

car b_i et σ_i sont supposés indépendants.

L'intégrale sur les effets aléatoires n'ayant pas de solution analytique nous pouvons l'estimer par la méthode de quasi-Monte-Carlo. L'intégrale pour les risques cumulés est calculée par Gauss-Kronrod. La maximisation de la fonction de vraisemblance se fait par l'algorithme Marquardt-Levenberg, variante robuste de l'algorithme de Newton-Raphson, grâce à la fonction "marLevAlg" du package éponyme [12].

4 Application

Le modèle a été appliqué à l'essai clinique PROGRESS afin d'étudier l'impact de la variabilité de la pression artérielle sur le risque d'AVC. L'essai clinique PROGRESS est un essai randomisé, contrôlé, en double aveugle, multicentrique et international, ayant pour objectif de déterminer l'effet d'un traitement hypotenseur, le Perindopril, associé ou non à l'Indapamide, pour prévenir la récidive des AVC. Pour cet essai, 6105 patients ont été randomisés en suivant un ratio 1:1 : 3054 ont été affectés au bras Placebo et 3051 ont été affectés au bras Traitement. Le protocole prévoyait cinq visites au cours de la première année de suivie puis deux visites annuelles pendant les 4.5 années suivantes. L'événement d'intérêt considéré est la première récidive d'un AVC mortel ou non mortel, hémorragique ou ischémique. Les décès dus à d'autres causes ont été considérés comme des événements concurrents.

L'estimation du modèle sur le bras Placebo montre que le risque de récidive d'AVC augmente avec la valeur courante de la pression artérielle mais il n'apparaît pas significativement associé à la variabilité de la pression artérielle.

Bibliographie

[1] GBD 2016 Stroke Collaborators (2019). Global, regional, and national burden of stroke, 1990-2016 : A systematic analysis for the global burden of disease study 2016. *Lancet Neurol*, 18, 439-458.

[2] De Pourville, G. (2016). Coût de la prise en charge des accidents vasculaires cérébraux en France, *Archives of Cardiovascular Diseases Supplements*, 8, 161-168.

[3] Lewington, S., Clarke R., Qizilbash, N., Peto, R., Collins, R.; Prospective Studies Collaboration. (2002). Age-specific relevance of usual blood pressure to vascular mortality : a meta-analysis of individual data for one million adults in 61 prospective studies. *The Lancet*, 360, 1903-1913.

[4] Pringle, E. et al, (2003). Systolic blood pressure variability as a risk factor for stroke and cardiovascular mortality in the elderly hypertensive population. *Journal of Hypertension*, 21, 2251-2257.

[5] Rothwell, PM., Howard, SC., Dolan, E., O'Brien, ET., Dobson, JE., Dahlöf, B., Sever, PS., Poulter, NR. (2010). Prognostic significance of visit-to-visit variability, maximum systolic blood pressure, and episodic hypertension. *The Lancet*, 375, 895-905.

[6] de Courson, H., Leffondre, K., Tzourio, C. (2018). Blood pressure variability and risk od cardiovascular event : is it appropriate to use the future for predicting the present ? *European heart journal*, 39, 4220-4220.

[7] Mehlum, MH., Liestøl, K., Kjeldsen, SE., Julius, S., Hua, TA., Rothwell, PM., Mancia, G., Parati, G, Weber, MA., Berge, E. (2018). Blood pressure variability and risk of cardiovascular events and death in patients with hypertension and different baseline risks. *European Hearth Journal*, 39, 2243-2251.

[8] Prentice, RL. (1982). Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika*, 69, 331-342.

[9] Kalbeisch, J. and Prentice, R. (2002). The Statistical Analysis of Failure Time Data, 2nd edition. *John Wiley Sons*.

[10] Barrett, JK., Huille, R., Parker, R., Yano, Y., Griswold, M. (2019). Estimating the association between blood pressure variability and cardiovascular disease : An application using the aric study. *Statistics in medicine*, 38, 1855-1868.

[11] PROGRESS Management Committee (1996). Blood pressure lowering for the secondary prevention of stroke: rationale and design for PROGRESS *Journal of Hypertension*, 14, S41-S46.

[12] Philipps, V., Heijblum, B., Prague, M., Commenges, D., Proust-Lima, C. (2021). Robust and efficient optimization using a Marquardt-Levenberg algorithm with R package marqlevalg. *The R Journal*, 13.

Utilisation des matrices emplois expositions pour les troubles musculo-squelettiques. Application aux douleurs sévères du genou.

Guillaume Deltreil [1] & Patrick Tardivel [2]
Alexis Descatha [3] & Mikael Escobar-Bach [1]
Piotr Graczyk [1]

[1] *LAREMA, Université d'Angers, France*
[2] *IMB, Université de Bourgogne, France*
[3] *Univ Angers Inserm, CHU Angers, Hofstra/ Northwell, France-USA*

Résumé. Les troubles musculo-squelettiques (TMS) ; appellation incluant les pathologies du dos, des membres supérieurs et des genoux, comptent parmi les premières causes de maladies professionnelles. Ces dernières sont liées à des facteurs de risques au travail tels que le port de charges lourdes. Durant cette présentation, nous adapterons un modèle de régression logistique *via* une sélection de modèle pénalisée afin d'étudier les effets du temps total individuel d'exposition et de la moyenne individuelle d'exposition sur la probabilité de développer une douleur sévère de genou.

Mots-clés. Troubles musculo-squelettiques, régression logistique, matrice emplois expositions.

Abstract. Musculoskeletal disorders (MSDs), including back pain, upper extremity and knee disorders, are the main cause of occupational disorders. MSDs are associated with many risk factors during working life, such as carrying heavy loads. During our presentation, we will tailor a logistic model by penalized model selection method in order to study the effects of the total individual exposure time and of the individual average exposure on the probability to contract severe knee pain.

Keywords. Musculoskeletal disorders, Logistic model, Job Exposure Matrix.

1 Introduction:

L'étude CONSTANCES est une large cohorte française de salariés affiliés à la Caisse Nationale d'Assurance Maladie. Notre échantillon pour cette étude se compose de 66 553 sujets issus de cette cohorte parmi des sujets recrutés entre 2012 et 2020 et âgés de 18 à 69 ans. Ces sujets sont en activité professionnelle et leurs calendriers professionnels étaient codés. Les douleurs sévères de genou ont été considérés comme indicateur de troubles musculo-squelettiques (douleur de genou d'intensité $> 5/10$ ou quotidiennes dans l'année d'inclusion). Les données ont été obtenues à partir de questionnaires et

d'entretiens médicaux dont les individus sont différenciés selon les critères suivants : le sexe (femmes : 46,2 %; hommes : 53,7 %), l'âge (âge moyen : 48,5 ans), l'existence connue d'une maladie inflammatoire des articulations (1,4 % des sujets), les activités de loisir (pratiquants : 43,1 %), le CES-D (questionnaire de dépression; 21,2 % des sujets) et l'indice de masse corporelle (IMC moyen : 25,07 kg/m^2). En plus de ces variables, la matrice emplois-expositions (JEM Constances) a pu être appliquée pour les expositions pertinentes sur la carrière professionnelle, ici le port de charges lourdes. Ainsi, nous avons pour chaque individu, le temps total individuel d'exposition et la moyenne individuelle d'exposition par emploi exercé et ce sur la totalité de sa carrière professionnelle. On parlera par la suite *d'intensité/fréquence* puisqu'il s'agit d'évaluer l'importance d'une contrainte biomécanique d'origine professionnelle.

2 Méthodologie

2.1 Sélection de modèle

Notre méthode est basée sur le modèle de régression logistique et un critère de minimisation pénalisé dans la sélection du modèle. Formellement, nous considérons un échantillon $(Y_i, X_{i,1}, \ldots, X_{i,p}, A_1, \ldots, A_6)_{i\in[\![1,n]\!]}$ *i.i.d.* à valeurs dans $\{0,1\}^{p+1} \times \mathbb{R}^6$ de taille $n \in \mathbb{N}^*$, tel que conditionnellement à $\mathbf{X}_i := (1, X_{i,1}, \ldots, X_{i,p}, A_{i,1}, \ldots, A_{i,6})$, la variable Y_i suit une loi de Bernoulli de paramètre

$$\pi_{\beta^*}(\mathbf{X}_i) := \frac{\exp(<\beta^*, \mathbf{X}_i>)}{1+\exp(<\beta^*, \mathbf{X}_i>)} \tag{1}$$

où $\beta^* \in \mathbb{R}^{p+7}$ est un vecteur inconnu de paramètres à estimer.
À l'exception des covariables d'ajustement $A_1, \ldots, A_6$ représentant le sexe, l'âge, l'IMC, les loisirs, l'arthrite et la dépression, la matrice de covariables $X := (X_{ij})_{1\leq i\leq n, 1\leq j\leq p}$ définit la matrice de planification du modèle et traduit l'appartenance de chaque individu à un groupe donné. Plus précisément, on suppose que $X_{ij} = 1$ si l'individu i appartient au groupe j et 0 sinon. À noter que les groupes sont disjoints avec $X_{ij}X_{ij'} = 0$ dès que $j \neq j'$. Ces groupes seront comparés au groupe témoin pour lequel $X_{ij} = 0$, $\forall j \in [\![1,p]\!]$. A priori, le nombre de groupes p n'est pas fixe et dépend du nombre de niveaux que l'on souhaite considérer pour l'exposition moyenne ainsi que pour la durée de la carrière. Notre sélection du nombre de groupes et de la matrice de planification X se base sur Kwemou et al. (2015), dont les garanties mathématiques reposent sur les inégalités oracles de Birgé et Massart (2001).
Plus précisément, soit $\mathcal{F} := \left\{X^{(1)}, \ldots, X^{(M)}\right\}$ une famille de matrices de planification donnée. Pour chaque $m \in \{1, \ldots, M\}$, $\hat{\beta}_m$ désigne l'estimateur du maximum de vraisemblance associé à la matrice de planification $X^{(m)}$ pour le modèle (1). Nous utilisons ensuite un critère du maximum de vraisemblance pénalisée pour choisir une matrice de

planification parmi $\mathcal{F}$. Ici, nous considérons le critère d'information d'Akaike (Kwemou et al. (2015)) où p_m, le nombre de colonnes de la matrice $X^{(m)}$, correspond au nombre de paramètres du modèle (1). On minimise alors le critère suivant

$$\widehat{m} := \arg \min_{m \in \{1,\dots,M\}} \left\{ -\gamma_n(\widehat{\beta}_m) + p_m \right\}.$$

où γ_n désigne la fonction de log-vraisemblance du modèle (1). La matrice sélectionnée est donnée par $X^{\widehat{m}}$ et l'estimateur de β^* est alors donné par $\beta_n := \widehat{\beta}_{\widehat{m}}$.

2.2 Construction de la matrice de planification

Dans le cadre de notre étude, la construction des groupes dépend de la valeur de durée totale des carrières et des expositions moyennes des individus. Ces dernières se décrivent dans la cohorte *via* les variables $T_i := (T_{i1}, \dots, T_{ik_i})$ et $N_i := (N_{i1}, \dots, N_{ik_i})$, respectivement pour les temps d'occupation et les niveaux d'exposition, de chacun des k_i emplois occupés par l'individu i. Sachant que les nombres d'emplois k_i peuvent différer, nous résumons la trajectoire professionnelle des sujets *via* les transformations

$$\phi_1(T_i) := \sum_{j=1}^{k_i} T_{ij}, \quad \text{et} \quad \phi_2(T_i, N_i) := \frac{\sum_{j=1}^{k_i} N_{ij} g(T_{ij})}{\sum_{j=1}^{k_i} g(T_{ij})}$$

où $g(x) = 1 - \exp(-x)$. La construction de la matrice de planification est alors basée sur un *semi-coding* des variables $\{\phi_1(T_i)\}_{1 \le i \le n}$ et $\{\phi_2(T_i, N_i)\}_{1 \le i \le n}$. Plus précisément, l'adhésion à un groupe est déterminée par l'appartenance des valeurs $\phi_1(T_i)$ et $\phi_2(T_i, N_i)$ à différents intervalles préalablement donnés. Pour cette étude, nous considérons des classes d'intervalles

$$\mathcal{G}_1 := \{[q_j, q_{j+1}[,\ j = 1, \dots, q\} \quad \text{et} \quad \mathcal{G}_2 := \{\,[j-1, j[\,\}_{1 \le j \le 4} \cup \{4\}$$

où q_j définit des quantiles empiriques de l'échantillon $\{\phi_1(T_i)\}_{1 \le i \le n}$. L'ensemble des groupes est défini comme la classe $\mathcal{G} := \{G_j,\ j = 1, \dots, p\} = \mathcal{G}_1 \times \mathcal{G}_2$ où $p = 5q$. Un individu est alors associé au groupe $j \in [\![1, p]\!]$, et $X_{i,j} = 1$, si et seulement si

$$(\phi_1(T_i), \phi_2(T_i, N_i)) \in G_j.$$

2.3 Tests

La seconde partie de la présentation sera dédiée à la comparaison entre la probabilité de développer un TMS pour un groupe d'exposition par rapport à la probabilité de développer un TMS pour le groupe témoin. Cette comparaison peut s'effectuer *via* l'*odds-ratio* car avoir des probabilités différentes est équivalent à avoir un *odds-ratio* différent de 1. En

effet, l'*odds-ratio* correspond au facteur multiplicateur entre la probabilité d'être malade pour un groupe d'exposition par rapport à la probabilité de développer une maladie pour le groupe témoin. Dans le cadre du modèle de régression logistique, l'*odds-ratio* a une formule simple : $OR_j = \exp(\beta_j^*)$ et il est facile de tester si l'*odds-ratio* est significativement différent de 1.
Nous souhaitons déterminer les groupes d'exposition pour lesquels la probabilité d'être malade est significativement différente de la probabilité d'être malade pour le groupe témoin. Spécifiquement, pour chaque groupe j, on formule le problème de test suivant :

$$H_0^j : OR_j = 1 \text{ vs } H_1^j : OR_j \neq 1.$$

3 Application aux données

Le procédé de sélection évoqué en section 2.3 nous a fourni une matrice de planification ayant 20 colonnes dont 14 colonnes correspondent à des groupes. Les groupes sont construits à partir de la durée totale de la carrière : courte (entre 0 et 13,125 ans), intermédiaire (entre 13,125 ans et 25,75 ans) et longue (entre 25,75 ans et 53 ans) couplée avec l'intensité/fréquence : $[0, 1[, [1, 2[, [2, 3[, [3, 4[$ et 4. Le groupe de référence est composé des sujets ayant une intensité/fréquence comprise entre 0 et 1 et une durée totale de la carrière comprise entre 0 et 13,125 ans.

Variable	OR	IC 95 %	p-valeur
Durée [0, 13.125] x Intensité/Fréquence [1,2[	1.56	[1.41 , 1.73]	$< 10^{-4}$
Durée [0, 13.125] x Intensité/Fréquence [2,3[	1.70	[1.49 , 1.94]	$< 10^{-4}$
Durée [0, 13.125] x Intensité/Fréquence [3,4[	1.58	[1.23 , 2.03]	$< 10^{-4}$
Durée [0, 13.125] x Intensité/Fréquence 4	2.29	[1.79 , 2.93]	$< 10^{-4}$
Durée [13.125, 25.75] x Intensité/Fréquence [0,1[*	0.94	[0.86 , 1.03]	0.19
Durée [13.125, 25.75] x Intensité/Fréquence [1,2[	1.29	[1.16 , 1.44]	$< 10^{-4}$
Durée [13.125, 25.75] x Intensité/Fréquence [2,3[	1.79	[1.56 , 2.07]	$< 10^{-4}$
Durée [13.125, 25.75] x Intensité/Fréquence [3,4[	1.82	[1.39 , 2.39]	$< 10^{-4}$
Durée [13.125, 25.75] x Intensité/Fréquence 4	2.10	[1.47 , 2.99]	$< 10^{-4}$
Durée [25.75, 53.0] x Intensité/Fréquence [0,1[*	0.96	[0.87 , 1.05]	0.35
Durée [25.75, 53.0] x Intensité/Fréquence [1,2[	1.31	[1.18 , 1.46]	$< 10^{-4}$
Durée [25.75, 53.0] x Intensité/Fréquence [2,3[	1.58	[1.39 , 1.79]	$< 10^{-4}$
Durée [25.75, 53.0] x Intensité/Fréquence [3,4[	1.80	[1.48 , 2.18]	$< 10^{-4}$
Durée [25.75, 53.0] x Intensité/Fréquence 4	2.73	[2.12 , 3.52]	$< 10^{-4}$

Table 1: Ce tableau concerne l'étude des douleurs sévères au genou selon la levée de poids. Pour chacun des groupes ce tableau fournit une p-valeur pour le problème de test: l'odds-ratio est égal à 1 contre l'odds-ratio est différent de 1 ainsi qu'un intervalle de confiance pour l'odds-ratio. La notation "*" indique les p-valeurs plus grandes que $0.05/14 \approx 0.0036$ (correction de Bonferroni).

Indépendamment de la durée de la carrière, les individus exerçant un emploi dont l'intensité/fréquence est faible (comprise entre 0 et 1) ont un odds-ratio qui n'est pas significativement différent de 1 (la p-valeur dépasse $0,05$ ou de façon équivalente 1 est dans l'intervalle de confiance de l'odds-ratio). À l'inverse, indépendamment de la durée de la carrière, les individus exerçant un emploi dont l'intensité/fréquence est forte (comprise entre 1 et 4) ont un odds-ratio qui est significativement différent de 1. En conclusion, un individu ayant effectué des tâches exigeantes physiquement (forte intensité/fréquence) a, d'après notre modélisation, une probabilité de développer une douleur de genou supérieure à un individu n'ayant pas ou peu effectué de tâches exigeantes physiquement (faible intensité/fréquence). Notre modélisation montre également que la durée de la carrière a un faible impact sur la probabilité de développer ce type de douleur.
D'autres études (en cours) portant sur d'autres expositions professionnelles et d'autres TMS sont nécessaires pour confirmer les résultats. En parallèle, des comparaisons de modèles avec des pondérations différentes sont en cours.

Bibliographie

Kwemou, M. and Taupin, M.L. and Tocquet, A.S. (2015). Model selection in logistic regression, *arXiv*.
Birgé, L. and Massart, P. (2001). Gaussian model selection, Journal of the European Mathematical Society, 3,3, pp. 203-268.

Optimal Individualized Treatment Regime of Plasma Transfusion for Severe Trauma

Pan Zhao [1] Nicolas Gatulle [2] Julie Josse [3] Antoine Chambaz [4]

[1] *Inria. pan.zhao@inria.fr*
[2] *Sorbonne University, GRC 29, AP-HP, DMU DREAM, Pitié-Salpêtrière Hospital. nicolas.gatulle@aphp.fr*
[3] *Inria. julie.josse@inria.fr*
[4] *Université de Paris. antoine.chambaz@u-paris.fr*

Résumé. Un régime de traitement individualisé (RTI) est une règle de décision qui attribue un traitement, parmi les options disponibles, à un patient en fonction de ses caractéristiques. Le choc hémorragique est la deuxième cause de décès chez les patients victimes de traumatismes sévères, et est considéré comme la première cause de mortalité évitable dans les hôpitaux. Nous effectuons une analyse des données de TraumaBase®, et développons une nouvelle méthode basée sur le super-learning pour apprendre l'ITR optimal de la transfusion de plasma pour les patients victimes de traumatismes sévères. En outre, nous nous attaquons au problème des données manquantes.

Mots-clés. Régime de Traitement Individualisé, Super Learner, Estimation Ciblée du Maximum de Vraisemblance, Données Manquantes, Traumatisme Sévère, Transfusion.

Abstract. An individualized treatment regime (ITR) is a decision rule that assigns a treatment, among the available options, to a patient based on the patient's characteristics. Hemorrhagic shock is the second leading cause of death in severe trauma patients, and considered as the leading cause of preventable mortality in hospitals. We conduct an analysis of TraumaBase® data, and develop a new super-learning based method to learn the optimal ITR of plasma transfusion for severe trauma patients. In addition, we tackle the issue of missing data.

Keywords. Individualized Treatment Regime, Super Learner, Targeted Maximum Likelihood Estimation, Missing Data, Severe Trauma, Transfusion

1 Introduction

Traumas take the lives of 4.4 million people around the world each year and constitute nearly 8% of all deaths [1]; in France, traumas cause approximately 40,000 deaths each year [2]. Hemorrhagic shock, present in about one-third of patients, is the second leading

[1] https://www.who.int/news-room/fact-sheets/detail/injuries-and-violence
[2] https://www.santepubliquefrance.fr/maladies-et-traumatismes/traumatismes

cause (about 15%) of death in severe trauma patients, behind head injuries (about 63.5%), and is considered as the leading cause of preventable mortality in hospitals (Pfeifer et al. (2009)).

Intensive care of severe trauma patients with hemorrhagic shock has evolved considerably over the last few decades (Holcomb et al. (2015)); among current practices including damage-control resuscitation, in this work we focus on the transfusion of fresh frozen plasma (FFP).

An individualized treatment regime is a decision rule that assigns a treatment, among the available options, to a patient based on the patient's characteristics. We conduct a retrospective study based on the TraumaBase® dataset, and develop a new super-learning based method to learn the optimal FFP quantity to administer to trauma patients, in order to maximize the in-hospital survival of patients after severe trauma.

2 Basic setup

2.1 Data

TraumaBase® is a national registry aiming to describe the characteristics, management and outcomes of severe trauma patients admitted to 20 level 1 trauma centers in France. All trauma patients admitted to one of the participating centers between 2010 and 2020 are included in our study.

The dataset contains extensive information on demographic characteristics, injury pattern (penetrating or blunt injury), medication before trauma, pre-hospital management, in-hospital management including trauma bay and initial surgery, ICU stay and patients' outcomes.

2.2 Notations and Assumptions

In our data structure, each patient contributes a vector of data consisting of baseline variables $W \in \mathbf{R}^k$, treatment $A \in \mathcal{A} = \{0, 1, 2, \ldots\}$ and outcome Y. We denote this vector of data for each patient i by $O_i = (W_i, A_i, Y_i)$, and we assume that O_i is i.i.d. from an unknown probability distribution P of a random variable O.

We aim to evaluate an individualized treatment regime $d : \mathcal{W} \to \mathcal{P}(\mathcal{A})$ mapping any $w \in \mathcal{W}$ to a subset $d(W)$ of $\mathcal{A}$, and define the parameter of interest (also known as the value of regime d) as

$$\Psi_d(P) = E_P\{E_P(Y|A \in d(W), W)\}. \tag{1}$$

To provide a causal interpretation of this parameter, we use the potential outcomes framework from Rubin (1974). Under the three common assumptions (consistency, unconfoundedness and positivity) in causal inference literature (Hernán & Robins (2020)), $\Psi_d(P)$ is the expected potential outcome of the population had all patients received the

treatment $A \in d(W)$ according to the regime d, where A is drawn from the conditional law of A given W and that $A \in d(W)$.

3 Methods

We outline here our main methods for optimal ITR estimation. Consider the outcome regression model:

$$Q_0(a, w) = P(Y = 1|A = a, W = w) \tag{2}$$

and the propensity score model:

$$g_0(a, w) = P(A = a|W = w) \tag{3}$$

To estimate the value 1 of a regime, both super learning algorithm and parametric marginal structural models are implemented, and then a targeted maximum likelihood estimation (TMLE) step with cross-validated risk evaluation is carried out in order to draw valid statistical inference for optimal ITR.

3.1 Super Learner with Missing Data

Super learning is a general loss based learning algorithm proposed and analyzed theoretically in Van der Laan et al. (2007). The powerful super learner first uses cross-validation to estimate the performance of multiple data-adaptive machine learning algorithms, and then creates an ensemble of these algorithms using test data, which has been shown in Polley & Van der Laan (2010) to be asymptotically equivalent to the best possible prediction algorithm.

However, the super learning algorithm only works for complete data. In the presence of missing data, two popular approaches are: (a) use only the set of complete cases; and (b) use imputation algorithms to complete the data for further analysis.

Instead of separating the imputation and prediction procedures, we propose a new super learning algorithm that can handle missing data, which evaluates the performance of the combination of both procedures; also a R package `missSuperLearner`[3] has been prepared.

3.2 Marginal Structural Models

Besides highly adaptive nonparametric methods, we are also interested in the performance of parsimonious parametric models, such as the popular marginal structural models (MSMs) proposed by Robins et al. (2000). Experts with domain knowledge can propose

[3]The R package will be public in GitHub soon.

different parametric models; and note that the choice of a working model defines a target parameter, but does not represent a model assumption.

Usually a working model of MSM is defined as:

$$m_\beta = \text{expit}(\beta_0 + \beta_1^T a + \beta_2^T w) \tag{4}$$

where $\beta \in \mathcal{B}$ fully parameterize the model for $P(Y = 1|A, W)$. As shown in Rosenblum & Van der Laan (2010), a targeted maximum likelihood estimator can be used to improve the MSM estimation.

3.3 Targeted Maximum Likelihood Estimation for ITR Value

Targeted maximum likelihood estimation (Van der Laan & Rose (2011)) incorporates the super learner and MSM as initial estimators Q_n and g_n for model 2 and 3, and then carries out a bias reduction step for the target parameter 1; define the clever covariate:

$$H(a, w) = \frac{I\{a = d(w)\}}{g_n(a, w)} w$$

and solve the parametric submodel $P(\epsilon)$:

$$P(\epsilon)(Y = 1|d(W), W) = \text{expit}\{\epsilon H + \text{logit}(Q_n(d(w), w))\}$$

then the estimated ϵ_n is used in the correction step, which has been shown to enjoy the so-called double robustness property and statistical efficiency.

3.4 Variable Importance

After learning the optimal ITR, we are also interested in understanding which variables play an important role in the ITR; that is, an variable importance measure for building the ITR is required. Williamson et al. (2021) introduced the analysis of variance (ANOVA) variable importance measure and discuss how it facilitates the use of machine learning techniques to flexibly estimate the variable importance.

For any vector v and any given set $s \subseteq \{1, 2, \ldots, k\}$, v_{-s} denotes the vector of all components of v with index not in s. We define the ANOVA-based variable importance measure for regime d:

$$\begin{aligned}\Psi_{d,s}(P) &= \frac{E_P\{(E_P(d(W)|W) - E_P(d(W)|W_{-s}))^2\}}{var_P\{d(w)\}} \\ &= \frac{E_P\{(d(W) - E_P(d(W)|W_{-s}))^2\}}{var_P\{d(w)\}}\end{aligned}$$

4 Results and Discussion

We already conducted a preliminary analysis based on the vanilla `SuperLearner` package (and simply median imputation for the TraumaBase® dataset). We are currently preparing the analysis using our newly developed `missSuperLearner` package, and also computing the variable importance measure.

Bibliographie

Hernán MA, Robins JM (2020). Causal Inference: What If. Boca Raton: Chapman & Hall/CRC.

Holcomb, J. B., Tilley, B. C., Baraniuk, S., Fox, E. E., Wade, C. E., Podbielski, J. M., ... & PROPPR Study Group. (2015). Transfusion of plasma, platelets, and red blood cells in a 1: 1: 1 vs a 1: 1: 2 ratio and mortality in patients with severe trauma: the PROPPR randomized clinical trial. *Journal of the American Medical Association*, 313(5), 471-482.

Pfeifer, R., Tarkin, I. S., Rocos, B., & Pape, H. C. (2009). Patterns of mortality and causes of death in polytrauma patients—has anything changed?. *Injury*, 40(9), 907-911.

Polley, E. C., & Van Der Laan, M. J. (2010). Super learner in prediction.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5), 688.

Robins, J. M., Hernan, M. A., & Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5), 550-560.

Rosenblum, M., & Van der Laan, M. J. (2010). Targeted maximum likelihood estimation of the parameter of a marginal structural model. *The international journal of biostatistics*, 6(2).

Van der Laan, M. J., Polley, E. C., & Hubbard, A. E. (2007). Super learner. *Statistical applications in genetics and molecular biology*, 6(1).

Van der Laan, M. J., & Rose, S. (2011). Targeted learning: causal inference for observational and experimental data (Vol. 4). New York: Springer.

Williamson, B. D., Gilbert, P. B., Carone, M., & Simon, N. (2021). Nonparametric variable importance assessment using machine learning techniques. *Biometrics*, 77(1), 9-22.

Données Fonctionnelles

Estimation adaptative dans le modèle linéaire fonctionnel à sortie fonctionnelle

Gaëlle Chagny [1] & Anouar Meynaoui [2] & Angelina Roche [3]

[1] *LMRS, UMR CNRS 6085, Université de Rouen Normandie,*
gaelle.chagny@univ-rouen.fr
[2] *LIG, UMR 5217, Université de Grenoble Alpes,*
anouar.meynaoui@univ-grenoble-alpes.fr
[3] *CEREMADE, UMR CNRS 7534, Université Paris Dauphine,*
roche@ceremade.dauphine.fr

Résumé.
On considère un modèle de régression linéaire fonctionnel à sortie fonctionnelle : les variables explicative et réponse sont des variables aléatoires "fonctionnelles", à valeurs dans un espace de Hilbert. On s'intéresse à la question de l'estimation non-paramétrique de l'opérateur intégral reliant ces deux variables, à partir d'un échantillon. Une collection d'estimateurs par projection sur la base de l'ACP empirique de la covariable est construite. On obtient une décomposition biais-variance pour un risque quadratique moyen de prédiction. Une procédure de sélection de modèle (minimisation de contraste pénalisé) permet ensuite un choix automatique du meilleur estimateur dans la collection. Celui-ci satisfait une inégalité de type oracle, et atteint des vitesses de convergences minimax sur des espaces de régularité de type ellipsoïde : la borne supérieure du risque de prédiction correspond à la borne inférieure, que nous calculons aussi. Ces résultats théoriques sont illustrés par des applications à des jeux de données simulées et réelles.

Mots-clés. Estimation non-paramétrique adaptative, Analyse de données fonctionnelles, sélection de modèles.

Abstract. We consider a functional linear regression model, where both the covariate and the response variable are functional random variables. We address the problem of nonparametric estimation of the conditional expectation operator in this model. A collection of projection estimators onto the functional PCA basis of the covariate is built. A nonasymptotic bias-variance decomposition for the Mean Square Prediction Error is proved. The automatic trade-off is realized thanks to a model selection device which selects the best projection dimensions: the penalized contrast estimator satisfies an oracle-type inequality and is thus optimal in an adaptive point of view. These upper-bounds allow us to derive convergence rates over ellipsoidal smoothness spaces. These rates are optimal in the minimax sense: they match with a lower bound, which is also proved. Finally, we conduct a numerical study, over simulated data and over two real-data sets.

Keywords. Adaptive nonparametric estimation, functional data analysis, model selection.

1 Introduction

Nous considérons dans ce travail un modèle de régression fonctionnel linéaire à sortie fonctionnelle, de la forme :

$$Y = SX + \varepsilon, \tag{1}$$

où X, Y et ε sont des variables aléatoires fonctionnelles, c'est-à-dire à valeurs dans un espace de dimension infinie, que l'on suppose être l'espace $\mathbb{H} = L^2([0,1])$, des fonctions de carré intégrable sur l'intervalle $[0,1]$, muni du produit scalaire usuel $\langle \cdot, \cdot \rangle$ défini $\langle f, g \rangle = \int_0^1 f(t)g(t)dt$, $f, g \in \mathbb{H}$ et de la norme associée $\| \cdot \| = \sqrt{\langle \cdot, \cdot \rangle}$. La covariable X et le bruit (non-observé) ε sont supposés indépendants et centrés. Le lien, à estimer, entre la variable d'intérêt Y et la variable explicative X est linéaire : S est un élément de $\mathcal{L}(\mathbb{H})$, l'espace des opérateurs linéaires continus sur $\mathbb{H}$. Cet opérateur, dit *opérateur de pente* est supposé être intégral, de noyau $\mathcal{S} \in L^2([0,1]^2)$, au sens où

$$\begin{array}{rcl} S : \mathbb{H} & \longrightarrow & \mathbb{H} \\ f & \longmapsto & Sf \,:\, t \in [0,1] \mapsto Sf(t) = \displaystyle\int_0^1 \mathcal{S}(s,t)f(s)ds. \end{array}$$

Notre objectif est de proposer une méthode d'estimation non-paramétrique pour S (ou son noyau $\mathcal{S}$) à partir de l'observation d'un échantillon $(X_i, Y_i)_{i \in \{1,\dots,n\}}$ ($n \in \mathbb{N} \backslash \{0\}$) distribué selon la loi de (X, Y), méthode optimale d'un point de vue minimax et adaptatif.

L'analyse de données fonctionnelles (voir par exemple Ferraty et Romain, 2011) est un domaine de recherche en plein essor ces dernières années, et l'intérêt pour des modèles de régression impliquant de telles données s'est beaucoup développé. Le cas du modèle linéaire fonctionnel à sortie scalaire (où seule la variable explicative est fonctionnelle, et la variable à prédire est réelle) est désormais bien compris : Cardot et Johannes (2010) ont calculé les vitesses minimax d'estimation dans ce modèle, et une procédure adaptative a par exemple été proposée par Brunel *et al.* (2016) (voir également d'autres références dans ce dernier article). À l'inverse, le cas du modèle linéaire fonctionnel à sortie fonctionnelle (1) qui nous intéresse a été moins étudié.

La majorité de la littérature consacrée au sujet se concentre sur des estimateurs par projection pour l'opérateur S, dans la base de l'Analyse en Composantes Principales (ACP) de la covariable X, soit la base de fonctions propres $(\varphi_j)_{j \geq 1}$ de l'opérateur de covariance Γ de X, défini par

$$\Gamma : \quad f \in \mathbb{H} \quad \longmapsto \quad \mathbb{E}[X \otimes X(f)] = \mathbb{E}[\langle X, f \rangle X],$$

où $\otimes$ désigne le produit tensoriel de $\mathbb{H}$, défini pour deux éléments a et b de $\mathbb{H}$ par $b \otimes a : f \mapsto \langle a, f \rangle b$.

Crambes et Mas (2013) sont parmi les premiers à proposer une étude théorique de tels estimateurs : ils proposent des vitesses de convergence, une borne inférieure pour le risque quadratique de prédiction, et la convergence en loi. Mais les résultats restent

non adaptatifs, au sens où la dimension optimale de l'espace de projection dépend des indices de régularité, a priori inconnus, de l'opérateur cible S et aussi de la covariable X, via la vitesse de décroissance des valeurs propres $(\lambda_j)_{j\geq 1}$ de l'opérateur de covariance Γ associées aux fonctions propres $(\varphi_j)_{j\geq 1}$ (et rangées par ordre décroissant $\lambda_1 \geq \lambda_2 \ldots$). Plus récemment, Imaizumi et Kato (2018) ont repris l'étude de l'estimateur de Crambes et Mas (2013), qu'ils nomment "simple truncature ", et proposent une seconde méthode, dite "double truncature". Leurs résultats sont asymptotiques, non-adaptatifs, et valides uniquement dans le cas d'une décroissance polynomiale des valeurs propres $(\lambda_j)_j$.

Nous proposons ici une stratégie d'inférence par minimisation de contraste pénalisé. Nous démontrons des résultats théoriques (borne supérieure de risque pour les estimateurs par projection, borne inférieure pour le risque minimax, et inégalité de type oracle) prouvant l'optimalité de la sélection de modèle, au sens minimax et adaptatif.

2 Collection d'estimateurs par projection

Nous définissons un contraste de type régression qui va correspondre à un équivalent empirique du risque que nous considérerons (voir ci-dessous): soit

$$\gamma_n : \quad T \in \mathcal{L}_2(\mathbb{H}) \quad \longmapsto \quad \frac{1}{n}\sum_{i=1}^{n} \|Y_i - T(X_i)\|^2 ,$$

où $\mathcal{L}_2(\mathbb{H})$ désigne le sous-espace de $\mathcal{L}(\mathbb{H})$ constitué des opérateurs de Hilbert-Schmidt ($S \in \mathcal{L}_2(\mathbb{H})$, puisque son noyau $\mathcal{S} \in L^2([0,1]^2)$). Soit également les sous-espaces vectoriels de dimension finie ou modèles

$$V_{m_1,m_2} = \mathrm{Span}\{\widehat{\varphi}_k \otimes \widehat{\varphi}_j,\ 1 \leq j \leq m_1, 1 \leq k \leq m_2\}, \ \ m_1, m_2 \in \mathbb{N}\backslash\{0\}.$$

Nous introduisons

$$\widehat{S}_{m_1,m_2} \in \arg\min_{T \in V_{m_1,m_2}} \gamma_n(T), \ \ m_1, m_2 \in \mathbb{N}\backslash\{0\}, \tag{2}$$

ainsi que les notations suivantes, pour l'opérateur de covariance empirique, et l'opérateur de covariance empirique croisée :

$$\Gamma_n : f \mapsto \frac{1}{n}\sum_{i=1}^{n} \langle f, X_i\rangle X_i \quad \text{ et } \quad \Delta_n : f \mapsto \frac{1}{n}\sum_{i=1}^{n} \langle f, X_i\rangle Y_i.$$

Soient aussi $(\widehat{\lambda}_j, \widehat{\varphi}_j)_{j\geq 1}$ les éléments propres de Γ_n. Le pseudo-inverse de Γ_n est défini par

$$\Gamma_{n,m_1}^{\dagger} = \sum_{j=1}^{m_1} \frac{1}{\widehat{\lambda}_j} \widehat{\varphi}_j \otimes \widehat{\varphi}_j, \quad m_1 \in \mathbb{N}\backslash\{0\}.$$

Nous prouvons alors le résultat suivant.

Proposition 1. *L'estimateur $\widehat{S}_{m_1,m_2}$ introduit en (2) est défini de manière unique, et se réécrit*

$$\widehat{S}_{m_1,m_2} = \widehat{\Pi}_{m_2}\Delta_n\Gamma^{\dagger}_{n,m_1}, \quad m_1, m_2 \in \mathbb{N}\backslash\{0\}, \tag{3}$$

où $\widehat{\Pi}_{m_2}$ est l'opérateur de projection orthogonale sur Vect$\{\widehat{\varphi}_1, \ldots, \widehat{\varphi}_{m_2}\}$; ou encore,

$$\widehat{S}_{m_1,m_2} = \sum_{\substack{1\leq j\leq m_1\\ 1\leq k\leq m_2}} \frac{1}{\widehat{\lambda}_j}\langle\Delta_n(\widehat{\varphi}_j), \widehat{\varphi}_k\rangle\widehat{\varphi}_k \otimes \widehat{\varphi}_j = \Delta_n\Gamma^{\dagger}_{n,m_1} + \sum_{\substack{1\leq j\leq m_1\\ 1\leq k\leq m_2\\ j\neq k}} \frac{1}{\widehat{\lambda}_j}\langle\Delta_n\widehat{\varphi}_j, \widehat{\varphi}_k\rangle\widehat{\varphi}_j \otimes \widehat{\varphi}_k.$$

Par comparaison avec les estimateurs de Crambes et Mas (2013), définis par $\widehat{S}^{CM}_{m_1} = \Delta_n\Gamma^{\dagger}_{n,m_1}$, nous avons une régularisation supplémentaire, naturelle au premier abord puisque estimer S revient à estimer son noyau $\mathcal{S}$ qui est une fonction de deux variables. L'expression (3) provient également heuristiquement de l'équation $\Delta = S\Gamma$, où Δ est l'opérateur de covariance croisée $\Delta : f \in L^2([0,1]) \mapsto \mathbb{E}[Y \otimes X(f)]$. Cette équation, déduite de (1), prouve aussi que l'estimation de S est un problème inverse mal posé.

3 Bornes supérieures et inférieures de risque

Nous considérons le risque quadratique de prédiction pour un estimateur $\widehat{S}_n$, défini par

$$\mathrm{MSPE}(\widehat{S}_n) = \mathbb{E}\left\|\widehat{S}_n(X_{n+1}) - S(X_{n+1})\right\|^2,$$

où X_{n+1} est une nouvelle observation de X, indépendante de $(X_i, \varepsilon_i), i = 1, \ldots, n$. On peut montrer que, si $\|\cdot\|_{HS}$ désigne la norme de Hilbert-Schmidt, $\mathrm{MSPE}(\widehat{S}_n) = \mathbb{E}\|(\widehat{S}_n - S)\Gamma^{1/2}\|^2_{HS}$.

Ceci explique pourquoi dans la suite notre hypothèse de régularité porte sur $S\Gamma^{1/2}$ et pas uniquement sur la cible S : on suppose

$$S\Gamma^{1/2} \in \mathcal{W}^R_{\alpha,\beta} = \left\{T \in \mathcal{L}_2(\mathbb{H}), \sum_{j=1}^{+\infty}\sum_{r=1}^{+\infty}\eta_\alpha(j)\psi_\beta(r)\langle T(\varphi_j), \varphi_r\rangle^2 \leq R^2\right\},$$

avec $\eta_\gamma(j) \asymp j^\gamma$ ou $\eta_\gamma(j) \asymp \exp(j^\gamma)$, (de même pour ψ_β).

Sous des hypothèses concernant la covariable X (des hypothèses de régularité : condition de séparabilité sur les valeurs propres $(\lambda_j)_{j\geq 1}$ et condition sur leur vitesse de décroissance - au maximum polynomiale; des hypothèses sur les moments), on démontre le résultat suivant.

Théorème 1. *Le risque de prédiction maximal est majoré par la décomposition biais-variante suivante, pour une constante $c > 0$ et pour $\sigma_\varepsilon^2 = \mathbb{E}\|\varepsilon\|^2$,*

$$\inf_{\substack{m_1,m_2\in\mathbb{N}\setminus\{0\}\\ m_1\leq n/\ln^2(n)}} \sup_{S\Gamma^{1/2}\in\mathcal{W}^R_{\alpha,\beta}} MSPE(\widehat{S}_{m_1,m_2}) \leq \inf_{\substack{m_1\in\mathbb{N}\setminus\{0\}\\ m_1\leq n/\ln^2(n)}} \left\{\sigma_\varepsilon^2\frac{m_1}{n} + \frac{3}{\eta_\alpha(m_1)}\right\} + \frac{c}{n}.$$

De manière surprenante, le paramètre m_2 n'apparait pas dans la borne supérieure, phénomène constaté également par Imaizumi et Kato (2018). Le terme de variance n'en dépend en fait pas du tout (il est en m_1/n), et l'on montre que le biais tend vers 0 quand m_2 tend vers ∞. Le contrôle du biais nécessite le recours à des outils issus de la théorie de la perturbation (Kato, 2013). On démontre que la borne du Théorème 1 est optimale au sens minimax : elle coïncide avec la borne inférieure suivante.

Théorème 2. *Soient $\alpha > 0$, $\beta > 0$ et $R > 0$. Si le bruit ε est un processus gaussien, alors, il existe une constante $C > 0$ telle que*

$$\inf_{\widehat{S}_n} \sup_{S\Gamma^{1/2}\in\mathcal{W}^R_{\alpha,\beta}} MSPE(\widehat{S}_n) \geq C \inf_{m_1\in\mathbb{N}\setminus\{0\}} \left\{\sigma_\varepsilon^2\frac{m_1}{n} + \frac{3}{\eta_\alpha(m_1)}\right\},$$

où l'infimum est pris sur l'ensemble des estimateurs $\widehat{S}_n$ calculés à partir de l'échantillon $\{(X_i, Y_i), i = 1, \ldots, n\}$. En particulier,

1. *si $\eta_\alpha(j) \asymp j^\alpha$ (cas polynomial),*

$$\inf_{\widehat{S}_n} \sup_{S\Gamma^{1/2}\in\mathcal{W}^R_{\alpha,\beta}} MSPE(\widehat{S}_n) \geq Cn^{-\frac{\alpha}{\alpha+1}},$$

2. *si $\eta_\alpha(j) \asymp \exp(j^\alpha)$ (cas exponentiel),*

$$\inf_{\widehat{S}_n} \sup_{S\Gamma^{1/2}\in\mathcal{W}^R_{\alpha,\beta}} MSPE(\widehat{S}_n) \geq C\frac{(\ln(n))^{1/\alpha}}{n}.$$

Ces vitesses sont cohérentes avec les résultats obtenus dans le cas du modèle linéaire fonctionnel à sortie scalaire (Brunel *et al.*, 2016), et sont atteintes par les estimateurs (2) dans le cas où les dimensions de projections sont bien choisies (par exemple $m_1 = n^{1/(1+\alpha)}$ et $m_2 = \infty$ pour le cas polynomial).

4 Sélection de modèle

La question suivante est celle de la sélection automatique des dimensions de projection pour que l'estimateur correspondant atteigne les vitesses ci-dessus, sans être définis à partir de l'indice de régularité α de l'opérateur inconnu $S\Gamma^{1/2}$.

Au vu des résultats précédents, on choisit $m_2 = \infty$. Ceci nous conduit à considérer la collection d'estimateurs $(\widehat{S}_{m_1,\infty})_{m_1}$, avec $\widehat{S}_{m_1,\infty} = \Delta_n \Gamma^{\dagger}_{n,m_1}$ (voir Proposition 1). On choisit m_1 dans une collection $\mathcal{M}_n = \{1, \ldots, N_n\}$ avec $N_n \leq \lfloor n/\ln^2(n) \rfloor$, par minimisation de contraste pénalisé (Barron *et al.* , 1999) :

$$\widehat{m}_1 = \arg \min_{m_1 \in \mathcal{M}_n} \left(\gamma_n(\widehat{S}_{m_1,\infty}) + \text{pen}(m_1) \right), \tag{4}$$

où pen est la fonction de pénalité définie par pen : $m_1 \mapsto \kappa \sigma_\varepsilon^2 m_1/n$ pour une constante κ à calibrer en pratique.

Le résultat final, sous une hypothèse sur les moments du bruit ε, est une inégalité de type oracle pour la norme empirique, définie pour tout opérateur T par $\|T\|_n^2 = n^{-1} \sum_{i=1}^n \|T(X_i)\|^2$ (un résultat similaire peut être obtenu pour le risque de prédiction au prix de calculs techniques supplémentaires, voir par exemple Brunel *et al.* 2016).

Théorème 3. *Pour tout $\zeta > 0$,*

$$\mathbb{E} \left\| S - \widehat{S}_{\widehat{m}_1,\infty} \right\|_n^2 \leq (1+\zeta) \inf_{m_1 \in \mathcal{M}_n} \left\{ \mathbb{E} \left\| S - \widehat{\Pi}^{op}_{m_1,\infty} S \right\|_n^2 + c(\zeta) pen(m_1) \right\} + \frac{C'}{n},$$

pour $c(\zeta) = (2+\zeta)/(1+\zeta)$, et $C' > 0$, avec $\widehat{\Pi}^{op}_{m_1,\infty}$ la projection orthognonale sur l'adhérence de $V_{m_1,\infty} = Vect\{\widehat{\varphi}_k \otimes \widehat{\varphi}_j,\ 1 \leq j \leq m_1,\ m_2 \geq 1\}$.

Des détails et commentaires concernant la procédure, les démonstrations ainsi que des applications sur des données simulées et réelles, peuvent être trouvés dans le preprint de Chagny *et al.* (2022).

Bibliographie

Barron, A., Birgé, L., et Massart, P. (1999). Risk bounds for model selection via penalization *Probability theory and related fields*, 113(3), pp. 301413.

Brunel, É., Mas, A., et Roche, A. (2016) Non-asymptotic adaptive prediction in functional linear models. *Journal of Multivariate Analysis*, 143, pp. 208232.

Cardot, H. and Johannes, J. (2010). Thresholding projection estimators in functional linear models. *Journal of Multivariate Analysis*, 101(2), pp. 395408.

Crambes, C. et Mas, A. (2013) Asymptotics of prediction in functional linear regression with functional outputs *Bernoulli*, 19(5B), pp. 26272651.

Chagny G., Meynaoui A., et Roche A. (2022) Adaptive nonparametric estimation in the functional linear model with functional output, *Preprint, hal-03579232*.

Ferraty, F. et Romain, Y. (2011) *The Oxford handbook of functional data analysis*, Oxford University Press, Oxford.

Imaizumi, M. et Kato, K. (2018) PCA-based estimation for functional linear regression with functional responses. Journal of multivariate analysis, 163, pp. 1536.

Kato, T. (2013) *Perturbation theory for linear operators*, volume 132. Springer Science & Business Media.

Procédure de classification plug-in pour des données fonctionnelles

Eddy Ella Mintsa [1]

[1]Laboratoire LAMA UMR 8050 – Université Gustave Eiffel – France

Abstract

Les récents progrès de la technologie moderne ont généré des données étiquetées regardées comme des réalisations d'une fonction aléatoire. Ce travail porte sur un problème de classification multiclasses pour des données fonctionnelles modélisées par une équation différentielle stochastique. Peu de travaux étudient le cas où les données fonctionnelles sont modélisées par des processus de diffusion, c'est pourquoi, la construction de procédures de classification adaptées à ce type de modèle est un enjeu majeur. Nous nous concentrons sur des processus de diffusion homogènes en temps avec des coefficients de dérive et de diffusion inconnus. L'objectif est de proposer une procédure de classification de type plug-in implémentable basée sur une estimation non-paramétrique à partir d'un échantillon d'apprentissage, des fonctions de dérive et du coefficient de diffusion par la minimisation d'un contraste des moindres carrés sur une base de fonctions B-splines. Nous établissons ensuite la consistance du classifieur obtenu.

Mots-clés. Classification supervisée, processus de diffusion, estimation non-paramétrique, classifieur plug-in

Fusion tardive en analyse de données fonctionnelles élastique

Julien Ah-Pine [1,2,3] & Noé Lebreton [1]

[1] *Univ Lyon, Univ Lyon 2, ERIC UR3083; julien.ah-pine@univ-lyon2.fr*
[2] *Université Clermont Auvergne, LMBP-CMRS, UMR 6620*
[3] *Université Clermont Auvergne, CERDI-CNRS-IRD, UMR 6587*

Résumé. Dans cette communication nous comparons la fusion précoce et la fusion tardive en classification supervisée de données fonctionnelles lorsqu'une séparation des variations d'amplitude et de phase est pertinente. Le cadre méthodologique utilisé est l'analyse de données fonctionnelles élastique qui passe par un alignement des fonctions basé sur la distance de Fisher-Rao. Ensuite, afin de considérer conjointement les variations d'amplitude et de phase, la fusion précoce concatène les fonctions alignées et les fonctions de déformation temporelle dans une fonction composite avant réduction de dimension et apprentissage du modèle de classification. Cette méthode est proposée dans la littérature récente. *A contrario*, nous examinons une fusion tardive dans laquelle, deux réductions de dimension et apprentissages de classifieurs sont appliquées indépendamment sur les deux types de fonctions d'amplitude et de phase, et ce sont les estimations des prédictions des deux modèles qui sont fusionnées par un opérateur d'agrégation.

Mots-clés. Analyse de données fonctionnelles, variabilité d'amplitude et de phase, métrique élastique, fusion de données.

Abstract. In this paper, we compare early and late fusion schemes in classification tasks of functional data when amplitude and phase separation is relevant. We rely on the elastic functional data analysis framework where functions alignment is carried out using the Fisher-Rao distance. Then, in order to jointly consider amplitude and phase variabilities, the early fusion concatenates in a composite function, both the aligned functions and the time warpings, before applying dimension reduction and learning a classifier. This approach is proposed in the recent literature. In contrast, we examine a late fusion method where two dimension reductions and classifiers are independently estimated using the two types of functions and it is the estimated predictions that are combined by means of an aggregation operator.

Keywords. Functional data analysis, amplitude and phase variabilities, elastic metric, data fusion.

1 Introduction

Nous nous intéressons à l'analyse de données fonctionnelles (*FunctionalData Analysis* - FDA par la suite-) définie sur $[0, 1]$ dans les cas où il y a un intérêt à séparer les variations

d'amplitude et de phase comme, par exemple, lorsque les observations présentent des problèmes d'alignement. Dans ce contexte, l'utilisation de la distance de Fisher-Rao, dite métrique élastique (voir par exemple [3]), permet d'aligner les fonctions de façon appropriée et d'isoler en deux vues distinctes, des fonctions mettant en perspective des disparités au niveau des amplitudes (variation sur l'axe vertical des ordonnées) d'une part; et des fonctions de re-paramétrisation encodant des variabilités de phase dans le temps (variation sur l'axe horizontal) d'autre part.

Dans ce cadre, nous nous intéressons au problème de la fusion de ces deux sources d'informations complémentaires pour la tâche de classification supervisée. Nous abordons ce sujet en opposant d'un côté une fusion précoce qui concatène les deux types de fonctions encodant respectivement l'amplitude et la phase; et d'un autre côté, une fusion tardive qui raisonne au niveau des estimations données par deux modèles appris de façon indépendante sur l'un et l'autre des deux types de fonctions.

Nous illustrons sur deux jeux de données réels, les avantages potentiels de la fusion tardive que nous proposons vis-à-vis de la fusion précoce déjà proposée dans la littérature.

2 Analyse de données fonctionnelles élastique

Dans ce qui suit, f est une fonction de $\mathcal{F}$, l'ensemble des fonctions absolument continues définies sur $[0,1]$; γ est une fonction de Γ, l'ensemble des difféomorphismes de $[0,1]$ avec $\gamma(0) = 0$ et $\gamma(1) = 1$. La fonction $f \circ \gamma$ correspond à une re-paramétrisation de f. Supposons, sans perte de généralité, que le support $[0,1]$ est un axe temporel, alors $f \circ \gamma$ s'interprète comme un recalage de f par la déformation temporelle γ. Nous notons par q une fonction de $\mathbb{L}^2$, l'espace de Hilbert des fonctions définies sur $[0,1]$. Les notations précédentes sont génériques et lorsque nous travaillons avec un échantillons de n fonctions, nous ajoutons un indice inférieur i (parcourant la suite $1, \ldots, n$) à chacune de ces notations, le cas échéant.

2.1 Alignement des fonctions basé sur la transformation SRVF

Le problème d'alignement en FDA se pose lorsque, par exemple, il existe un décalage dans les observations des différentes fonctions, ce qui produit des imprécisions sur les mesures de distance entre celles-ci. Pour tenir compte de ce problème, les auteurs de [3] proposent d'utiliser la distance de Fisher-Rao, notée d_{FR}, entre fonctions. Cette métrique Riemannienne est adaptée au problème traité en raison de son invariance par rapport aux re-paramétrisations de Γ: $d_{FR}(f_1, f_2) = d_{FR}(f_1 \circ \gamma, f_2 \circ \gamma)$.

Cette métrique étant Riemannienne, elle est difficile à évaluer dans le cas général. Or, il existe une isométrie transformant les fonctions f de $\mathcal{F}$ en des fonctions q de $\mathbb{L}^2$ avec $d_{FR}(f_1, f_2) = \|q_1 - q_2\|_{\mathbb{L}^2}$. Cette propriété permet ainsi de réduire le coût computationnel des calculs. Cette transformation est donnée par $q(t) = \text{sign}(\dot{f}(t))\sqrt{|\dot{f}(t)|}$ où $\dot{f}$ est la

dérivée de f. q est appelée la *Square Root Velocity Function (SRVF)* de f [3] et nous avons par ailleurs la relation suivante: $f(t) = f(0) + \int_0^t q(s)|q(s)|ds$.

Nous renvoyons le lecteur à [3] pour le détail des procédures mettant en oeuvre, en théorie et en pratique, l'alignement dans le cadre précédent. Nous présentons plus dans le détail ce qui nous intéresse pour la suite et qui concerne les fonctions issues de ces procédures.

Si nous avons un échantillon de n fonctions de $\mathcal{F}$, $\{f_i\}_{i=1,\ldots,n}$, ce qui nous intéresse sont d'une part, les fonctions alignées dénotées $\{\tilde{f}_i\}_i$ et leurs SRVFs associées $\{\tilde{q}_i\}_i$ et d'autre part, les fonctions de re-paramétrisation ayant permis le recalage des fonctions et qui sont dénotées $\{\gamma_i\}_i$. En effet, $\{\tilde{f}_i\}_i$ et $\{\tilde{q}_i\}_i$ permettent d'isoler l'étude de la variabilité d'amplitude de celle de la phase qui est permise par l'analyse de l'ensemble $\{\gamma_i\}_i$.

2.2 Analyse séparée de l'amplitude et de la phase

Dans [4], un cadre méthodologique basé sur l'analyse en composantes principales fonctionnelle (*Functional Principal Component Analysis* -FPCA par la suite-) des SRVFs dans $\mathbb{L}_2$ est proposé, afin d'analyser séparément l'amplitude et la phase.

Dans le premier cas, il s'agit d'étudier les fonctions SRVF alignées $\{\tilde{q}_i\}_i$, augmentées des valeurs $\{f_i(0)\}_i$. L'ajout de ces dernières valeurs fait écho à la relation liant les fonctions et leurs SRVFs rappelée précédemment. Pour analyser l'amplitude, [4] propose de réaliser dans $\mathbb{L}_2 \times \mathbb{R}$ une FPCA sur les fonctions $\{[\tilde{q}_i \quad f_i(0)]\}_i$.

Concernant la phase, l'approche prend comme éléments de départ les fonctions $\{\gamma_i\}_i$ ayant conduit au recalage des $\{f_i\}_i$. La distance de Fisher-Rao est utilisée à nouveau (Γ n'est d'ailleurs pas un espace linéaire) et on calcule ainsi les SRVF des $\{\gamma_i\}_i$ que l'on dénote $\{\psi_i\}_i$. Dans ce cas particulier, nous avons la propriété suivante: $\psi_i = \sqrt{\dot{\gamma}_i}$. Ceci implique que dans $\mathbb{L}_2$, les SRVF $\{\psi_i\}_i$ appartiennent à l'hypersphère de rayon unité. L'espace n'étant pas linéaire, il n'est pas possible d'appliquer une FPCA directement. Pour contourner le problème, [4] propose de projeter (il s'agit d'une bijection) les $\{\psi_i\}_i$ sur l'espace tangent à la moyenne de Karcher des $\{\psi_i\}_i$. Ces *shooting vectors* dénotés $\{v_i\}_i$ sont dans un espace linéaire et il est alors possible d'appliquer la FPCA.

3 Fusion des vues amplitude et phase dans la classification supervisée de données fonctionnelles

La plupart des travaux qui étudient des données fonctionnelles (*Functional Data* -FD par la suite-) présentant des variations d'amplitude et de phase, alignent les fonctions puis analysent les fonctions alignées sans tenir compte des déformations temporelles. Or, la variabilité de phase peut apporter une information discriminante dans l'analyse. Dans ce contexte, [2] combine les fonctions alignées aux fonctions de re-paramétrisation en amont d'une FPCA. En termes de fusion d'information, l'approche peut être qualifiée de fusion

précoce (*early fusion*). Nous explorons *a contrario*, une alternative dite de fusion tardive (*late fusion*). Nous étudions ce thème dans le cadre de la classification supervisée de FD. Dans ce cas, une fonction f est associée à un label y, élément d'un espace fini discret $\mathcal{Y}$.

3.1 Fusion précoce

Dans la fusion précoce, on concatène les représentations des différentes vues en une représentation augmentée et composite avant d'appliquer une quelconque analyse. Dans notre contexte, [2] propose la fonction composite g^C définie sur $[0, 2]$ de la façon suivante:

$$g^C(t) = \begin{cases} \tilde{f}(t), & t \in [0, 1[\\ Cv(t-1), & t \in [1, 2] \end{cases}$$

où le paramètre $C > 0$ permet de gérer le déséquilibre d'échelle entre $\tilde{f}$ et v.

Nous nous intéressons aux travaux de [5] qui utilisent la fusion précoce précédente dans la tâche de classification supervisée. Contrairement à [2], [5] préconise d'utiliser $\tilde{q}$ plutôt que $\tilde{f}$ dans la définition de g^C car ces premières sont garanties d'être dans $\mathbb{L}_2$. Nous retenons cette proposition dans ce qui suit et dans nos expériences. Une fois la FPCA appliquée aux fonctions $\{g_i^C\}_i$, [5] utilise la régression logistique sur composantes principales comme modèle de classification (voir également [1]).

3.2 Fusion tardive

Dans l'approche que nous promouvons, la combinaison des informations d'amplitude et de phase se fait tardivement. La démarche consiste à apprendre deux modèles de prédiction, l'un centré sur l'amplitude et l'autre basé sur la phase et la classification finale s'effectue en combinant les estimations de prédiction des deux classifieurs. L'avantage ici est que chaque modèle se focalise sur un seul type d'information ce qui permet de différencier, si besoin, les traitements de chacune des deux vues. De plus, en fusionnant *a posteriori* les prédictions, cela permet d'avoir du recul sur laquelle des deux vues est la plus pertinente et si le mélange permet par ailleurs de meilleures performances. Dans notre contexte, en pratique, nous réutilisons les résultats des FPCA introduites dans la sous-section 2.2 et qui traitent les fonctions $\{\{[\tilde{q}_i \quad f_i(0)]\}_i\}$ indépendamment des fonctions $\{\gamma_i\}_i$. Nous obtenons deux dimensions de réduction distinctes, l'une pour l'amplitude, l'autre pour la phase. Ces deux espaces peuvent par ailleurs être de dimensions différentes. Ensuite, nous apprenons deux régressions logistiques sur composantes principales dans ces deux espaces de description.

Supposons que le modèle d'alignement et les deux régressions logistiques sur composantes principales aient été estimés à partir des $\{f_i\}$. Étant donné $f \in \mathcal{F}$, voici successivement les étapes pour obtenir sa prédiction: (1) calcul de la SRVF q, alignement et obtention de $\tilde{q}$ et γ; (2) projection de $[\tilde{q} \quad f(0)]$ dans l'espace réduit de l'amplitude et γ dans celui de la phase; (3) calcul des probabilités d'appartenance à chaque classe y pour

chacun des deux modèles de régression logistique; (4) fusion des probabilités de prédiction des deux classifieurs.

Nous détaillons l'étape (4) qui est au coeur de ce travail. Notons $P_V(y|f)$ la probabilité conditionnelle d'observer y, donnée par le classifieur basé sur l'amplitude et $P_H(y|f)$ celle donnée par le classifieur basé sur la phase. L'opérateur de fusion tardive que nous proposons est la moyenne pondérée:

$$P(y|x) = \alpha P_V(y|x) + (1-\alpha)P_H(y|x)$$

où $\alpha \in [0,1]$ est un paramètre permettant de donner des poids différents aux deux vues.

La règle de classification finale est alors $\arg\max_{y\in\mathcal{Y}} P(y|f)$. Afin de proposer une approche simple pour fixer α, nous utilisons également une fusion guidée par les performances respectives de chaque vue. Notons par acc_V et acc_H les taux de reconnaissance obtenus, sur un jeu de données de validation par exemple, par l'amplitude et la phase. Nous proposons d'estimer α et $\beta = 1-\alpha$ par $\text{acc}_V/(\text{acc}_V + \text{acc}_H)$ et $\text{acc}_H/(\text{acc}_V + \text{acc}_H)$.

4 Illustration sur deux jeux de données réels

Nous avons appliqué les méthodes de fusion précoce et tardive décrites précédemment sur deux jeux de données réels connus de la communauté: *Growth* et *Tecator*. Le premier concerne des courbes de croissance et il faut discriminer celles correspondant à des filles et celles correspondant à des garçons. Le deuxième représente des courbes de spectrométrie de viande et il faut indiquer si le taux de graisse est faible ou élevé.

Nous représentons dans la Figure 1, les différentes fonctions initiales, les fonctions de déformation temporelle et les fonctions alignées.

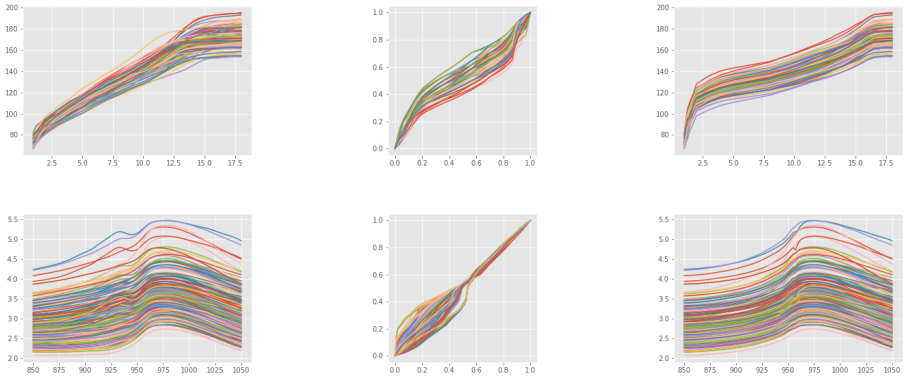


Figure 1: Illustrations des jeux de données: *Growth* 1ère ligne, *Tecator* 2ème ligne. De gauche à droite: $\{f_i\}$, $\{\gamma_i\}$ et $\{\tilde{f}_i\}$

Dans la Figure 2, nous montrons les résultats obtenus pour la fusion précoce avec un paramètre $C = [8.11, 0.06]$, la fusion tardive avec $\alpha = \{0, 0.1, \ldots, 1\}$ et la fusion tardive

guidée par les performances des deux vues. Ces résultats illustrent l'intérêt de la fusion tardive dans la mesure où elle permet d'obtenir des taux de reconnaissance meilleurs que la fusion précoce étant donné que la courbe des points (fusion tardive avec différentes valeurs de α) peut dépasser la ligne de *baseline* (fusion précoce). Les performances reportées pour les données *Tecator* sont particulièrement convaincantes. La fusion tardive guidée par les performances (point jaune) ne donne pas nécessairement les meilleurs résultats et la question de l'estimation du paramètre de mélange α reste donc ouverte.

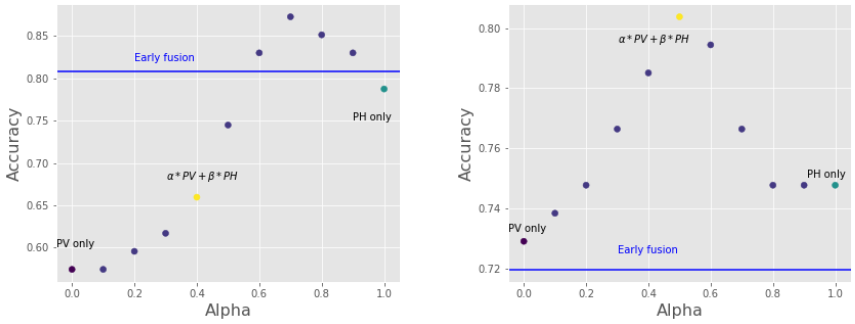


Figure 2: Fusion précoce *vs* fusion tardive: *Growth* à gauche, *Tecator* à droite

References

[1] Ana M Aguilera, Manuel Escabias, and Mariano J Valderrama. Using principal components for estimating logistic regression with high-dimensional multicollinear data. *Computational Statistics & Data Analysis*, 50(8):1905–1924, 2006.

[2] Sungwon Lee and Sungkyu Jung. Combined analysis of amplitude and phase variations in functional data. *arXiv preprint arXiv:1603.01775*, 2016.

[3] Anuj Srivastava, Wei Wu, Sebastian Kurtek, Eric Klassen, and James Stephen Marron. Registration of functional data using fisher-rao metric. *arXiv preprint arXiv:1103.3817*, 2011.

[4] J Derek Tucker. *Functional component analysis and regression using elastic methods*. PhD thesis, The Florida State University, 2014.

[5] J Derek Tucker, John R Lewis, and Anuj Srivastava. Elastic functional principal component regression. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 12(2):101–115, 2019.

Prédiction de la Qualité d'Expérience dans les Réseaux Mobiles : Cas de la VoIP

Jean Steve Tamo Tchomgui[1] & Julien Jacques[1] & Stéphane Chretien[1] & Guillaume Fraysse[2] & Vincent Barriac[2]

[1] *Univ Lyon, Univ Lyon 2, ERIC, Lyon.*
{*jean-steve.tamo-tchomgui, julien.jacques, stephane.chretien*}*@univ-lyon2.fr*
[2] *Orange Innovation, France.* {*guillaume.fraysse, vincent.barriac*}*@orange.com*

Résumé: Ce travail propose une étude comparative entre les techniques récentes de l'analyse des données fonctionnelles et celles des signatures pour la résolution d'un problème de prédiction de la Qualité d'Expérience (QoE), une mesure qui reflète la perception de l'utilisateur final d'un service de télécommunications. La QoE ainsi que les facteurs pouvant l'influencer étant mesurés à haute fréquence, le problème que nous considérons trouve donc sa formulation naturelle dans le contexte de la régression linéaire fonctionnelle, où la variable à prédire et les variables explicatives sont toutes fonctionnelles. Notre contribution principale est celle de montrer que l'utilisation des signatures permet de résumer de façon efficace l'information contenue dans les variables explicatives et de produire des prédictions souvent meilleures que celles obtenues par des méthodes classiques.
Mots-clés. Données fonctionnelles, Modèles de régression fonctionnels, Signatures.

Abstract: This work presents a comparative study of the functional data analysis approach and the approach based on the theory of signatures for a prediction problem in a Quality of Experience (QoE) monitoring projet. The problem we consider is a functional linear regression problem where both the dependent variables and the independent variables are functional variables. Our experiments demonstrate the efficiency of the signature approach for prediction in a difficult functional regression problem.
Keywords. Functional data, Functional regression models, Signatures.

Introduction

La plupart des problèmes d'apprentissage rencontrés aujourd'hui n'ont plus de données collectées sous forme classique, c'est-à-dire une sortie $Y \in \mathbb{R}$ décrite par un nombre fini de prédicteurs $X \in \mathbb{R}^p$. On observe plutôt de multiples enregistrements pour chaque prédicteur au cours du temps, autrement dit une fonction du temps de la forme $X : T \longrightarrow \mathbb{R}^p$. L'approche naturelle pour mener une inférence sur ce type de données est d'étendre le modèle linéaire classique à ce cadre plus général : c'est l'Analyse des Données

Fonctionnelles (ADF) et particulièrement la régression linéaire fonctionnelle. Comme le soulignent les auteurs de [1], qui constitue une excellente introduction au domaine, l'analyse fonctionnelle s'applique aux données dont la structure peut être représentée par une ou plusieurs fonctions et repose sur l'hypothèse selon laquelle les données à traiter possèdent une structure sous-jacente plus ou moins apparente dont l'identification et la prise en compte permettent d'étendre efficacement les techniques de l'analyse classique.

En pratique, on n'observe pas directement une fonction mais les valeurs de cette dernière à différents instants. Les données collectées se présentent donc sous forme vectorielle et nécessitent un premier traitement dans l'analyse fonctionnelle, il consiste à reconstruire la nature fonctionnelle des données. Une approche récente a été introduite en analyse fonctionnelle, basée sur la notion de signatures [2, 7] qui sont des séries (infinies) d'intégrales itératives permettant de résumer l'information contenue dans des données fonctionnelles.

Dans cet article, la Section 1 est consacrée à présenter brièvement le cadre d'analyse des données fonctionnelles, la Section 2 celle des signatures. La Section 3 décrit l'inférence menée et présente l'interprétation des résultats obtenus.

1 Inférence Fonctionnelle

Nous allons nous extraire dans ce travail du cadre d'analyse classique. Les observations ne seront plus des réalisations indépendantes et identiquement distribuées (i.i.d.) d'un processus aléatoire réel (ou vectoriel plus généralement), mais plutôt d'une variable aléatoire à valeur dans un espace de fonctions : c'est la notion de fonction aléatoire. Faire de l'ADF sur des observations collectées sous forme de séries pour chaque prédicteur revient d'abord à accepter l'hypothèse fondamentale d'existence du processus aléatoire sous-jacent et ensuite considérer chaque série observée comme les valeurs en un nombre fini de pas de temps d'une réalisation de ce dernier. Explicitement, pour une fonction aléatoire Y (également valable pour les variables explicatives X), nous disposons en réalité de n réalisations $\mathrm{Y_i}$ sous la forme :

$$\mathrm{Y_i} = \Big\{(\mathrm{y}_{i,1}, t_{i,1}),\ (\mathrm{y}_{i,2}, t_{i,2}),\ \ldots,\ (\mathrm{y}_{i,m_i}, t_{i,m_i})\Big\},\ \text{pour}\ \begin{array}{c} 1\leq i\leq n \text{ et } 1\leq j\leq m_i \\ t_{i,1}<t_{i,2}<\cdots<t_{i,m_i}\in \mathrm{T} \end{array}$$

Pour i et j fixés, $\mathrm{y}_{i,j}$ est la capture de la courbe $\mathrm{Y_i}$ au point $t_{i,j}$. Et nous devons reconstituer pour chaque i, la trajectoire $\mathrm{Y_i}(.)$ de sorte que :

$$\mathrm{y}_{i,j} = \mathrm{Y}_i(t_{i,j}) + \varepsilon_{i,j} \quad \text{avec } \varepsilon_{i,j} \text{ bruit blanc i.i.d. non observé.}$$

Une méthode courante pour reconstruire $\mathrm{Y_i}$ est de l'exprimer dans une base de fonctions (B-Splines, Fourier, ...).

Dans le cas d'une régression de variable fonctionnelle sur variables fonctionnelles, l'un des modèles proposé par [1] consiste à prendre en compte l'influence de chaque régresseur

fonctionnel sur tout son domaine de définition. Ce modèle appelé *intégral* s'écrit :

$$\mathrm{Y_i}(t_{i,j}) = \beta_0(t_{i,j}) \, + \, \sum_{k=1}^{p} \int_{\mathrm{T}} \beta_k(s, t_{i,j}) \, \mathrm{X_{k,i}}(s) \, ds. \tag{1}$$

Il est alors d'usage [5] de supposer que les coefficients fonctionnels $\beta_0 : \mathrm{T} \longrightarrow \mathbb{R}$ et $\beta_k : \mathrm{T} \times \mathrm{T} \longrightarrow \mathbb{R}$ s'expriment eux aussi dans une base de fonctions, sous la forme :

$$\beta_0(t_{i,j}) = \sum_{l=1}^{\mathrm{L_0}} \mathrm{b}_{0,l} \, \varphi_{0,l}(t_j) \qquad \text{et} \qquad \beta_k(s, t_{i,j}) \;\; = \sum_{l=1}^{\mathrm{L_1^k} \times \mathrm{L_2^k}} \mathrm{b}_{k,l} \; \Phi^k_{1,l}(t_{i,j}) \otimes \Phi^k_{2,l}(t_{i,j})$$

où $(\varphi_{0,l})_l$, $(\Phi^k_{1,l})_l$, et $(\Phi^k_{2,l})_l$ sont des bases de fonctions connues, ou tout du moins choisies par l'utilisateur.
Dans ce cas, l'équation (1) s'écrit en termes de coefficients fonctionnels $\{\mathrm{b}_{kl}\}_{k,l}$ sous forme du modèle de régression multivariée dont le procédé d'estimation est bien connu [5].

2 Modèle de régression sur les signatures

L'utilisation des signatures dans les modèles d'apprentissage statistique s'est beaucoup développé ces dernières années et semble spécialement approprié aux situations où les données sont définies sur les intervalles de tailles diverses. Les références [2, 3] fournissent une excellente introduction et mettent en exergue le fait que cette approche est très adaptée aux données collectées sous forme de fonctions multidimensionnelles du temps.
Soit $\mathrm{X} : \mathrm{T} = [a, b] \longrightarrow \mathbb{R}^d$ avec $d \geq 2$. On appelle signature de X, la collection (infinie) $\mathrm{S(X)}_{a,b}$ de tenseurs d'ordre croissant définis par les intégrales itérées :

$$\mathrm{S(X)}_{a,b} = \Big\{ 1 \,;\, \underbrace{\mathrm{S(X)}^1_{a,b} \,;\, \ldots \,;\, \mathrm{S(X)}^d_{a,b}}_{\text{coef. d'ordre 1}} \,;\, \underbrace{\mathrm{S(X)}^{1,1}_{a,b} \,;\, \mathrm{S(X)}^{1,2}_{a,b} \,;\, \ldots \,;\, \mathrm{S(X)}^{d,d}_{a,b}}_{\text{coef. d'ordre 2}} \,; \ldots \Big\} \tag{2}$$

où les coefficients d'ordre 1 sont définis par :

$$\mathrm{S(X)}^i_{a,b} = \int_{a<s<b} \mathrm{dX}^i(s) = \mathrm{X}^i(b) - \mathrm{X}^i(a), \quad 1 \leq i \leq d,$$

les coefficients d'ordre 2 par :

$$\mathrm{S(X)}^{i,j}_{a,b} = \int_{a<s<b} \mathrm{S(X)}^i_{a,s} \, \mathrm{dX}^j(s) = \int_{a<r<s<b} \mathrm{dX}^i(r) \, \mathrm{dX}^j(s), \quad 1 \leq i, j \leq d,$$

et ainsi de suite, ceux d'ordre k se calculant récursivement par la relation :

$$\mathrm{S(X)}^{i_1, i_2, \ldots, i_k}_{a,b} = \int_{a<s<b} \mathrm{S(X)}^{i_1, i_2, \ldots, i_{k-1}}_{a,s} \, \mathrm{dX}^{i_k}(s), \quad 1 \leq i_1, \ldots i_k \leq d.$$

On définit également la signature tronquée à l'ordre m de X, notée $\mathrm{S}^m(\mathrm{X})_{a,b}$, qui est la série (finie) contenant tous les coefficients de la signature d'ordre inférieur ou égal à m. Cette série est de dimension $s_d(m) = \sum_{l=0}^{m} d^l = \frac{d^{m+1}-1}{d-1}$; cf [3].
Le modèle de régression linéaire sur les signatures définie dans [3] cherche à modéliser une relation linéaire entre une cible $\mathrm{Y} \in \mathbb{R}$ et une variable fonctionnelle multidimensionnelle X. Des conditions garantissent l'existence de $m \in \mathbb{N}$ et $\beta_m^\star \in \mathbb{R}^{s_d(m)}$ telles que :

$$\mathbb{E}\big[\mathrm{Y} \,|\, \mathrm{X}\big] \;= \langle\, \beta_m^\star \,,\, \mathrm{S}^m(\mathrm{X})_{a,b} \,\rangle \quad \text{et} \quad \mathrm{Var}(\mathrm{Y} \,|\, \mathrm{X}) = \sigma^2 \leq \infty \tag{3}$$

Il faut noter une simplification dans notre utilisation des signatures par rapport au modèle fonctionnel (1) où le choix de la base d'expansion est une question naturelle et assez complexe, alors que dans le problème (3) seule se pose la question du choix de l'ordre maximal m considéré. En adoptant les signatures, nous avons donc naturellement beaucoup moins d'hyperparamètres à calibrer que dans le modèle (1).
Un des objectifs de notre travail utilisant les signatures est d'aborder le cas où Y n'est plus scalaire comme dans [3] mais fonctionnel. Nous proposons alors de construire au préalable la forme fonctionnelle de $\mathrm{Y_i}$ pour $1 \leq i \leq n$ dans une base de fonctions connue de la forme $\mathrm{Y_i}(t) = \sum_{k=1}^{\mathrm{K}} \theta_{i,k}\ \phi_k(t)$ [1] de sorte que Y soit complètement identifiable par les coefficients $\big\{\Theta_i = (\theta_{i,1}, \ldots, \theta_{i,\mathrm{K}}),\ 1 \leq i \leq n\big\}$. Nous nous retrouvons ensuite avec un problème de régression multi-output sur les signatures de la forme :

$$\Theta_i = \langle\, \beta_m^\star \,,\, \mathrm{S}^m(\mathrm{X_i}) \,\rangle + \varepsilon_i \quad 1 \leq i \leq n.$$

Les coefficients fonctionnels $(\widehat{\Theta}_i)_i$ estimés pour ce modèle seront utilisés pour reconstruire la fonction cible (output) $\widehat{\mathrm{Y}}_\mathrm{i}$.

3 Application: la prédiction de la QoE

Nous nous tournons à présent vers l'implémentation proprement dite des modèles de régression présentés ci-dessus. Les données d'application sont issues du simulateur de réseau OMNeT[8] sur des conversations téléphoniques d'environ 50 secondes pour 800 clients mobiles à l'intérieur d'une cellule 4G. Différentes conditions de simulation ont été réalisées pour un total d'environ 10 000 clients. Tout au long de chaque conversation, dix métriques réseaux ont été collectées à une fréquence inférieure à la seconde et une note MOS (Mean Opinion Score) réelle comprise entre 1 et 5 (normalisé dans $[0, 1]$) traduisant cette QoE est produite après chaque prise de parole de l'auditeur de la conversation à l'aide du modèle de la norme UIT-T G.107.2 [9]. La fréquence d'enregistrement étant propre à la métrique et au client, nous avons pour chaque client $\mathrm{x_i} = (\mathrm{x_{i,1}}, \ldots, \mathrm{x_{i,10}})$ avec $\mathrm{x_{i,j}} \in \mathbb{R}^{l_{i,j}}$ avec $l_{i,j}$ le nombre d'enregistrements du client i sur la métrique j.

Dans le cas du modèle (1) nous avons opté pour la représentation en base de splines de nos différentes variables fonctionnelles. Le nombre K des fonctions de base obtenu par validation croisée leave-one-out est égal à 7. Une fois les fonctions représentées, nous lançons aisément la fonction `pffr` (Penalized Flexible Functional Regression) du package R `refund` [4].
Dans le cas du modèle (3) de régression sur les signatures, c'est le choix de l'ordre m qui contrôle le nombre de coefficients et la faisabilité du modèle. Son estimation telle que présentée dans [3] ne peut se faire rigoureusement ici du fait que le jeu de données à notre disposition ne soit pas exactement sous la même forme ie régulièrement échantillonnée sur les variables pour un même individu. La réflexion en cours sur ce problème, les valeurs de la signatures que nous utiliserons ici seront dépourvues d'indices croisés. Nous ne travaillerons donc qu'avec des termes $\mathrm{S(X)}_{a,b}^{i_1,i_2,\ldots,i_k}$ tels que $i_1 = i_2 = \cdots = i_k \in \{1 \ldots d\}$.
Nous avons effectué une validation croisée pour obtenir $m = 5$.
Pour comparer les performances de ces deux modèles, nous calculons l'erreur quadratique moyenne de prédiction (RMSPE) sur différents échantillons test grâce à la validation croisée K-fold ($K = 10$). Nous obtenons une erreur moyenne de 0.20 pour le modèle fonctionnel traditionnel et de 0.11 pour le modèle utilisant les signatures.
Une observation de quelques trajectoires prédites représentée dans la Figure 1 ci-dessous confirme le RMSPE observé qui montre les meilleurs performances du modèle sur les signatures. On note principalement sa plus grande flexibilité dans l'ajustement des grosses variations de MOS.

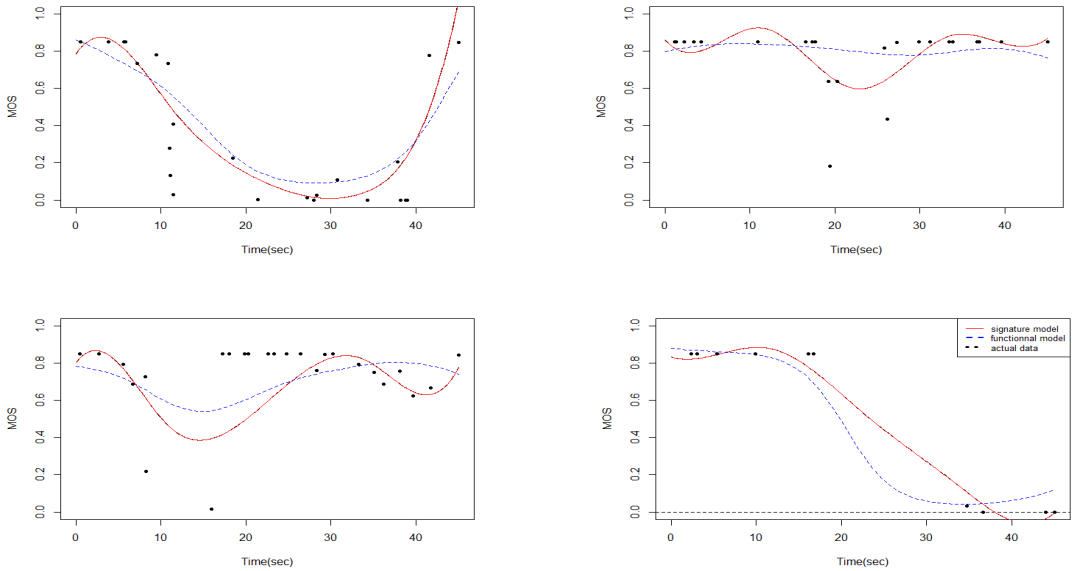


Figure 1: Prédiction des trajectoires de 4 clients. Observations (points noirs), prédiction avec le modèle fonctionnel (pointillés bleu), avec le modèle sur signatures (rouge)

Conclusion

L'objet du présent travail était l'étude comparée d'un modèle fonctionnel traditionnel et d'un modèle basé sur les signatures pour la prédiction de la qualité d'un conversation téléphonique. Nous avons principalement constaté que les signatures ont le bénéfice de nous affranchir de l'étape de la représentation des fonctions de l'ADF alors que cette étape reste incontournable dans l'approche fonctionnelle classique. Les résultats de prédiction nous conduisent à penser que les signatures sont une très bonne piste à poursuivre pour les problèmes de prédiction fonctionnelle avec covariables fonctionnelles. Les signatures présentent également l'avantage de permettre l'analyse de données irrégulièrement échantillonnées au sein d'un même individu.
L'un des défis à affronter dans la suite de notre travail est de tirer profit de la structure des données à analyser pour comprendre comment passer à l'échelle lorsque nous auront à traiter de nombreuses cohortes conjointement.

References

[1] J. O. Ramsay and B. W. Silverman. Functional Data Analysis 2nd ed., 2005. ser. Springer Series in Statistics. New York: Springer-Verlag.

[2] I. Chevyrev and A. Kormilitzin. A primer on the signature method in machine learning, 2016. http ://arxiv.org/abs/1603.03788 arXiv :1603.03788.

[3] A. Fermanian. Learning time-dependent data with the signature transform, PhD Thesis, 2021, Sorbonne Université, Paris.

[4] A. E. Ivanescu, A.-M. Staicu, F. Scheipl, and S. Greven. Penalized function-on-function regression, 2015 Comput Stat 30:539–568

[5] D. Ruppert, M. P. Wand and R. J. Carrol. Semiparametric Regression, 2003. Cambridge University Press, Cambridge, UK.

[6] S. N. Wood. Generalized additive models: An introduction with R, 2006. Chapman & Hall/CRC, New York

[7] K.-T. Chen. Integration of paths - a faithful representation of paths by non-commutative formal power series, 1958. Transactions of the American Mathematical Society 89 395–407.

[8] A. Varga and OpenSim Ltd. OMNeT++ Simulation Manual version 6.5.1, 2016

[9] Union Internationale des Télécommunications, Recommandation G.107.2 "Fullband E-model", 06/2019

Sélection d'intervalles pour des prédicteurs fonctionnels à partir de forêts aléatoires

Rémi Servien [1,2] & Nathalie Vialaneix [3]

[1] *INRAE, Univ. Montpellier, LBE, 102 Avenue des étangs, F-11000 Narbonne, France*
[2] *ChemHouse Research Group, Montpellier, France*
[3] *Université de Toulouse, INRAE, UR MIAT, F-31320, Castanet-Tolosan, France*

Résumé. Nous nous intéressons ici au problème de sélection de variables dans un cadre de régression fonctionnelle. Le but est de sélectionner des points de mesure consécutifs afin de déterminer les intervalles importants dans la prédiction de la variable cible. Pour cela nous nous basons sur les forêts aléatoires et évaluons des variantes possibles pour trois étapes de l'approche générale que nous proposons (création de groupes, définition des résumés, sélection) que nous comparons sur des données simulées et réelles.

Mots-clés. Régression fonctionnelle, sélection d'intervalles, forêts aléatoires.

Abstract. In this communication, we focus on the problem of variable selection in a functional regression framework. Our goal is to select consecutive measurement points to define the most important intervals for prediction in functional regression. The proposed approach is based on random forest and we evaluate different variants for three steps defining our approach (interval creation, summary definition, selection), which are compared on both simulated and real datasets.

Keywords. Functional regression, interval selection, random forest.

1 Contexte

Dans de nombreuses applications, les données se présentent sous la forme de vecteurs de très grande dimension qui sont la discrétisation d'un phénomène continu pour lesquelles des mesures ont été réalisées : données météorologiques (« courbes » de températures et précipitations), spectres en chimiométrie haut-débit, ... Ces données sont généralement désignées sous le terme commun de *données fonctionnelles* et, depuis longtemps, tout un pan de la littérature statistique s'intéresse à adapter les analyses statistiques aux spécificités de ce type de données [Ramsay and Silverman, 1997, Ferraty and Vieu, 2006].

Dans cette proposition de communication, nous nous intéresserons, en particulier, à un problème de sélection de variables adapté au cadre fonctionnel. Ce travail se place dans le contexte de la régression fonctionnelle, dans lequel une variable réelle, Y, doit être prédite par une variable aléatoire fonctionnelle, X. Les approches de sélections multi-dimensionnelles de sélection de variables peuvent être utilisées sur la discrétisation observée de la variable X mais donnent alors des résultats peu interprétables et qui ne

rendent pas compte de la nature spécifique des données. En effet, cette approche tend à produire des sélections de points de mesure de X isolés alors que la variable cible Y est intrinsèquement dépendante d'un ou plusieurs intervalles constitués de points de mesure consécutifs.

Nous proposons ici une approche basée sur la méthode des forêts aléatoires [Breiman, 2001] qui a montré ses bonnes performances dans les problèmes de prédiction en grande dimension. Nous proposons une approche en trois étapes et évaluons des variantes possibles pour ces trois étapes. L'intégralité de la procédure est testée sur un problème simulé (utilisant des courbes de données météorologiques réelles) et sur un problème réel dans lequel la vérité terrain est connue.

2 Description de l'approche proposée

Dans la suite, on notera (X, Y), le couple de variables aléatoires dans lequel X est une covariable fonctionnelle et Y est une variable catégorielle ou continue à prédire. On dispose de n observations i.i.d. de (X, Y), $(x_i, y_i)_{i=1,\dots,n}$ pour lequelles x_i est observée partiellement aux temps de mesure t_1, t_2, $\dots$, t_d. On note alors $\mathbf{x}_i = (x_i(t_1), \dots, x_i(t_d))^\top \in \mathbb{R}^d$.

L'approche proposée se définit en trois étapes :

— la **création de groupes** de prédicteurs ;
— la **définition de variables résumées** pour chacun de ces groupes ;
— la **sélection de variables**.

2.1 Création de groupes de prédicteurs

L'objectif de cette étape est de fournir aux modèles prédictifs une partition de $\{t_1, \dots, t_d\}$ en groupes de temps de mesures adjacents. La difficulté réside ici dans le fait que, pour d temps de mesure, le nombre de partitions possibles des prédicteurs (2^d) est rapidement impossible à explorer de manière exhaustive. Aussi, nous proposons d'utiliser des approches gloutonnes permettant d'explorer une hiérarchie de regroupements.

Parmi les approches sélectionnées, explicitées dans l'algorithme 1 , nous testons :

— la classification ascendante hiérarchique sous contrainte de contiguïté qui utilise la perte d'inertie intra-groupes minimale basée sur la corrélation entre temps de mesure, $\mathrm{Cor}((x_i(t_j))_i^\top, (x_i(t_{j+1}))^\top)$ comme implémentée dans **adjclust** [Ambroise et al., 2019] ;
— une version sous contrainte de contiguïté de l'approche **ClustOfVar** [Chavent et al., 2012] ;
— une approche basée sur la régression linéaire dans laquelle l'étape 1 est basée sur le critère

$$\arg \min_{j=1,\dots,p-T} \mathrm{MSE}(X(t_j), X(t_{j+1}) - \max\left(\mathrm{MSE}(X(t_j)),\ \mathrm{MSE}(X(t_{j+1}))\right.$$

où MSE(.) est la valeur du critère des moindres carrés pour la prédiction de Y à partir de, respectivement $(X(t_j), X(t_{j+1})$, $X(t_j)$ et $X(t_{j+1})$.

Algorithm 1 Création de groupes par approche gloutonne

$\mathcal{D}^0 = (C_i^0)_{1 \leq i \leq p}$ with $C_i^0 = \{x_i\}$ ▷ Initialisation
for $T = 1$ to $p - 1$ **do**
    Trouver le temps de mesure $t_{j^\star}$ ($j^\star \in \{1, \ldots, p - T\}$) tel que le groupe de temps de mesure $\{t_{j^\star}, t_{j^\star+1}\}$ est optimal pour un certain critère ▷ Meilleur candidat
    for $j = 1$ to $p - T - 1$ **do** ▷ Mise à jour de $\mathcal{D}^{T-1}$ en $\mathcal{D}^T$
        if $j < j^\star$ **then** $C_j^T = C_j^{T-1}$
        else if $j = j^*$ **then** $C_j^T = C_j^{T-1} \cup C_{j+1}^{T-1}$
        else if $j > j^\star$ **then** $C_j^T = C_{j+1}^{T-1}$
        end if
    end for
end for

Notons que les deux premières approches définissent des groupes de points d'observations de X complètement non supervisés alors que la dernière utilise la capacité prédictive des points d'observations pour définir les groupes. Cette étape fournit donc D partitions de prédicteurs, $\mathcal{D}_1, \ldots, \mathcal{D}_D$, où, dans notre cas, D est toujours égal à d.

2.2 Définition de variables résumées

Pour chaque partition $\mathcal{D}_k$ ainsi obtenue, la **définition de variables résumées** consiste à résumer les variables $X(t_j)_{j \in C_\ell}$ pour un $C_\ell \in \mathcal{D}_k$. Pour chacune des partitions issues de chacune des trois méthodes présentées plus haut, nous définissons trois approches permettant de définir un résumé :

— comme dans [Deng et al., 2013] ("Time Series Forest"), chaque groupe est résumé par la valeur moyenne et l'écart type sur les temps de mesure inclus dans le groupe (deux variables créés par groupe, approche appelée « basique » par la suite) :

$$\tilde{X}_\ell^1 = \frac{1}{|C_\ell|} \sum_{j \in C_\ell} X(t_j) \qquad \text{et} \qquad \tilde{X}_\ell^2 = \sqrt{\frac{1}{|C_\ell|} \sum_{j \in C_\ell} (X(t_j) - \tilde{X}_\ell^1)^2} \; ;$$

— comme dans [Poterie et al., 2019, Menze et al., 2011], chaque groupe est résumé en utilisant une information utilisant la variable cible Y : meilleure combinaison linéaire au sens de la régression linéaire, ridge, ou PLS ;

— lorsque les groupes ont été définis avec la version contrainte de **ClustOfVar**, comme [Chavent et al., 2021], il est aussi possible de résumer les points de mesure du groupe en utilisant la variable synthétique du groupe [Chavent et al., 2012].

2.3 Sélection de variables

Pour sélectionner les intervalles importants, nous utilisons deux approches de sélection de variables adaptées aux forêts aléatoires : l'approche **VSURF** [Genuer et al., 2010] (essentiellement basé sur l'utilisation de la notion d'importance) et l'approche **Boruta** [Kursa and Rudnicki, 2010], basée sur une approche de knockoffs [Barber and Candès, 2015]. Ces approches de sélection sont confrontées à une approche utilisant directement l'importance des variables pour repérer les intervalles importants.

3 Simulations

3.1 Données

L'intégralité des comparaisons est effectuée sur deux problèmes :

— **Données simulées** : nous avons utilisé les mêmes séries climatiques que décrites dans [Picheny et al., 2019] ($n = 1000$ séries temporelles mesurées à $t \in [\![287, 730]\!]$). Ces données ont été générées à l'aide du simulateur de données météorologiques WACSGen [Flecher et al., 2010], reproduisant le climat de Lleida, Catalogne, Espagne entre 1981 et 1982, pour générer les données selon le modèle (inspiré par [Grollemund et al., 2019]) :

$$\forall\, i = 1, \dots, 1000, \qquad y_i = \log\left(1 + |\langle x_i, \beta \rangle|\right) + \epsilon_i \tag{1}$$

où $\beta(t) = 4 \times \mathbf{1}_{\{t \in [320,410]\}} + 2 \times \mathbf{1}_{\{t \in [500,550]\}} - \mathbf{1}_{\{t \in [680,730]\}}$ et $\epsilon_i \sim \mathcal{N}(0, 0.5)$ sont i.i.d. ;

— **Truffes** : ce jeu de données est décrit en détails dans [Baragatti et al., 2019]. Il consiste en 25 années de production de truffes (rendement en kilos) pour 4 variables climatiques associées (température maximale, précipitation, bilan hydrique et déficit hydrique) fournies sous la forme de séries temporelles et mesurées mensuellement du mois de janvier de l'année n au mois de mars de l'année $n + 1$. Une connaissance experte sur ces données permet de savoir quelles périodes doivent être retrouvées comme importantes, celles-ci étant potentiellement différentes pour chaque variable explicative.

3.2 Quelques résultats et discussions

Nous comparons les différentes combinaisons des étapes 1 et 2 par un test de différence des valeurs d'importance entre les points de mesure connus pour être importants et ceux connus pour ne pas l'être (test de Wilcoxon). Le tableau 1 donne la p-valeur minimale (sur l'ensemble de la hiérarchie de groups telle que définie dans la section 2.1) pour chaque combinaison de méthodes (ainsi que le nombre de groupes associés) pour une des variables des données simulées.

Création	Définition	p-valeur ($-\log_{10}$)	Nombre de groupes
adjclust	basique	-48.3	24
	modèle linéaire	46.1	43
	PLS	34.7	64
	ridge	40.6	69
ClustOfVar	basique	-45.9	101
	ClustOfVar	51.8	109
	modèle linéaire	44.4	118
	PLS	43.0	112
	ridge	46.4	118
modèle linéaire	basique	-12.0	369
	modèle linéaire	17.2	52
	PLS	17.2	52
	ridge	14.8	55

TABLE 1 – **Données simulées.** Résultats comparatifs en terme de p-valeurs (voir le texte) et de nombre de groupes..

Les p-valeurs sont très petites, ce qui est encourageant. La meilleure méthode semble être l'enchaînement des deux étapes avec ClustOfVar. Des résultats similaires ont été obtenus pour toutes les variables des données « truffes » : la méthode ClustOfVar est systématiquement la meilleure pour créer les groupes. Si le nombre de groupes semblent, en revanche, important, il faut noter qu'il ne correspond pas complètement aux nombres d'intervalles qui seront finalement considérés comme « importants » (plusieurs groupes « importants » pouvant être adjacents et ne constituer qu'un seul intervalle « important »). La figure 1 montre les valeurs d'importance sur tout l'ensemble de définition, pour l'approche ClustOfVar, soit en utilisant l'importance initiale, soit ne conservant que les valeurs d'importance supérieure à $|\min_{C^t \in \mathcal{D}^t} \mathcal{I}_{\mathcal{C}^t}|$ où $\mathcal{I}_c$ est l'importance du résumé correspondant au groupe c (approche dite « seuillée »). Celles-ci sont comparées à la vraie variable β telle que définie dans l'équation (1).

Ces résultats sont prometteurs car nous retrouvons globalement les bons intervalles : les deux premiers intervalles, en partculier, sont presque parfaitement retrouvés mais le dernier est, par contre, perdu. Nous espérons que l'étape de sélection, que nous n'avons pas encore évaluée à l'heure d'écriture de cette présentation, permettra d'améliorer les résultats qui seront présentés de manière complète lors de la communication orale.

Remerciements

Nous remercions la plateforme bioinformatique Genotoul Bioinfo `http://bioinfo.genotoul.fr/` pour la mise à disposition des ressources de calcul. Nous remercions également le métaprogramme INRAE DIGIT-BIO pour le financement du projet PhenoDyn qui soutient ces recherches. Enfin, nous remercions Pierre Casadebaig, Ronan Trepos, Victor Picheny, Meili Baragatti et François Le Tacon pour la mise à disposition des données de climats simulées et des données « truffes ».

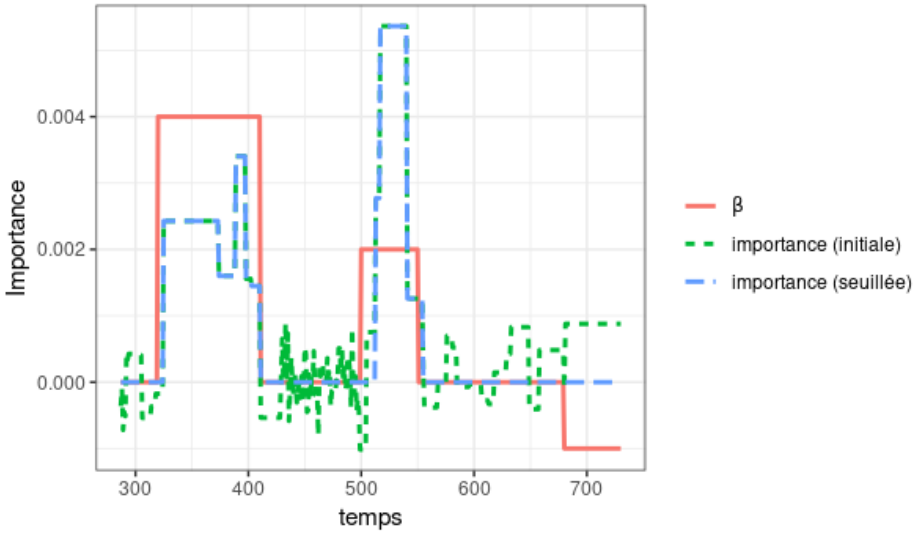


FIGURE 1 – **Données simulées.** Importances initiale et seuillée, comparées à la valeur de β, prédicteur ayant permis de générer les données.

Bibliographie

[Ambroise et al., 2019] Ambroise, C. et al. (2019). *Algorithm Mol Biol*, 14 :22. doi.

[Baragatti et al., 2019] Baragatti, M. et al. (2019). *Mycorrhiza*, 29(2) :113–125. doi.

[Barber and Candès, 2015] Barber, R. and Candès, E. (2015). *Ann Stat*, 43(5) :2055–2085. doi.

[Breiman, 2001] Breiman, L. (2001). Random forests. *Mach Learn*, 45(1) :5–32. doi.

[Chavent et al., 2012] Chavent, M. et al. (2012). *J Stat Softw*, 50(13) :1–16. doi.

[Chavent et al., 2021] Chavent, M. et al. (2021). *Commun Stat Simul Comput*, 50(2) :426–445. doi.

[Deng et al., 2013] Deng, H. et al. (2013). *Inf Sci*, 239 :142–153. doi.

[Ferraty and Vieu, 2006] Ferraty, F. and Vieu, P. (2006). *NonParametric Functional Data Analysis*. Springer. doi.

[Flecher et al., 2010] Flecher, C. et al. (2010). *Wat Resour Res*, 46(7) :W07519. doi.

[Genuer et al., 2010] Genuer, R. et al. (2010). *Patt Recognit Lett*, 31(14) :2225–2236. doi.

[Grollemund et al., 2019] Grollemund, P.-M. et al. (2019). *Bay Anal*, 14(1) :111–135. doi.

[Kursa and Rudnicki, 2010] Kursa, M. and Rudnicki, W. (2010). *J Stat Softw*, 36(11) :1–13. doi.

[Menze et al., 2011] Menze, B. H. et al. (2011). In *Proc of Mach Learn and Know Disc*, volume 6912, pages 453–469. Springer Berlin Heidelberg. doi.

[Picheny et al., 2019] Picheny, V. et al. (2019). *Stat Comput*, 29(2) :255–267. doi.

[Poterie et al., 2019] Poterie, A. et al. (2019). *Comp Stat*, 34 :1613–1648. doi.

[Ramsay and Silverman, 1997] Ramsay, J. and Silverman, B. (1997). *Functional Data Analysis*. Springer Verlag, New York. doi.

Statistique mathématique

Multivariate Poisson Lognormal model: optimisation, inference and application to high dimensional data

Bastien Batardière [1,†] & Julien chiquet [2,†] & Joon Kwon[3,†] & Laure Sansonnet [4,†]

[1] *bastien.batardiere@inrae.fr,* [2] *julien.chiquet@inrae.fr,* [3] *joon.kwon@inrae.fr ,* [4] *laure.sansonnet@agroparistech.fr*

[†] *Université Paris-Saclay, AgroParisTech, INRAE, UMR MIA Paris-Saclay*

Résumé. En analyse unicellulaire, il est habituel de traiter des observations non gaussiennes de grande dimension. Un exemple typique est la distribution de comptage transcriptomique unicellulaire brute, qui vise à comprendre les covariations entre l'activité de chaque gène (nombre de fois où le gène est exprimé) dans chaque cellule. Alors que le cadre gaussien fournit une façon canonique de modéliser de telles dépendances avec des données continues, il ne s'applique pas aux données de comptage, contrairement au modèle Poisson lognormal (PLN). De plus, des techniques efficaces de réduction de dimension doivent être utilisées afin de traiter des données à haute dimension. Le modèle Poisson lognormal PCA (PLN-PCA) nous permet de traiter ces deux problématiques. Cependant, l'inférence est délicate car elle porte sur des intégrales insolubles. Nous nous concentrons sur deux approches d'inférence différentes : un algorithme variationnel qui maximise une borne inférieure de la vraisemblance (*Evidence Lower BOund* – ELBO) et une autre utilisant des méthodes de Monte-Carlo. Aucune garantie théorique n'existe quant à l'erreur faite par l'approximation variationnelle. Nous montrons numériquement que l'approximation faite dans le cadre de PLN-PCA est efficace, en prenant comme référence la vraisemblance calculée par Monte-Carlo. Les complexités algorithmiques des deux approches évoluent linéairement en fonction du nombre de gènes et de cellules pour la partie PCA. Nous étudions également la différence entre les techniques avec et sans réduction de dimension (PLN/PLN-PCA).

Mots-clés. Apprentissage statistique, données de comptage, statistiques, grande dimension, génomique, réduction de dimension, inférence variationnelle.

Abstract. In Single-cell analysis, it is usual to deal with high-dimensional non-Gaussian observations. A typical example is the raw single-cell transcriptomic count distribution, which aims to understand the co-variations of each gene's activity (number of times the gene is expressed) in each cell. While the Gaussian setting provides a canonical way of modeling such dependencies with continuous data, it does not apply to count data. The Poisson lognormal model (PLN) does. Moreover, efficient dimension reduction techniques need to be performed to handle high-dimensional data. The Poisson lognormal PCA (PLN-PCA) model address this additional issue. However, inference is tricky since it deals with intractable integrals. We focus on two different inference approaches: a variational algorithm maximizing an Evidence Lower BOund (ELBO) and a Monte-Carlo approach. No theoretical guarantee exists for the error made by the ELBO maximization. We study this error thanks to the estimation of the log-likelihood and approve the validity of the approximation. The algorithmic complexities of both approaches evolve linearly with the number of genes and cells for the PCA part. We also study the difference between techniques with dimension reduction (PLN-PCA) and those without (PLN).

Keywords. Machine learning, count data, statistics, high dimension, genomics, dimension reduction, variational inference

1 Context

In Single-cell analysis, it is usual to deal with high-dimensional non-Gaussian observations. The Gaussian setting is natural to describe dependencies in continuous data. The Poisson lognormal (PLN, see Aitchison and Ho, 1989) relies on a latent Gaussian layer to model such dependencies. This layer is then sent in the integer space with a Poisson layer.

2 The Poisson lognormal PCA model

The model is the following: let $(n, p, q, d) \in \mathbb{N}_\star^4$, such as n is the number of cells ($i = 1, \ldots, n$), p the number of genes ($j = 1, \ldots, p$), q the number of latent dimensions ($k = 1, \ldots, q$) and d the number of covariates ($h = 1, \ldots, d$). We observe

- n measures $Y_i = (Y_{ij})_{1 \leqslant j \leqslant p}$ for the vector of counts, Y_{ij} giving the number of times gene j is expressed in cell i.
- n measures $X_i = (x_{ih})_{1 \leqslant h \leqslant d}$ for the vector of covariates, X_{ih} giving the value of descriptor h for cell i.

We assume that for all $1 \leqslant i \leqslant n$, the observed counts $(Y_{ij})_{1 \leqslant j \leqslant p}$ are independent conditionally on a latent variable $W_i \in \mathbb{R}^q$ such that :

$$\begin{aligned} W_i &\sim \mathcal{N}(0, I_q) \\ Z_i &= \beta^\top X_i + C W_i \\ (Y_{ij} \mid Z_{ij}) &\sim \mathcal{P}(\exp(Z_{ij})) \end{aligned} \tag{1}$$

where the model parameters encompass

- $\beta = (\beta_{kj})_{1 \leqslant k \leqslant d, 1 \leqslant j \leqslant p}$, the unknown regression parameters and
- $C \in \mathbb{R}^{p \times q}$, an unkwon matrix sending the latent variable W_i from a space of dimension q to a space of dimension p.

We denote $Y \in \mathbb{R}^{n \times p}$ (resp. $X \in \mathbb{R}^{n \times d}, Z \in \mathbb{R}^{n \times p}$) the matrix obtained by vertically stacking the Y_i's (resp. X_i, Z_i). The unknown (and identifiable) parameter is $\theta = \{\Sigma, \beta\}$, where $\Sigma = CC^\top$ is the covariance matrix of each Z_i. The model latent variable can either be seen as W or Z. However, Z depends on θ, which is not the case of C. Note that multiplying C by an orthogonal matrix does not modify the model so C is not identifiable. In practice, the dimension $q \leqslant p$ is a hyperparameter that must be tuned. We will consider two very different cases:

- $q < p$. In this case, $\Sigma = CC^\top$ is not full rank and Z does not have a continuous density. We are performing dimension reduction since the p variables are defined through only $q < p$ variables. The model is then called PLN-PCA.
- $p = q$. In this case, Σ is invertible. It does not perform any dimension reduction. It is called PLN.

3 Inference

We aim at approximately solving

$$\theta^\star = \underset{\theta}{\operatorname{argmax}} \log p_\theta(Y) = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{n} \log p_\theta(Y_i) \tag{2}$$

We consider two approaches here:

1. Direct optimization by means of the gradient of the objective:

$$\nabla_\theta \sum_{i=1}^{n} \log p_\theta(Y_i) = \sum_{i=1}^{n} \nabla_\theta \log \left(\int_{R^q} p_\theta(Y_i|W)p(W) \right).$$

   The integral being intractable, we can rely on Monte-Carlo estimation and importance sampling.

2. Approximated Expectation-Maximization (EM): the model involving latent variables, we want to apply the EM algorithm. Given a starting point θ^0, it solves iteratively

$$\theta^{t+1} = \arg\max_\theta \mathbb{E}_{\theta^t} \left[\log p_\theta(Y, W) \mid Y \right].$$

   However, the expectation is hard to compute because of the conditional distribution of $W|Y$ which has no analytical form. We fall back on variational approximation and variational EM to circumvent this issue (VEM, see Blei et al., 2017). The VEM maximizes a lower bound of the log-likelihood called Evidence Lower BOund (ELBO):

$$J_Y(\theta, q) \triangleq \log p_\theta(Y) - KL\left[q(W) \| p_\theta(W \mid Y)\right]$$

   where KL is the Kullback-Leibler divergence and $q(.)$ is a distribution chosen among a family of candidate distributions denoted by $\mathcal{Q}$.

3.1 Importance sampling approach

We use Monte-Carlo methods to approximate the quantity $p_\theta(Y_i) = \mathbb{E}_{\mathcal{N}(0,I_q)}\left[p_\theta(Y_i \mid W)|Y_i\right]$. We then apply the logarithm and differentiate with respect to θ. However, the function $\tilde{p}_{\theta,i}^{(u)}$:

$$\begin{array}{rccl} \tilde{p}_{\theta,i}^{(u)} & : & \mathbb{R}^q & \to \mathbb{R}^+ \\ & & W & \mapsto p_\theta(Y_i|W)p(W) \end{array}$$

has little mass near zero, so that using the basic Monte-Carlo approach

$$p_\theta(Y_i) \approx \frac{1}{N_s} \sum_{j=1}^{N_s} p_\theta(Y_i \mid W_j), W_j \sim \mathcal{N}(0, I_q)$$

(where $N_s \in \mathbb{N}$ is the Monte-Carlo effort) would lead to a bad approximation with a very high variance. We use Importance sampling to circumvent this issue by putting mass where $\tilde{p}_{\theta,i}^{(u)}$ has some. It takes advantage of the equality $\mathbb{E}_{W\sim p}\left[p_\theta(Y_i|W)\right] = \mathbb{E}_g\left[\frac{p_\theta(Y_i|W)p(W)}{g(W)}\right]$ for any density g. The resulting estimator

$$\frac{1}{n_s}\sum_{k=1}^{n_s}\frac{\tilde{p}_{\theta,i}^{(u)}(V_k)}{g(V_k)}, \quad (V_k)_{1\leqslant k\leqslant n_s} \overset{iid}{\sim} g \tag{3}$$

is always consistent, but the variance of this estimator depends on the choice of g. One can show that the lowest variance is obtained by g defined as

$$g_{opt}(V) = \frac{\tilde{p}_{\theta,i}(V)}{p_\theta(Y_i)}, V \in \mathbb{R}^q, \tag{4}$$

which is in fact a zero-variance estimator. However, this quantity is untractable. We will approximate this function with a Gaussian density since $\tilde{p}_{\theta,i}^{(u)}$ is well Gaussian-approximated. We define the mean of the proposed Gaussian as the mode μ of $\tilde{p}_{\theta,i}^{(u)}$(computed performing gradient ascent) and the variance as the inverse second order of $\log(\tilde{p}_{\theta,i}^{(u)}(.))$ evaluated in μ. Indeed, if it were Gaussian, the mean and variance would be the mode and the inverse second order of the log-likelihood. We call IMPS_PLNPCA the resulting algorithm.

3.2 Variational EM

Variational inference for the PLN-PCA model has already been studied in Chiquet et al. (2018). We develop here a variational inference for the PLN model for comparison with PLN-PCA. The VEM for PLN is very fast (convergence in 50 iteration) for low p since it admits some analytical forms, unlike PLN-PCA (Σ is invertible in the PLN case, contrary to PLN-PCA). However, the number of parameters being p^2, the computational complexity is quadratic so that the time per iteration is considerable for large p. By contrast, the VEM for PLN-PCA needs to perform much more iterations (around 15000 when p is large), but each iteration is cheaper ($p \times q$ parameter). We call the VEM algorithm for the PLN (resp PLN-PCA) fastPLN (resp fastPLNPCA).

4 Results

We compared each method for $n = p = 1000, q = 10, d = 1$. The heatmaps of the resulting matrix Σ for each algorithm are displayed in Figure (1). The variational algorithms perform well in both cases, as well as the importance sampling-based algorithm. Obviously the latter is slower than the formers since it performs both gradient ascent (to find g_{opt}) and importance sampling estimation at each step. However, the achievable maximization of the log-likelihood remains its biggest advantage. Indeed, we do not have any theoretical guarantee about the validity of the ELBO maximization in contrary to log-likelihood maximization (variance estimation, central limit theorem for potential statistical tests).

We compared fastPLN and fastPLNPCA for large p in Figure (2). We also found a fast optimization for each algorithm to scale it to large amounts of data by comparing various algorithms (in Figure. 3). Rprop from Pytorch (Paszke et al., 2019) performed very well for

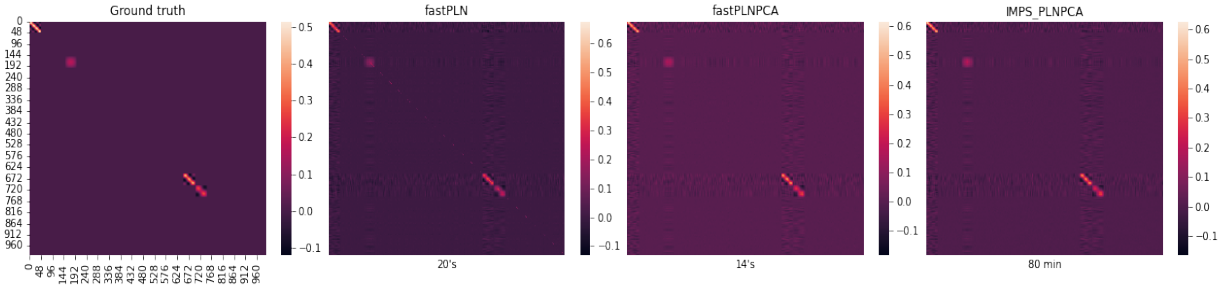


Figure 1: Heatmap of $\hat{\Sigma}$ for each algorithm. The ground truth is on the left. The simulation parameters were $n = p = 1000, q = 10, d = 1$. For each algorithm, the convergence time is given at the bottom.

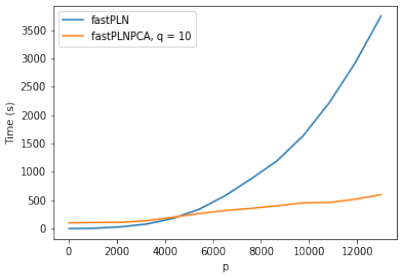


Figure 2: Convergence time comparison with respect to p for the variational algorithms, for $n = 1000, q = 10, d = 1$. We launch fastPLN (resp fastPLNPCA) for 60 (resp 15000) iterations. We clearly see the asymptotic convergence rate of both algorithms. Even for very large $p > 12000$, the convergence time remains quite low for fastPLNPCA.

fastPLN and fastPLNPCA compared to others optimizers. For IMPS_PLNPCA optimization, we used an adaptive stochastic gradient algorithm (AdaGrad, see Duchi et al., 2011) to increase the convergence rate significantly.

4.1 ELBO versus log-likelihood

The ELBO is a lower bound on the log-likelihood. We have no guarantee about its closeness to the true log-likelihood. In Section 3.1, we computed an approximated log-likelihood. We are then able to estimate the variational approximation error. We display in Figure 4 both the error made between ELBO and log-likelihood and the error made by the ELBO maximization compared to the log-likelihood maximization. As expected, the ELBO is always lower. The gap seems to be smaller at convergence. The maximum ELBO is actually better than the maximum obtained with the log-likelihood maximization. We would expect the opposite since we are finding a better log-likelihood when maximizing the ELBO. The error done when estimating the gradients must be greater than the variational approximation error, which could explain such a phenomenon.

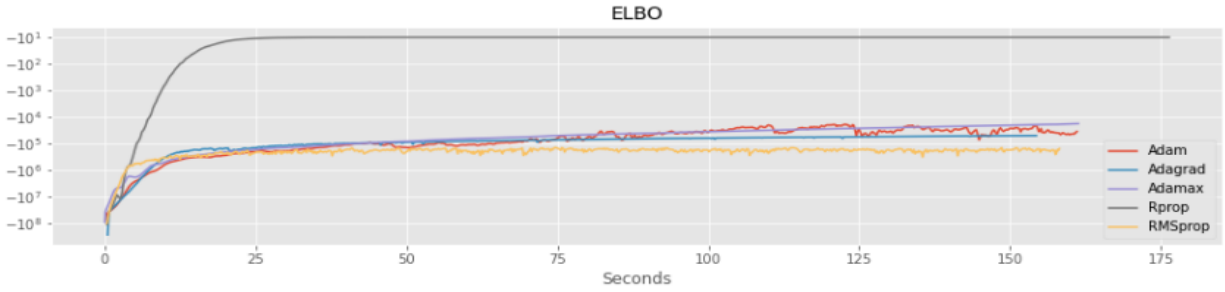


Figure 3: Simulation with $n = 1000, p = q = 200, d = 2$. The learning rate of each optimizer has been tuned. Rprop outperforms the other optimizers.

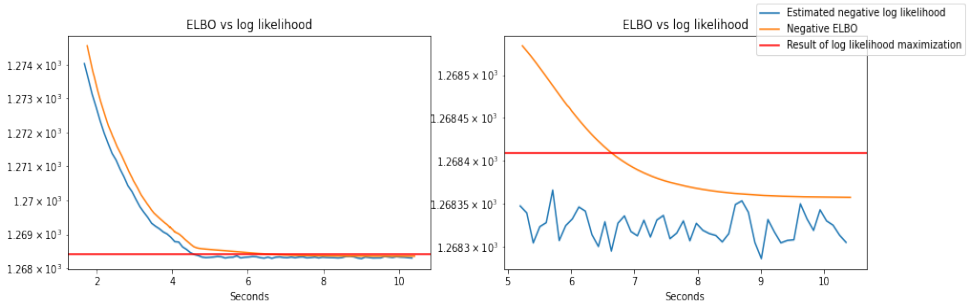


Figure 4: Simulation with $n = p = 1000, q = 10, d = 1$. The orange plot corresponds to the ELBO computed in fastPLNPCA. The blue plot corresponds to the log-likelihood computed with the current parameters of the algorithm. The right plot is the left one near convergence. We plot the negative values for a logarithmic scale. We estimate the distance between the ELBO and the log-likelihood at each point. In red, we display the maximum log-likelihood found by IMPS_PLNPCA.

References

J. Aitchison and C. H. Ho. The multivariate Poisson-log normal distribution. *Biometrika*, 76(4), 12 1989. ISSN 0006-3444. doi: 10.1093/biomet/76.4.643.

D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, page 859–877, Apr 2017.

J. Chiquet, M. Mariadassou, and S. Robin. Variational inference for probabilistic Poisson PCA. *The Annals of Applied Statistics*, 12(4):2674 – 2698, 2018.

J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.*, 12:2121–2159, jul 2011.

A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. Pytorch: An imperative style, high-performance deep learning library. pages 8024–8035. Curran Associates, Inc., 2019.

Les approximations à noyau à base de processus ponctuels déterminantaux

Ayoub Belhadji [1]

[1] *Univ Lyon, ENS de Lyon*
Inria, CNRS, UCBL
LIP UMR 5668, Lyon, France
ayoub.belhadji@ens-lyon.fr

Résumé.
On étudie l'approximation d'une fonction, qui vit dans un espace à noyau, par un mélange finie de translatées de noyau quand les noeuds de l'approximation suivent la distribution d'un processus ponctuel determinantal (DPP) adapté au noyau. On montre que le taux de convergence est presque optimal et dépend principalement des valeurs propres de l'opérateur d'intégration correspondant. En particulier, cette analyse unifiée permet d'avoir un nouveau regard sur l'utilisation des DPPs dans le domaine de l'intégration numérique.

Mots-clés. Processus ponctuels déterminantaux, Méthodes Monte Carlo, Interpolation, Quadrature, Methodes à noyau

Abstract. We study approximation problems in reproducing kernel Hilbert spaces (RKHS) using random nodes. More precisely, we focus on kernel quadrature and kernel interpolation for smooth functions living in an RKHS using nodes that follow the distribution of a determinantal point process (DPP) and mixtures of DPPs. The definition of these DPPs is tailored to the RKHS. We prove fast convergence rates that depend on the eigenvalues of the RKHS kernel. This unified analysis gives new insights on the rates of the quadratures and interpolations based on DPPs, especially for high dimensional numerical integration problems.

Keywords. Determinantal point processes, Monte Carlo methods, interpolation, quadrature, kernel methods

1 Introduction

We study the problem of approximating a function μ defined on some metric space $\mathcal{X}$ by a finite mixture of kernel translates $k(x_n, .)$

$$\mu \approx \sum_{n\in[N]} w_n k(x_n, .) \tag{1}$$

More precisely, $\mathcal{X}$ is assumed to be a measurable space equipped with a measure ω, and μ is assumed to write as

$$\mu = \mu_g := \int_{\mathcal{X}} k(x, .)g(x)\mathrm{d}\omega(x), \tag{2}$$

with $g \in \mathbb{L}_2(\omega)$. By construction, μ belongs to the RKHS $\mathcal{F}$ associated to the kernel k, which we assume to be embedded in $\mathbb{L}_2(\omega)$. Ensuring a "good" approximation in (1) is particulary useful for quadrature. Indeed, we have

$$\forall f \in \mathcal{F}, \ \left| \int_{\mathcal{X}} f(x)g(x)\omega(x) - \sum_{n\in N} w_n f(x_n) \right| \leq \|f\|_{\mathcal{F}} \left\| \mu_g - \sum_{n\in[N]} w_n k(x_n, .) \right\|_{\mathcal{F}}, \tag{3}$$

so that controlling the approximation error in r.h.s. of (3) yields a control on the integration error in the l.h.s. Moreover, $\|\mu_g - \sum_{n\in[N]} w_n k(x_n, .)\|_{\mathcal{F}}$ is actually the worst case integration error on the unit ball of $\mathcal{F}$.

Bach (2017) proposed to choose the *nodes* x_n to be independent random variables that follow a density q with respect to the measure ω, and to choose the weights $(w_n)_{n\in N}$ as the solution of the optimization problem

$$\min_{\boldsymbol{w}\in\mathbb{R}^N} \left\| \mu_g - \sum_{n\in[N]} \frac{w_n}{q(x_n)^{1/2}} k(x_n, .) \right\|_{\mathcal{F}}^2 + N\lambda\|\boldsymbol{w}\|^2 \tag{4}$$

This optimization problem admits a unique solution

$$\boldsymbol{w}^* = (\overline{\boldsymbol{K}}(\boldsymbol{x}) + \lambda\mathbb{I}_N)^{-1}\mu_g(\boldsymbol{x}), \tag{5}$$

where $\mu_g(\boldsymbol{x}) = (\mu_g(x_n))_{n\in[N]} \in \mathbb{R}^N$, and $\overline{\boldsymbol{K}}(\boldsymbol{x}) = (q(x_n)^{-1/2}k(x_n, x_{n'})q(x_{n'})^{-1/2})_{n,n'\in[N]}$. In particular, there exists a density q_λ^* called the *continuous ridge leverage score density* for which

$$\sup_{\substack{g\in\mathbb{L}_2(\omega)\\ \|g\|_\omega\leq 1}} \left\| \mu_g - \sum_{n\in[N]} \frac{w_n^*}{q_\lambda^*(x_n)^{1/2}} k(x_n, .) \right\|_{\mathcal{F}}^2 \leq \lambda, \tag{6}$$

with high probability, whenever $N \geq C d_{\text{eff}}(\lambda)\log d_{\text{eff}}(\lambda)$, with $d_{\text{eff}}(\lambda)$ is the effective degree of freedom defined by

$$d_{\text{eff}}(\lambda) = \operatorname{Tr} \boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \lambda\mathbb{I}_{\mathcal{F}})^{-1}, \tag{7}$$

and $\boldsymbol{\Sigma}$ is the integration operator associated to the kernel k and the measure ω, defined on $\mathbb{L}_2(\omega)$ by $\boldsymbol{\Sigma}g = \int_{\mathcal{X}} g(.)k(x, .)\mathrm{d}\omega(x)$. The continuous ridge leverage score density has a spectral expression

$$q_\lambda^*(x) = \sum_{n\in\mathbb{N}^*} \frac{\sigma_n}{\sigma_n + \lambda} e_n(x)^2, \tag{8}$$

where the (σ_n, e_n) are the eigenpairs of the integration operator $\boldsymbol{\Sigma}$.

In a nutshell, this result implies that the resulting quadrature $(\boldsymbol{x}, \boldsymbol{w}^*)$ has a worst case integration error that goes below λ, with high probability, if $N \geq C d_{\text{eff}}(\lambda) \log d_{\text{eff}}(\lambda)$. Yet, for a fixed level of regularization λ, sampling more nodes from q_λ^* does not improve the worst case integration error. Nevertheless, one can build up an infinite sequence of quadratures $(\boldsymbol{x}_t, \boldsymbol{w}_t^*)_{t \in \mathbb{N}^*}$, each associated to a regularization parameter λ_t, such that $\lim_{t \to +\infty} \lambda_t \to 0$. This way, Bach (2017) was able to derive the rates of convergence of these sequence of quadratures in the limiting asymptotic $\lambda \to 0$ in different RKHSs. While intricate, this method allows to prove that the rates of the resulting quadratures are almost optimal, up to logarithmic terms. Moreover, these rates depend on the eigenvalues σ_n of the integration operator: the smoother the kernel, the faster the convergence to 0. However, the ridge leverage score density (8) may not have a tractable expression, in general; even worse the spectral decomposition of $\boldsymbol{\Sigma}$ may not be available.

We try to solve the problem of the intractability of the continuous ridge leverage score density using DPPs and one of its variants, continuous volume sampling.

2 Kernel approximations using determinantal sampling

In Belhadji et al. (2019), we introduced a new class of quadratures based on projection DPPs. These quadratures are suited for functions living in a reproducing kernel Hilbert space. Compared to the work of Bach (2017), our quadrature is ridgeless ($\lambda = 0$), i.e. we solve the optimization problem

$$\min_{\boldsymbol{w} \in \mathbb{R}^N} \|\mu_g - \sum_{n \in [N]} w_n k(x_n, .)\|_{\mathcal{F}}^2 \tag{9}$$

The solution of (9) writes $\hat{\boldsymbol{w}} = \boldsymbol{K}(\boldsymbol{x})^{-1} \mu_g(\boldsymbol{x})$, and the optimal mixture $\sum_{n \in [N]} \hat{w}_n k(x_n, .)$ is the projection of μ_g on the subspace $\mathcal{T}(\boldsymbol{x}) = \text{Span}(k(x_n, .)_{n \in [N]})$ denoted $\boldsymbol{\Pi}_{\mathcal{T}(\boldsymbol{x})} \mu_g$, and the optimal value of (9) is called the interpolation error of μ_g. As for the nodes, we take a random set $\boldsymbol{x} = \{x_1, \dots, x_N\}$ that follows the distribution of a projection DPP that depends only on the first eigenfunctions of $\boldsymbol{\Sigma}$: this projection DPP provides an implementable alternative to the continuous ridge leverage scores distribution when the spectral decomposition of $\boldsymbol{\Sigma}$ is known.

We give a theoretical analysis of this class of quadratures and we show that the convergence rate depends on the eigenvalues $(\sigma_n)_{n \in \mathbb{N}^*}$ of the integration operator as $\mathcal{O}(N \sum_{n \geq N+1} \sigma_n)$, where N is the number of quadrature nodes. Numerical simulations hint that the rate of convergence actually scales as fast as $\mathcal{O}(\sigma_N)$. This is illustrated in Figure 1, where we compare the worst-case interpolation error $\sup_{\|g\|_\omega \leq 1} \|\mu_g - \boldsymbol{\Pi}_{\mathcal{T}(\boldsymbol{x})} \mu_g\|_{\mathcal{F}}^2$ of the following quadratures: 1) our quadrature (DPPKQ), 2) the quadrature based on continuous ridge leverage score sampling (LVSQ), 3) the quadrature based on uniform

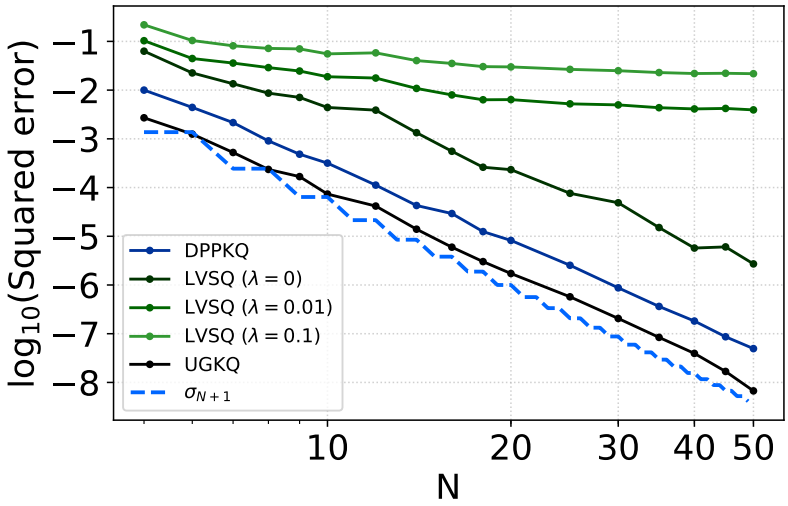


Figure 1: The worst-case interpolation error on the unit ball of $\mathbb{L}_2(\omega)$ under projection DPP (DPPKQ), continuous ridge leverage score density (LVSQ) and the uniform grid (UGKQ) compared to the eigenvalue of order σ_{N+1}. The RKHS is the periodic Sobolev space of order $s = 3$.

grid (UGKQ) and the weights that solve the optimization problem (9) in the case of the periodic Sobolev space of order $s = 3$.

In Belhadji et al. (2020), we continue the line of research initiated in Belhadji et al. (2019), and we study kernel quadrature and kernel interpolation based on nodes that follow the continuous volume sampling distribution, a mixture of DPPs. Contrary to the case of the projection DPP, we show that the expected value of the squared interpolation error has a closed formula under the distribution of continuous volume sampling

$$\mathbb{E}_{\mathrm{CVS}} \|\mu_g - \mathbf{\Pi}_{\mathcal{T}(\boldsymbol{x})}\mu_g\|_{\mathcal{F}}^2 = \sum_{m\in\mathbb{N}^*} \langle g, e_m\rangle_\omega^2 \epsilon_m(N), \tag{10}$$

where

$$\epsilon_m(N) = \sum_{\substack{|T|=N \\ m\notin T}} \prod_{t\in T} \sigma_t \Big/ \sum_{|T|=N} \prod_{t\in T} \sigma_t. \tag{11}$$

Moreover, we show that the $\epsilon_m(N)$ scales as σ_{N+1} for several RKHSs, and we give convergence guarantees for interpolation outside the scope of kernel quadrature.

The advantage of this distribution is that it is amenable to approximation using an MCMC algorithm that does not require the spectral decomposition of $\mathbf{\Sigma}$, unlike the projection DPP or the continuous ridge leverage score distribution.

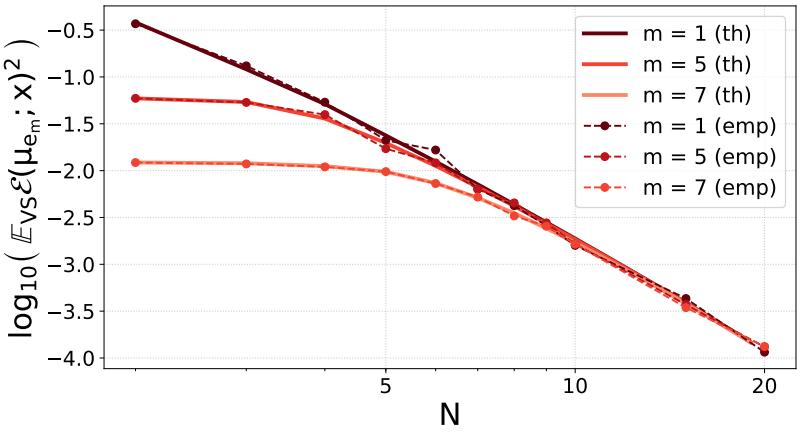


Figure 2: The theoretical values of the $\epsilon_m(N)$ compared to the empirical estimation using a fully kernelized MCMC. The RKHS is the periodic Sobolev space of order $s = 2$.

Figure 2 illustrates the comparison between the theoretical values of the $\epsilon_m(N)$ and their estimation using the MCMC sampler of Rezaei and Gharan (2019) in the case of the periodic Sobolev space of order $s = 2$.

Bibliographie

Bach, F. (2017). “On the equivalence between kernel quadrature rules and random feature expansions”. In: The Journal of Machine Learning Research 18.1, pp. 714–751.

Belhadji, A., R. Bardenet, and P. Chainais (2019). “Kernel quadrature with DPPs”. In: Advances in Neural Information Processing Systems 32, pp. 12907–12917.

Belhadji, A., R. Bardenet, and P. Chainais (2020). “Kernel interpolation with continuous volume sampling”. In: Proceedings of the 37th International Coference on International Conference on Machine Learning.

Rezaei, A. and S. O. Gharan (2019). “A Polynomial Time MCMC Method for Sampling from Continuous Determinantal Point Processes”. In: International Conference on Machine Learning, pp. 5438–5447.

Un multi-critère pour contrôler la multicolinéarité dans les modèles linéaires de régression multiple

Christian Derquenne

Electricité de France - Recherche et Développement - 7, boulevard Gaspard Monge - 91120 Palaiseau - christian.derquenne@edf.fr

Résumé. Cet article se place dans le cadre de la multicolinéarité entre les prédicteurs au sein d'un modèle linéaire de régression multiple. Le phénomène de multicolinéarité peut entrainer des incohérences sur les coefficients de régression et des oublis de prédicteurs, cela peut par conséquent poser des problèmes d'interprétation qui peuvent entrainer de mauvaises décisions. Le critère proposé revient à régulariser la matrice des corrélations simples entre les variables candidates à l'explication de manière à respecter conjointement la cohérence des signes entre les coefficients de régression multiple et de corrélation simple, ainsi que l'ordre de liaison entre les prédicteurs et la réponse, et les niveaux de significativité, tout en préservant la qualité de reconstitution des données à l'aide du modèle estimé.

Mots-clés. Modèle linéaire, Moindres Carrés Ordinaires, Multicolinéarité, régression linéaire régularisée, régression PLS

Abstract. This article is placed within the framework of the multicollinearity between the predictors within a linear model of multiple regression. The phenomenon of multicollinearity can lead to inconsistencies in the regression coefficients and omissions of predictors this can therefore pose problems of interpretation which can lead to poor decisions. The proposed criterion amounts to regularizing the matrix of simple correlations between the predictors so as to jointly respect the consistency of the signs between the coefficients of multiple regression and simple correlation, as well as the order of connection between the predictors and the response, and the levels of significance, while preserving the quality of reconstruction of the data using the estimated model.

Keywords. Linear model, Ordinary Least Squares, Multicollinearity, regularized linear regression, PLS regression

1 Contexte - objectif

Comme le disait Georges Box : "Tous les modèles sont faux, mais certains sont utiles" Cette phrase pleine de bon sens est à la base de cet article. En effet, l'utilisation d'un modèle peut être vue sous trois angles : l'interprétation, la prévision et la simulation. Le premier angle a pour objectif de répondre à la question suivante : "comment se forme le phénomène à expliquer ?". L'analyste s'attache tout d'abord aux sens des coefficients de régression du modèle estimé qui doivent être en adéquation avec les premières constatations qu'il a faites lors de la phase d'exploration des données (scatterplots, coefficients de corrélation linéaire simple). La force de liaison des prédicteurs avec la réponse est également un point primordial pour l'analyste. En d'autres termes, les niveaux des corrélations simples et multiples doivent être

proches. Cependant, ces propriétés requises peuvent ne pas être obtenues si la multicolinéarité entre les prédicteurs est trop élevée. L'interprétation de modèles est utilisée dans le domaine du marketing, par exemple quels leviers agissent sur la satisfaction, la fidélité à l'égard d'un produit ou d'un service ? Sous cet angle, l'analyste n'a pas pour objectif de prévoir au mieux le phénomène à expliquer. Par contre, le prévisionniste aura pour objectif de construire des modèles prévoyant la réponse avec la plus petite erreur possible (RMSE et MAPE, les plus faibles). L'interprétation de la forme et du sens des relations des prédicteurs ne sont pas privilégiées. Par exemple, la prévision de consommation d'électricité doit être la plus précise possible afin de respecter l'équilibre offre-demande. Enfin, la simulation se situe à la croisée des chemins de l'interprétation et de la prévision. En effet, les résultats de simulations peuvent être en entrée de modèles complexes ayant pour objectif de prévoir des phénomènes sur le long terme. Il est alors nécessaire que les simulations obtenues revêtent des propriétés d'explicativité du processus à générer et une qualité de reproduction efficace afin de ne pas diverger de scenarii vraisemblables du futur.

Dans cet article, nous avons choisi l'angle de l'interprétabilité des résultats de régression à destination des experts de domaines d'applications. Comme indiqué, La qualité de l'interprétation dépend du niveau de multicolinéarité. Les questions qui se posent en pratique sont : Doit-on sélectionner tous les prédicteurs ? Que faire si le commanditaire de l'étude désire garder toutes les variables "explicatives" ? Dans ce cas, comment les pondérer, sans altérer l'interprétation métier du modèle et la signification statistique ? Un certain nombre d'approches permettent de contourner ces problèmes, mais de façon partielle. Les méthodes de sélection de prédicteurs permettent de garder les variables les plus influentes dans le modèle, mais pas toujours de façon efficace face à la multicolinéarité. Le modèle choisi peut alors être inintéressant sous l'aspect métier, voire incohérent. Les méthodes de régularisation, telles que les régressions Ridge, Lars [Hoerl et al., 1970], Lasso [Tibshirani et al., 1996] ou ElasticNet [Zou et al., 2005] répondent en partie à la question de prédicteurs liés entre eux, mais elles exigent des paramètres non analytiques, donc difficiles à évaluer a priori. Le recours à la validation croisée est donc nécessaire. La régression sur composantes principales, qui consiste à réaliser une ACP sur toutes les variables explicatives, permet de garder au moyen de différentes stratégies certaines composantes principales qui résument au mieux les prédicteurs initiaux. Cependant, cette méthode n'est pas toujours optimale à l'égard du sens des coefficients et de l'importance des prédicteurs. La régression PLS (Partial Least Squares) [Wold H. et al., 1984] permet de mieux tenir compte de ces deux propriétés. En effet, cette méthode repose sur le principe de l'algorithme NIPALS (Non linear Iterative Partial Least Squares) qui est très puissant en termes de robustesse.

Nous proposons un critère permettant de maximixer les trois propriétés suivantes : reconstitution de l'inertie expliquée de l'estimateur MCO, adéquation des signes et de l'ordre des coefficients de régression multiple par rapport aux coefficients de corrélation linéaire simple associés. Une quatrième propriété peut également être recherchée : correspondance de la significativité des paramètres de régression multiple et simple. La première partie de cet article introduit le critère proposé, puis nous l'appliquons sur quelques exemples et le comparons aux méthodes évoquées précédemment. Nous concluons sur les apports et les faiblesses de l'approche proposée, et nous fournissons quelques voies futures en termes d'amélioration et de nouveaux développements.

2 Un multi-critère pour tenir compte de la multicolinéarité

La multicolinéarité entre les prédicteurs d'un modèle de régression multiple peut entrainer d'une part, des signes contraires des coefficients de régression par rapport aux corrélations linéaires simples et d'autre part, des résultats de tests marginaux des coefficients (test t) en contradiction avec ce qui aurait pu être attendu d'après les tests sur les coefficients de régression simple. Pour formaliser ces deux problèmes, prenons un exemple simple dans lequel il y a deux prédicteurs X_1 et X_2 et une réponse Y, avec le modèle de régression multiple : $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$. Le vecteur des coefficients $(\beta_0, \beta_1, \beta_2)$ avec X_1 et X_2 centrées sera :

$$(\bar{Y}, (r_{YX_1} - r_{X_1X_2} r_{YX_2})/(1 - r^2_{X_1X_2})(S_{X_1}/S_Y), (r_{YX_2} - r_{X_1X_2} r_{YX_1})/(1 - r^2_{X_1X_2})(S_{X_2}/S_Y)$$

où r_{YX_1}, r_{YX_2} et $r_{X_1X_2}$ sont les coefficients de corrélation simple pour les trois variables. Ce modèle peut fournir des signes contradictoires entre les coefficients de régression multiple et de régression simple. Considérons par exemple que les valeurs des corrélations soient $r_{YX_1} = 0,5$, $r_{YX_2} = 0,8$ et $r_{X_1X_2} = 0,9$ alors le numérateur de β_1 sera négatif, en contradiction avec le signe de r_{YX_1}. En effet, la corrélation entre les deux prédicteurs est plus élevée que le rapport entre r_{YX_1} et r_{YX_2}. Un deuxième problème peut subvenir telle que l'apparition de coefficients trop fortement élevés en valeur absolue. Cela est dû au dénominateur des coefficients : $1 - r^2_{X_1X_2}$ qui tendra d'autant plus vers 0 que $r_{X_1X_2}$ sera proche de l'unité en valeur absolue. Le dernier problème concerne les tests marginaux des coefficients de régression dont la significativité peut ne pas apparaitre, alors qu'elle était présente au sein des tests de corrélation linéaire simple. Cela est dû au phénomène d'inflation de la variance des coefficients de régression multiple. Dans notre exemple, la variance de $\sigma^2_{\hat{\beta}_j}$ associée à X_j, $(j = 1, 2)$ est de la forme $\hat{\sigma}^2_e/[(n-1)S_{X_j}(1 - r^2_{X_1X_2})]$. Plus la corrélation entre les deux prédicteurs sera élevée, plus la variance sera grande, plus la statistique de Student : $t_{\beta_j} = \beta_j/\sigma_{\hat{\beta}_j}$ sera faible, alors l'hypothèse de nullité du coefficient aura d'autant plus de chance d'être gardée. Un prédicteur risque donc d'être oublié.

Pour p prédicteurs le vecteur des coefficients de régression multiple est de la forme : $\hat{\beta} = S_Y \mathbf{D}^{-1/2} \Sigma^{-1} \mathbf{r}_{\mathbf{X}Y} = \Sigma^{-1} \hat{\beta}^{(s)}$, où $\mathbf{D}$, Σ^{-1}, $\mathbf{r}_{\mathbf{X}Y}$ et $\hat{\beta}^{(s)}$ sont respectivement la matrice diagonale des variances des prédicteurs, l'inverse de la matrice de corrélation des prédicteurs, le vecteur des coefficients de corrélation simple entre la réponse Y et chaque prédicteur et le vecteur des coefficients de régression simple associé. La matrice de covariance généralisée de $\hat{\beta}$ est : $\Sigma(\hat{\beta}) = \hat{\sigma}^2_e \mathbf{D}^{-1/2} \Sigma^{-1} \mathbf{D}^{-1/2}/(n-1)$. Plus Σ est chargée en fortes corrélations entre les prédicteurs, plus l'inverse de celle-ci influera sur l'estimateur des MCO et sur la matrice de covariance des coefficients. En effet, les éléments de celle-ci contient des sommes de produits de corrélations pouvant entrainer des poids d'autant plus importants qu'il y a présence de multicolinéarité.

Une solution naturelle pour pallier ce problème est d'affaiblir les corrélations dans la matrice Σ, comme le fait la régression PLS dans laquelle, la recherche de la première composante PLS passe par le remplacement pur et simple de toutes les corrélations par 0. Cela entraine par conséquent que le vecteur de coefficients obtenus soit égal à celui des coefficients de régression simple. Mais cela peut ne pas suffire en termes de variance expliquée, alors d'autres composantes PLS sont calculées et sélectionnées par validation croisée. Les coefficients de régression obtenus permettent généralement de préserver les signes et les forces des corrélations simples entre les prédicteurs et la réponse.

Le multi-critère proposé a pour objectif de maximiser : la reconstitution de l'inertie expliquée obtenue par le critère MCO, la préservation des signes des coefficients de régression multiple par rapport aux coefficients de régression simple, la préservation de l'ordre de la force de corrélation et du niveau de significativité. Pour cela, de façon formelle, le critère est le suivant :

$$\max_{\tilde{\beta}} \left[\left(\frac{\tilde{R}^2}{R^2_{MCO}} + \frac{\sum_{j=1}^{p} |r_{YX_j}| 1_{[r_{YX_j} \times \tilde{\beta}_j \geq 0]}}{\sum_{j=1}^{p} |r_{YX_j}|} + \tau_{Kendall}(\tilde{\beta}, \mathbf{r_{XY}}) + \sum_{j=1}^{p} \tilde{\pi}_j \right) \Big/ 4 \right] \quad (1)$$

où $\tilde{\beta}$, $\tilde{R}^2$ et R^2_{MCO} sont respectivement le vecteur de coefficients de régression multiple du multi-critère, le coefficient de détermination associé et celui du critère MCO, $\tau_{Kendall}$ est le coefficient de corrélation de Kendall mesurant l'adéquation entre les rangs de $\tilde{\beta}$ et de $\mathbf{r_{XY}}$. Enfin, $\tilde{\pi}_j = 1_{[pval(r_{YX_j}) \leq \alpha]} . 1_{[pval(\tilde{\beta}_j) \leq \alpha]} + 1_{[pval(r_{YX_j}) > \alpha]} . 1_{[pval(\tilde{\beta}_j) > \alpha]}$ mesure l'adéquation entre les tests significatifs et non significatifs communs des coefficients de régression multiple et les coefficients de corrélation simple. $pval(.)$ correspond à la p-valeur associée au test.

Le premier élément de (1) mesure la reconstitution de la qualité de l'inertie expliquée du modèle obtenu et celle du modèle MCO qui est optimale en termes de variance minimum, le deuxième terme correspond à l'adéquation des signes des coefficients de régression multiple et ceux des corrélations simples et de leurs niveaux. En effet, la valeur obtenue sera d'autant plus élevée qu'un signe bien reconstitué correspondra à une corrélation simple associée importante. Le troisième objet cherche à maximiser l'ordre entre les coefficients de régression multiple et les corrélations linéaires simples. Le quatrième terme évalue la cohérence d'acceptation ou de rejet entre les tests d'absence de liaison des coefficients du modèle de régression multiple et de non corrélation simple. Chacun de ces objets varie entre 0 et 1.

Comment est estimé $\tilde{\beta}$ ***?*** Comme indiqué précédemment, de fortes corrélations entre les prédicteurs entrainent la multicolinéarité. Afin de la réduire mécaniquement, la solution de la régression PLS pour le calcul de la première composante est simplement de multiplier par 0, l'ensemble des corrélations, sauf la diagonale. L'approche proposée dans cet article consiste à réduire les corrélations trop élevées afin de répondre au critère (1). Pour cela, celles-ci sont remplacées par un coefficient $\delta \in [0; 1]$ si elles sont supérieures à celui-ci, sinon la valeur de la corrélation est gardée telle quelle. De façon formelle, nous avons : $\Sigma = \begin{pmatrix} 1 & \mathbf{r_{X,X}} \\ \mathbf{r_{X,X}} & 1 \end{pmatrix}$, la matrice de corrélations usuelle, alors la matrice transformée est de la forme :

$$\tilde{\Sigma}_\delta = \begin{pmatrix} 1 & \mathbf{r_{X,X}} 1_{[\mathbf{r_{X,X}} \leq \delta]} \\ \mathbf{r_{X,X}} 1_{[\mathbf{r_{X,X}} \leq \delta]} & 1 \end{pmatrix} + \begin{pmatrix} 0 & \delta 1_{[\mathbf{r_{X,X}} > \delta]} \\ \delta 1_{[\mathbf{r_{X,X}} > \delta]} & 0 \end{pmatrix} \quad (2)$$

La procédure numérique d'estimation du vecteur $\tilde{\beta}$ consiste à faire varier δ dans $[0; 1]$ en calculant la forme usuelle du coefficient de régression, telle que $\tilde{\beta} = S_Y \mathbf{D}^{-1/2} \tilde{\Sigma}^{-1} \mathbf{r_{XY}}$ où Σ est remplacé par $\tilde{\Sigma}$. La valeur choisie pour $\tilde{\beta}$ est celle qui maximise le critère (1). Remarquons que si $\delta = 0$, alors $\tilde{\Sigma} = I$ et $\tilde{\beta} = \hat{\beta}^{PLS(1cp)}$ (1^{iere} composante PLS) ; si $\delta = 1$, alors $\tilde{\Sigma} = \Sigma$ et $\tilde{\beta} = \hat{\beta}^{MCO}$.

Terminons par l'évaluation d'autres critères (C_μ) par rapport au critère proposé dans cet article. Pour cela, nous calculons la valeur du multi-critère (1) sans le dernier élément, en remplaçant $\tilde{R}^2$ par $R^2_{C_\mu}$ et $\tilde{\beta}$ par $\hat{\beta}_\mu$. Cette évaluation est nommée $Eval$.

3 Application du multi-critère proposé et comparaisons

Afin d'évaluer la qualité de l'approche proposée et de la comparer à d'autres méthodes, nous avons simulé un jeu de données. Ses caractéristiques sont les suivantes : $X_1 \sim \mathcal{N}(0,1)$, $X_2 = X_1 + \mathcal{N}(0, 1.96)$, $X_3 = 2X_2 + 3 + \mathcal{N}(0, 0.04)$, $X_4 \sim \mathcal{N}(0, 2.25)$, $X_5 \sim \mathcal{N}(0,1)$, $X_6 = X_5 + \mathcal{N}(0, 0.04)$ et $Y = -1.5X_1 + 2X_2 + 0.5X_3 - 0.5X_4 + 0.1X_5 - 0.1X_6 + 4 + \mathcal{N}(0, 0.25)$. Nous avons appliqué, (i) le critère classique des MCO, (ii) la régression sur composantes principales avec une sélection pas à pas de celles-ci reposant sur des tests statistiques de nullité des coefficients entrant et sortant, (iii) la régression sur premières composantes principales (RFPC) fondées sur une étape préalable de classification des prédicteurs [Derquenne et al., 2002], (iv) la régression PLS, (v) la sélection pas à pas des prédicteurs initiaux, (vi), la sélection ascendante de prédicteurs, (vii) la méthode Stagewise, (viii) le critère Lars, (ix) le critère Lasso et (x) le critère ElasticNet.

La figure 1 montre l'évolution du multi-critère en orange, en fonction du coefficient δ variant dans $[0; 1]$, conjointement avec les quatre éléments de (1). Le rapport des R^2 est caractérisé en bleu, l'adéquation des signes des coefficients est matérialisée en rose, le τ de Kendall est dessiné en vert et la correspondance acceptation/rejet des tests est fournie en rouge. La valeur maximum observée du critère est de 0,9625 pour $\delta = 0,059$, matérialisée par le croisement des traits vertical et horizontal sur la figure 1. Nous pouvons constater que les 3 éléments sur 4 sont égaux à 1, le rapport des deux R^2 est égal à 0,8501, ce qui signifie que l'inertie expliquée associée au modèle du multi-critère reconstitue 85% de l'inertie expliquée du modèle MCO.

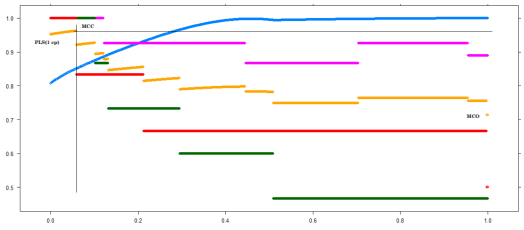


Figure 1: Evolution du critère (1)

La table 1 fournit les coefficients de régression standardisés pour chacune des méthodes, ainsi que les corrélations simples. Premièrement, seuls RFPC et le multi-critère fournissent des signes cohérents pour les six prédicteurs. Les forces de liaison sont relativement bien respectées pour la sélection pas à pas par p-valeurs, la régression PLS, avec un petit avantage pour le critère proposé. Seuls RFPC et le critère proposé sont en parfaite adéquation pour l'ordre des coefficients de régression et des corrélations simples. Globalement, les R^2 ajustés pour la plupart des méthodes sont supérieurs à 0,96, sauf pour RFPC (=0,66) et le multi-critère (=0,83). Ce dernier résultat est logique car notre critère n'optimise pas seulement le critère des MCO. Enfin, $Eval$ montre que le critère obtient la plus grande valeur ($Eval$=0,95), puis viennent Lasso, Lars, RFPC, ElasticNet, la sélection ascendante, la régression sur composantes principales et PLS ($Eval \in [0,83; 0,85]$), enfin, MCO et les deux méthodes de régression pas à pas ($Eval \in [0,73; 0,79]$). Par conséquent, seul le critère proposé parait efficace face à la multicolinéarité en termes de reconstitution de force et de sens des liaisons, tout en préservant la qualité du modèle.

	$\hat{\beta}^{MCO}$	$\hat{\beta}^{PCR_1}$	$\hat{\beta}^{RFPC}$	$\hat{\beta}^{PLS}$	$\hat{\beta}^{Spval}$	$\hat{\beta}^{Forw}$	$\hat{\beta}^{Swise}$	$\hat{\beta}^{Lars}$	$\hat{\beta}^{Lasso}$	$\hat{\beta}^{ENet}_{\lambda=0,5}$	$\hat{\beta}^{New}$	r_{y_X}
X_1	-0.379	-0.381	0.246	-0.323	-0.383	-0.377	-0.380	-0.263	-0.263	-0.209	0.027	0.171
X_2	0.708	0.560	0.332	0.549	0.695	0.265	0.000	0.393	0.393	0.589	0.845	0.910
X_3	0.419	0.569	0.332	0.557	0.435	0.861	1.130	0.608	0.608	0.594	0.849	0.913
X_4	-0.195	-0.194	-0.204	-0.263	-0.194	-0.191	-0.189	-0.125	-0.125	-0.167	-0.259	-0.084
X_5	0.099	0.064	0.041	0.051	0.051	0.053	0.000	0.000	0.000	0.071	0.026	0.136
X_6	0.050	0.013	-0.041	0.021	0.000	0.000	0.000	0.000	0.000	0.000	-0.013	-0.083
R^2_{adj}	0.987	0.987	0.663	0.980	0.987	0.990	0.981	0.962	0.962	0.953	0.829	n.a
$Eval$	0.786	0.830	0.850	0.827	0.774	0.839	0.727	0.854	0.854	0.830	0.950	1.000

Table 1: Coefficients de régression des méthodes

4 Apports, applications et voies futures

Le critère proposé permet de tenir compte de la multicolinéarité entre les prédicteurs dans un modèle linéaire de régression multiple. Il revient à régulariser la matrice des corrélations simples entre les variables candidates à l'explication de manière à respecter conjointement la cohérence des signes entre les coefficients de régression multiple et de corrélation simple, ainsi que l'ordre de liaison et les niveaux de significativité, tout en préservant la qualité de reconstitution des données à l'aide du modèle estimé. L'exemple simulé dans l'application fait apparaitre une très bonne efficacité par rapport aux méthodes concurrentes. Le critère introduit dans cet article a été évalué sur différents cas simulés (pas de corrélation entre prédicteurs, corrélations modérées, très fortes corrélations), ainsi que sur des exemples issus de la littérature, et des cas réels d'applications dans le domaine du marketing et du management de l'énergie. Les voies futures vont consister dans un premier temps à établir les propriétés mathématiques du critère proposé, puis de l'étendre à la sélection de prédicteurs afin d'obtenir des modèles parcimonieux tenant compte de façon robuste de la multicolinéarité, ce qui n'est pas toujours le cas de la plupart des méthodes de sélection pas à pas, voire même de méthodes de régularisation.

Bibliographie

[1] Derquenne Ch. et Hallais C., (2003), Une méthode alternative à l'approche PLS : comparaison et application aux modèles conceptuels marketing, *Revue de Statistique Appliquée*, **52**, 37-72.
[2] Hoerl, A.E. & Kennard R.W., (1970), Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics 42 (1)*, 80-86.
[3] Tibshirani, R., (1996), Regression Shrinkage and Selection via the Lasso, *J. R. Statist. Soc. B*, **58**, *No.* 1, 267-288.
[4] Wold S., Ruhe A., Wold H. & Dunn III W.J., (1984), The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses, *SIAM J. Sci. Stat. Comput.*, **5**, n°3, 735-743.
[5] Zou H. & Hastie T., (2005), Regularization and Variable Selection via the Elastic Net, *J. R. Statist. Soc. B*, **67**, *No.* 2, 301-320.

High-dimensional Two-sample Rank Statistics as Scalar Performance Criteria

Myrto Limnios[1] & Stephan Clémençon[2] & Nicolas Vayatis[1]

[1] *Université Paris-Saclay, ENS Paris-Saclay, CNRS, Centre Borelli, 91190 Gif-sur-Yvette, France,*
myli@math.ku.dk,nicolas.vayatis@ens-paris-saclay.fr
[2] *Telecom Paris, LTCI, Institut Polytechnique de Paris, 19 place Marguerite Perey, Palaiseau, 91120, France,* stephan.clemencon@telecom-paris.fr

Résumé. Nous proposons une généralisation des statistiques de rang à deux échantillons en grande dimension. Cette approche s'inspire de modèles issus de l'apprentissage d'ordonnancement. Nous détaillons comment ces collections de statistiques peuvent être utilisées comme critères scalaires de performance et, en particulier, pour les modèles d'ordonnancement biparti.
Mots-clés. Ordonnancement biparti, minimisation du risque empirique, bornes de généralisation, processus de rang, théorie de l'apprentissage statistique, statistiques linéaires de rang à deux échantillons

Abstract. We propose a generalization of two-sample linear rank statistics for high-dimensional data. The approach is inspired by learning-to-rank models. We detail how these collections of statistics can be used as scalar performance criteria and, in particular, for bipartite ranking models.
Keywords. Bipartite ranking, empirical risk minimization, generalization bounds, rank process, statistical learning theory, two-sample linear rank statistics

1 Introduction

In the context of high-dimensional statistics, ordering observations valued in a multivariate feature space $\mathcal{Z}$, *e.g.* of $\mathbb{R}^d$ with $d \geq 2$, is far from being straightforward. This is particularly needed in learning-to-rank problems, aiming to learn an order from a set of observations according, *e.g.* to their relevance, to predict the rank of any new instance. The simplest model, known as bipartite ranking, has gained popularity thanks to its adaptability to numerous applications, ranging from information retrieval, recommendation systems, to biomedical studies. Ranking models induce a relation order in $\mathcal{Z}$ as follows. Given a set of observations with binary labels, referred to as either 'positive' or 'negative', the goal is to learn an optimal scoring (measurable) function $s : \mathcal{Z} \to]-\infty, \infty]$, such that the ranks of univariate mapping by $s(z)$ of the observations minimize a statistical loss. The resulting pairwise loss corresponds to the probability of misranking pairs

of observations drawn at random from each label by a scoring function. Refer to [10] for a comprehensive review and approaches. In fact, the empirical binary (0-1) risk can be formulated thanks to the classical two-sample linear rank statistic: the *Mann-Whitney-Wilcoxon* (MWW, or *ranksum*) statistic [12], as the obtained ranks are univariate.
In this line, we propose a generalization of linear rank statistics when based on two independent and multivariate samples, of unknown underlying distributions, in Section 2. In Section 4, we show how collections of such statistics are generic scalar performance criteria for measuring the dissimilarity between two distribution functions, when indexed by a class of scoring functions. Properties are outlined from our work in [6]. We lastly apply these results to bipartite ranking and we refer to [9] for the application to anomaly ranking, and to [2, 1] for homogeneity testing.

2 Two-sample linear rank statistics

This section introduces a nonparametric generalization of two-sample rank statistics for multivariate observations. Let two independent *r.v.* $\mathbf{X}$ and $\mathbf{Y}$ *resp.* drawn from G and H, defined on a probability space $(\Omega,\ \mathcal{A},\ \mathbb{P})$ and valued in the multivariate feature space $\mathcal{Z}$, *e.g.*, of $\mathbb{R}^d$ with $d \in \mathbb{N}^*$. We consider two independent samples as follows. Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$, with $n \in \mathbb{N}^*$, *i.i.d.* observations drawn from G, and $\mathbf{Y}_1, \ldots, \mathbf{Y}_m$, with $m \in \mathbb{N}^*$, *i.i.d.* drawn from H, such that $n/N \to p \in (0,1)$, with $N = n + m$. The parameter p is interpreted as the asymptotic proportion of the $\mathbf{X}$s among the pooled sample, such that the pooled sample has asymptotic mixture distribution equal to $F = pG + (1-p)H$.

We propose a generalization of rank statistics based on bipartite ranking approaches, see [3]. Precisely, the multivariate observations are analyzed through their univariate value obtained by a given scoring function $s(z)$ from a class $\mathcal{S} = \{s : \mathcal{Z} \to \mathbb{R} \cup \{+\infty\},\ s \text{ measurable}\}$. The ranks are defined for all $s \in \mathcal{S}$, by

$$\text{Rank}(t) = \sum_{i=1}^{n} \mathbb{I}\{s(\mathbf{X}_i) \leq t\} + \sum_{j=1}^{m} \mathbb{I}\{s(\mathbf{Y}_j) \leq t\}, \quad \text{for all} \quad t \in \mathbb{R}\ . \tag{2.1}$$

The distribution of $s(\mathbf{X})$ (*i.e.* the image of G by s) is denoted by G_s, that of $s(\mathbf{Y})$ (*i.e.* the image of H by s) by H_s. Their empirical versions are defined by $\widehat{G}_{s,n}(t) = (1/n)\sum_{i=1}^{n} \mathbb{I}\{s(\mathbf{X}_i) \leq t\}$ and $\widehat{H}_{s,m}(t) = (1/m)\sum_{j=1}^{m} \mathbb{I}\{s(\mathbf{Y}_j) \leq t\}$, for all $t \in \mathcal{Z}$. We denote by F_s the image of the mixture distribution F by s, and its estimator being $\widehat{F}_{s,N}(t) = (n/N)\widehat{G}_{s,n}(t) + (m/N)\widehat{H}_{s,m}(t)$. The ranks based on the pooled sample are proportional to the empirical mixture distribution, *i.e.*, $\text{Rank}(t) = N\widehat{F}_{s,N}(t)$, for all $t \in \mathbb{R}$. We define the generalization of linear rank statistics below.

Definition 1. (Two-sample linear rank statistics) *Let $\phi : [0,1] \to \mathbb{R}$ be a nondecreasing function. The two-sample linear rank statistics with 'score-generating function'*

$\phi(u)$ based on the random samples $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ and $\{\mathbf{Y}_1, \ldots, \mathbf{Y}_m\}$ are given by:

$$\widehat{W}^{\phi}_{n,m}(s) = \sum_{i=1}^{n} \phi\left(\frac{\mathrm{Rank}(s(\mathbf{X}_i))}{N+1}\right) . \tag{2.2}$$

Example 1. (Two-sample hypothesis testing.) *The proposed formulation recovers classic two-sample univariate test statistics, depending on the scores of the statistics of Eq. (2.2). For instance, $\phi(u) = 2\sqrt{3}(u-1/2)$ recovers the logistic statistic, $\phi(u) = -\log(1-u)$ the logrank statistic, $\phi(u) = \mathrm{sgn}(u-1/2)$ the median one, and $\phi(u) = \Phi^{-1}(u)$, with Φ the normal quantile function, leads to van der Waerden test. Refer to [8] for additional properties and examples. In particular, choosing $\phi(u) = u$ coincides with the classical Mann-Whitney-Wilcoxon (MWW) statistic [12], counting the ranks of the $\mathbf{X}$s among the pooled sample $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\} \cup \{\mathbf{Y}_1, \ldots, \mathbf{Y}_m\}$, as follows*

$$\widehat{W}_{n,m}(s) = \sum_{i=1}^{n} \mathrm{Rank}(s(\mathbf{X}_i)) . \tag{2.3}$$

In the following section, we consider collections of these statistics as scalar performance criteria for ranking models, when indexed by classes of scoring functions.

3 Rank statistics as scalar performance criteria

We consider the statistics (Eq. (2.2)) introduced in Section 2. They are empirical counterparts of the functional W_ϕ-performance measure defined below. This section details their optimal elements and theoretical guarantees. Refer to [6] for detailed results and proofs.

Definition 2. *Let $p \in (0,1)$, two c.d.f. G and H, and $\mathcal{S}_0 \subset \mathcal{S}$ a class of scoring functions. For a given score-generating function ϕ, for all $s \in \mathcal{S}_0$, the functional*

$$W_\phi(s) = \mathbb{E}[(\phi \circ F_s)(s(\mathbf{X}))] , \tag{3.1}$$

is referred to as the "W_ϕ-ranking performance measure", where $F_s = pG_s + (1-p)H_s$.

Let $N \geq 1/p$, we suppose that $n = \lfloor pN \rfloor$ and $m = \lceil (1-p)N \rceil = N - n$. From the R-statistic based on two independent *i.i.d.* samples $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ and $\{\mathbf{Y}_1, \ldots, \mathbf{Y}_m\}$, the W_ϕ-functional is obtained by replacing $\widehat{F}_{s,N}(s(\mathbf{X}_i))$ by $F_s(s(\mathbf{X}_i))$ in Eq. (2.2), and then taking the expectation. In [5], Proposition 2 proves that the optimal scoring functions maximizes the W_ϕ-performance measure. They are given by nondecreasing transforms of the likelihood ratio $\Psi(z) = dG/dH(z)$, *i.e.*,

$$\mathcal{S}^* = \{s \in \mathcal{S} \;\; s.t. \text{ for all } (z,\, z') \text{ in } \mathcal{Z} : \;\; \Psi(z) < \Psi(z') \Rightarrow s^*(z) < s^*(z')\} . \tag{3.2}$$

The following paragraph presents generalization guarantees for maximizers of the empirical counterpart $\widehat{W}^{\phi}_{n,m}(s)$, proved and extensively detailed in [6].

Performance of maximizers of two-sample rank processes. Let a class of scoring functions $\mathcal{S}_0 \subset \mathcal{S}$ and consider any maximizer of the empirical W_ϕ-ranking performance measure over that class, as follows

$$\widehat{s} \in \arg\max_{s \in \mathcal{S}_0} \widehat{W}^{\phi}_{n,m}(s) \ . \tag{3.3}$$

Such a scoring function aims at maximizing the ranks of the $\mathbf{X}$s ('positive' instances) *w.r.t.* the pooled sample. The generalization error of $\hat{s}$ is upperbounded, similarly to classical ERM formulations, by

$$W^*_\phi - W_\phi(\widehat{s}) \leq 2 \sup_{s \in \mathcal{S}_0} \left| \frac{1}{n} \widehat{W}^{\phi}_{n,m}(s) - W_\phi(s) \right| + \left(W^*_\phi - \sup_{s \in \mathcal{S}_0} W_\phi(s) \right) \ , \tag{3.4}$$

where $W^*_\phi := W_\phi(s)$, for all $s \in \mathcal{S}^*$. Classical ERM methods for M-estimators are not applicable here, due to the complex structure of the statistic (sum of correlated and non-*i.i.d.* terms, see Def. 1). In [6], the collection of statistics $\{\widehat{W}^{\phi}_{n,m}(s)\}_{s \in \mathcal{S}_0}$ was linearly decomposed in empirical processes plus a remainder term of higher order, for which a nonasymptotic uniform control over $\mathcal{S}_0$ was proved, see Proposition 4 therein. The required hypothesis are the following.

(H1) Let $M > 0$. For all $s \in \mathcal{S}_0$, the random variables $s(\mathbf{X})$ and $s(\mathbf{Y})$ are continuous, with density functions that are twice differentiable and have Sobolev $\mathcal{W}^{2,\infty}$-norms bounded by $M < +\infty$.

(H2) The score-generating function $\phi : [0,1] \mapsto \mathbb{R}$, is nondecreasing and twice continuously differentiable.

(H3) The class of scoring functions $\mathcal{S}_0$ is a VC class of finite VC dimension $\mathcal{V} < +\infty$.

We state the generalization bound proved in Corollary 7 ([6]), of order $\mathcal{O}(N^{-1/2})$. A series of similar classical analysis can be found in [6], *e.g.*, model selection, smooth counterparts, wherein the constants depend only on $\mathcal{S}_0$, ϕ and p.

Theorem 3 (Corollary 7, [6])**.** *Let $\hat{s}$ be an empirical W_ϕ-ranking performance maximizer over the class $\mathcal{S}_0$, i.e., solution of Eq. (3.3). Under Assumptions (H1,2,3), for any $\delta \in (0,1)$, there exist constants C_1, $C_3 > 0$, $C_2 \geq 24$ depending on ϕ, K, R, $\mathcal{V}$, $C_4 \geq C_1$, and $C_5 > 0$ is a constant depending on ϕ, K and R, such that we have with probability at least $1 - \delta$:*

$$W^*_\phi - W_\phi(\hat{s}) \leq 2\sqrt{\frac{\log(C_2/\delta)}{pC_3N}} + C_5 h^2 + \left\{ W^*_\phi - \sup_{s \in \mathcal{S}_0} W_\phi(s) \right\}, \tag{3.5}$$

as soon as $N \geq 1/(p \min(p, 1-p)^2 C_3 C_4^2) \log(C_2/\delta)$ and $\delta \leq C_2 e^{-C_1^2 C_3}$.

4 Kernel regularization for the maximization of ranking performance measures

The optimization problem for learning the empirical function $\widehat{s}$ face regularity issues due to the non-smoothness of the R-statistic $\widehat{W}^{\phi}_{n,m}$, inherited from the empirical *c.d.f.* $\widehat{F}_{s,N}$. We propose a kernel-smoothed version theoretical criterion (3.1). Consider $\kappa(t) = \int_{-\infty}^{t} K(u)du$ with $K : \mathbb{R} \to \mathbb{R}$ a second-order Parzen-Rosenblatt kernel, then, for any $s \in \mathcal{S}_0$, for any kernel bandwidth $h > 0$ and all $t \in \mathbb{R}$,

$$\widetilde{W}_{\phi,h}(s) = \mathbb{E}[(\phi \circ \widetilde{F}_{s,h})(s(\mathbf{X}))] \ , \tag{4.1}$$

where the surrogate *c.d.f.* is defined by $\widetilde{F}_{s,h}(t) = \int_{\mathbb{R}} \kappa\left((t-u)/h\right) F_s(du)$. In [6], an empirical version of the $\widetilde{W}_{\phi,h}$-measure is considered, when based on the two samples $\{\mathbf{X}_1, \ \ldots, \ \mathbf{X}_n\}$ and $\{\mathbf{Y}_1, \ \ldots, \ \mathbf{Y}_m\}$, and under regularity assumptions on K and F, for all $s \in \mathcal{S}$,

$$\frac{1}{n}\widehat{W}^{\phi}_{n,m,h}(s) = \frac{1}{n}\sum_{i=1}^{n}(\phi \circ \widehat{F}_{s,N,h})(s(\mathbf{X}_i)) \ . \tag{4.2}$$

where $\widehat{F}_{s,N,h}(t) = (1/N)\sum_{i=1}^{n} \kappa\left((t - s(\mathbf{X}_i))/h\right) + (1/N)\sum_{j=1}^{m} \kappa\left((t - s(\mathbf{Y}_j))/h\right)$. A similar generalization bound to Theorem 3 with an additional bias term of order $\mathcal{O}(h^2)$ is proved, see Proposition 9 in [6].

Application to bipartite ranking. This paragraph relates the W_{ϕ}-functional measure to bipartite ranking models when used as scalar performance criterion. Tailoring the ranks by the choice of score-generating function leads to classical losses. For instance, choosing $\phi(u) = u$ recovers the classical binary (0-1) loss, [4] considered $\phi(u) = u\mathbb{I}\{u \geq u_0\}$, with $u_0 \in (0, 1)$, for considering only the $1 - u_0$ 'best observations' *i.e.* the ones having highest ranks. Refer also to classical criteria in ranking problems in [11, 7]. A simple gradient ascent algorithm for learning the optimal empirical function $\widehat{s}$ in Eq. (3.3) can be implemented for the maximization of the smoothed criterion Eq. (4.2). This is possible when the class of scoring functions $\mathcal{S}_0$ is parametrized by a set Θ. For instance, for multivariate Gaussian samples valued in $\mathcal{Z} = \mathbb{R}^d$, $d \geq 2$, the below probabilistic models recover the oracle class of scoring functions.

- $\Theta = \mathbb{R}^d$ and $\mathcal{S}_0 = \{s_\theta(z) = \langle z, \theta \rangle, \quad \text{for all} \quad z \in \mathcal{Z}, \quad \theta \in \Theta\}$, when the two samples differ only through their mean, *i.e.*, $\mathbf{X} \sim \mathcal{N}_d(\mu_X, \Sigma)$ and $\mathbf{Y} \sim \mathcal{N}_d(\mu_Y, \Sigma)$ where $\Sigma \in S_d^+(\mathbb{R})$, and $(\mu_X, \ \mu_Y) \in \mathcal{Z}$,
- $\Theta = S_d^+(\mathbb{R})$ and $\mathcal{S}_0 = \{s_\theta(z) = \langle z, \theta^{-1} z \rangle, \quad \text{for all} \quad z \in \mathcal{Z}, \quad \theta \in \Theta\}$, when the two samples differ only through their covariance matrix, *i.e.*, $\mathbf{X} \sim \mathcal{N}_d(\mu, \Sigma_X)$ and $\mathbf{Y} \sim \mathcal{N}_d(\mu, \Sigma_Y)$, with $(\Sigma_X, \ \Sigma_Y) \in S_d^+(\mathbb{R})$, and $\mu \in \mathcal{Z}$.

Numerical results are detailed in [6], where we compared the efficiency of the ranking performance criteria depending on the choice of ϕ.

References

[1] I. Bargiotas, A. Kalogeratos, M. Limnios, P-P. Vidal, D. Ricard, and N. Vayatis. Revealing posturographic profile of patients with Parkinsonian syndromes through a novel hypothesis testing framework based on machine learning. *PLOS ONE*, 16(2):1–22, 02 2021.

[2] S. Clémençon, M. Depecker, and N. Vayatis. AUC maximization and the two-sample problem. In *Advances in Neural Information Processing Systems*, volume 3559 of *Lecture Notes in Computer Science*, pages 1–15. Springer, 2009.

[3] S. Clémençon, G. Lugosi, and N. Vayatis. Ranking and empirical risk minimization of U-statistics. *The Annals of Statistics*, 36(2):844–874, 2008.

[4] S. Clémençon and N. Vayatis. Ranking the best instances. *Journal of Machine Learning Research*, 8:2671–2699, 2007.

[5] S. Clémençon and N. Vayatis. Tree-based ranking methods. *IEEE Transactions on Information Theory*, 55(9):4316–4336, 2009.

[6] S. Clémençon, M. Limnios, and N. Vayatis. Concentration inequalities for two-sample rank processes with application to bipartite ranking. *Electronic Journal of Statistics*, 15(2):4659 – 4717, 2021.

[7] D. Cossock and T. Zhang. Subset ranking using regression. In *Proceedings of COLT 2006*, volume 4005 of *Lecture Notes in Computer Science*, pages 605–619, 2006.

[8] E.L. Lehmann and J. P. Romano. *Testing Statistical Hypotheses*. Springer, 2005.

[9] M. Limnios, N. Noiry, and S. Clémençon. Learning to rank anomalies: Scalar performance criteria and maximization of two-sample rank statistics. In *Proceedings of the Third International Workshop on Learning with Imbalanced Domains: Theory and Applications*, volume 154 of *Proceedings of Machine Learning Research*, pages 63–75, 2021.

[10] A.K. Menon and R.C. Williamson. Bipartite ranking: A risk theoretic perspective. *Journal of Machine Learning Research*, 7:1–102, 2016.

[11] C. Rudin. Ranking with a P-Norm Push. In *Proceedings of COLT 2006*, volume 4005 of *Lecture Notes in Computer Science*, pages 589–604, 2006.

[12] F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics*, 1:80–83, 1945.

Clustering de modalités en régression logistique

Slimane Makhlouf [1], François-Xavier Jollois[2], and Avner Bar-Hen[3]

[1]CEDRIC. Méthodes statistiques de data-mining et apprentissage – Centre d´tudes et de recherche en informatique et communications – France
[2]Paris Descartes – Université Paris V - Paris Descartes – France
[3]Conservatoire National des Arts et Métiers [CNAM] – Conservatoire National des Arts et Métiers (CNAM), Conservatoire National des Arts et Métiers [CNAM] – France

Résumé

Ce travail s'intéresse à la réduction de dimension de données catégorielles par regroupement de modalités, dans le cadre d'une régression logistique. L'inflation du nombre de modalités augmente rapidement le coût de calcul et dégrade fortement les capacités prédictives des modèles. La réduction de dimensionalité peut s'effectuer par sélection et/ou par regroupement de variables (clustering). L'interprétation des groupes ainsi constitués peut apporter des informations supplémentaires sur la structure des données étudiées. De nombreux travaux de recherche se concentrent sur la sélection ou le regroupement de variables quantitatives. Nous présentons dans cet article nos travaux sur le regroupement de modalités à l'intérieur de variables catégorielles et tout particulièrement en grande dimension. Nos résultats sont appliqués sur un jeu d'enchères en ligne.

PLENIERE : Clémentine Prieur

(Non)linear dimension reduction of input parameter space using gradient information

Clémentine Prieur

Univ. Grenoble

Many problems that arise in uncertainty quantification, e.g., integrating or approximating multivariate functions, suffer from the curse of dimensionality. The cost of computing a sufficiently accurate approximation grows indeed dramatically with the dimension of input parameter space. It thus seems important to identify and exploit some notion of low-dimensional structure as, e.g., the intrinsic dimension of the model. A function varying primarily along a a low dimensional manifold embedded in the high-dimensional input parameter space is said of low intrinsic dimension. In that setting, algorithms for quantifying uncertainty focusing on the most relevant features of input parameter space are expected to reduce the overall cost. Our presentation goes from global sensitivity analysis to (non)linear gradient-based dimension reduction, generalizing the active subspace methodology.

PLENIERE ANNULÉE : Chloé-Agathe Azencott – ANNULEE

Structured feature selection in high-dimensional genomic data

Chloé-Agathe Azencott

Cbio Mines Paris Tech

Many problems in genomics require the ability to identify relevant features in data sets containing many more orders of magnitude than samples. One such example is genome-wide association studies (GWAS), in which hundreds of thousands of variants are measured for orders of magnitude fewer samples. Even for the most classical approaches, where one tests for the association of each variant with the phenotype, these studies are severely underpowered. Accounting for multiple effects of several variants is even more challenging. This talk will describe several approaches that alleviate this difficulty by incorporating prior knowledge (such as biological networks, population membership, or linkage disequilibrium) as structure on the data.

PLENIERE ANNULEE : Nicolas Papadakis

Gradient Step Denoiser for convergent Plug-and-Play

Nicolas Papadakis

CNRS Bordeaux

In image sciences, Plug-and-Play methods constitute a class of iterative algorithms for solving Bayesian inverse problems where regularization is performed by an off-the-shelf denoiser. Although Plug-and-Play methods can lead to tremendous visual performance for various image problems, the few existing convergence guarantees are based on unrealistic (or suboptimal) hypotheses on the denoiser, or limited to strongly convex data terms. In this work, we propose a new type of Plug-and-Play methods, based on half-quadratic splitting, for which the denoiser is realized as a gradient descent step on a functional parameterized by a deep neural network. Exploiting convergence results for proximal gradient descent algorithms in the non-convex setting, we show that the proposed Plug-and-Play algorithm is a convergent iterative scheme that targets stationary points of an explicit global functional. Besides, experiments show that it is possible to learn such a deep denoiser while not compromising the performance in comparison to other state-of-the-art deep denoisers used in Plug-and-Play schemes. We apply our proximal gradient algorithm to various ill-posed inverse problems, e.g. deblurring, super-resolution and inpainting. For all these applications, numerical results empirically confirm the convergence results. Experiments also show that this new algorithm reaches state-of-the-art performance, both quantitatively and qualitatively.

SFdS - Fiabilité

Degradation Model Selection Using Depth Functions

Mitra Fouladirad [1] & Diego Rodolpho Tomasi [2] & Arefe Asadi [2]

[1] *mitra.fouladirad@centrale-marseille.fr*
[2] *diego.tomassi@biofortis.fr*
[3] *arefe.asadi@utt.fr*

Abstract.
For lifetime prediction or maintenance planning of complex systems, degradation modeling is essential. The reason is that for highly reliable systems that failure times are difficult to observe, degradation measurements often provide more information than failure time to improve system reliability [1]. According to Lehmann [2], the stochastic-process-based models shows great flexibility in describing the failure mechanisms caused by degradation.

The aim of degradation modeling is to select a model from a set of competing models capturing the features of the underlying degradation process. An efficient statistical tool is able to discard irrelevant models.

The concept of statistical depth could be considered as a statistical tool for model suggestion. Tukey [3] introduced a data depth to extend the notion of median to multi-variate random variables. Considering functional data, the data which are recorded densely over time with one observed function per subject [4], the notion of depth has been extended by [5, 6, 7]. An alternative point of view based on the graphic representation of curves is proposed in [8].

A depth function reflects the centrality of the observation to a statistical population [9]. The models that show high values of depth function are compared based on different statistical criteria, namely Failure Time Distribution, and the best model is selected to predict failure time for the system under study.

Keywords. Degradation modeling, failure time, stochastic processes, reliability, depth function.

References

[1] W. Q. Meeker and L. A. Escobar, *Statistical methods for reliability data.* John Wiley & Sons, 2014.

[2] A. Lehmann, "Joint modeling of degradation and failure time data," *Journal of Statistical Planning and Inference*, vol. 139, no. 5, pp. 1693–1706, 2009. Special Issue on Degradation, Damage, Fatigue and Accelerated Life Models in Reliability Testing.

[3] J. W. Tukey, "Mathematics and the picturing of data," in *Proceedings of the International Congress of Mathematicians, Vancouver, 1975*, vol. 2, pp. 523–531, 1975.

[4] P. Hall, H.-G. Müller, and J.-L. Wang, "Properties of principal component methods for functional and longitudinal data analysis," *The annals of statistics*, pp. 1493–1517, 2006.

[5] R. Fraiman and G. Muniz, "Trimmed means for functional data," *Test*, vol. 10, no. 2, pp. 419–440, 2001.

[6] A. Cuevas, M. Febrero, and R. Fraiman, "On the use of the bootstrap for estimating functions with functional data," *Computational statistics & data analysis*, vol. 51, no. 2, pp. 1063–1074, 2006.

[7] A. Cuevas, M. Febrero, and R. Fraiman, "Robust estimation and classification for functional data via projection-based depth notions," *Computational Statistics*, vol. 22, no. 3, pp. 481–496, 2007.

[8] S. López-Pintado and J. Romo, "On the concept of depth for functional data," *Journal of the American statistical Association*, vol. 104, no. 486, pp. 718–734, 2009.

[9] G. Staerman, P. Mozharovskyi, S. Clémen, *et al.*, "The area of the convex hull of sampled curves: a robust functional statistical depth measure," in *International Conference on Artificial Intelligence and Statistics*, pp. 570–579, PMLR, 2020.

ON MISSPECIFICATION OF A GAMMA WITH AN INVERSE GAUSSIAN PROCESS AND ITS EFFECTS IN TERMS OF PROGNOSTIC AND RELATED DECISION PROCESSES

Nicola Esposito[1], Bruno Castanier[2], Massimiliano Giorgio[3]

[1] *Université d'Angers/Laris – Università di Napoli Federico II. Email: nicola.esposito@etud.univ-angers.fr*
[2] *Université d'Angers/Laris. Email: bruno.castanier@univ-angers.fr*
[3]*Università di Napoli Federico II. Email: massimiliano.giorgio@unina.it*

Keywords. Gamma process, misspecification issue, prognostic index, maintenance planning.

The gamma and inverse Gaussian processes are two common choices of models for increasing degradation phenomena in reliability field (see Van Nortwijk (2009) and Ye and Chen (2014)). These models have similar features. In fact, both have independent increments and are especially suitable to describe degradation phenomena that take place gradually over time in a sequence of tiny increments, such as crack growth or corrosion. Factually, there are many experimental situations where these models show comparable fitting ability. Hence, based on all these reasons, in the current literature, they are often treated as equivalent to each other. Nonetheless, this is not true. This circumstance makes the misspecification issue a problem of concern.

In fact, in this paper we focus on this misspecification issue and on the effect of a misspecification on the estimates of the remaining useful lifetime and on the long-run average maintenance cost rate, computed under a maintenance policy recently proposed in the literature.

Model selection is performed by using the Akaike Information Criterion (AIC). It is assumed that a misspecification occurs when the AIC leads to select the wrong model.

This study extends Esposito *et al.* (2021) and Esposito *et al.* (2022), where it is examined the misspecification of a gamma process with an inverse Gaussian, by considering the symmetric case where the true model is the inverse Gaussian process.

The new results give evidence and confirm that, under both settings, when the model used to analyze the available data is selected by using the Akaike information criterion, the (possible) misspecification of one of the considered competing model with the other one produces a negligible impact both in terms of remaining useful life estimation and in terms of maintenance costs. Conversely, it was observed that the consequences could be severe if the model is selected without adopting a formal inferential procedure.

Bibliography

Van Noortwijk J. M. (2009). *A survey of the application of gamma processes in maintenance*. Reliability Engineering & System Safety, vol. 94, no. 1, pp. 2-21.
Ye Z.-S. and Chen, N. (2014). *The inverse Gaussian process as a degradation model*. Technometrics, vol. 56, no. 3, pp. 302-311.
Esposito N., Mele A., Castanier B., and Giorgio M. (2021). *Misspecification analysis of a gamma-with an inverse Gaussian-based perturbed degradation model by using a new expectation-*

maximization particle filter algorithm. In Proceedings of the 31st European Safety and Reliability Conference (ESREL2021), Research Publishing, Singapore.
Esposito N., Castanier B., and Giorgio M. (2022). *Impacts on performances of a condition-based maintenance policy of misspecification of gamma with inverse Gaussian degradation processes*.
by using a new expectation-maximization particle filter algorithm. Submitted to 8th International Symposium on Reliability Engineering and Risk Management (ISRERM), Hannover.

Sequential Interval Reliability for semi-Markov systems

Vlad Stefan Barbu [1] & Guglielmo D'Amico [2] & Thomas Gkelsinis [3]

[1] *LMRS UMR 6085, Université de Rouen Normandie, Avenue de l'Université, BP.12, F76801 Saint-Étienne-du-Rouvray, France; barbu@univ-rouen.fr*
[2] *Department of Economics, University G. d'Annunzio of Chieti-Pescara, Italy; g.damico@unich.it*
[3] *LMRS UMR 6085, Université de Rouen Normandie, Avenue de l'Université, BP.12, F76801 Saint-Étienne-du-Rouvray, France; thomas.gkelsinis@univ-rouen.fr*

Résumé. Le travail que nous présentons, basé sur Barbu et al. (2021), concerne la modélisation de la fiabilité des systèmes multi-états avec dépendance temporelle. Plus précisément, nous proposons une généralisation des mesures de fiabilité, comme la fonction de fiabilité, la disponibilité ou bien la fiabilité sur un intervalle, pour des systèmes réparables semi-markoviens homogènes à temps discret. Cette nouvelle mesure, appelée la fiabilité séquentielle sur intervalles, est définie comme la probabilité que le système fonctionne sur une séquence des intervalles qui ne se chevauchent pas.

Mots-clés. processus semi-markoviens, fiabilité, analyse transitoire

Abstract. The present work, based on Barbu et al. (2021), is dealing with reliability modelling for multi-state systems with time dependence. We propose a generalization of reliability indexes such as reliability, interval reliability and availability for homogeneous semi-Markov repairable systems in discrete time. That measure, called sequential interval reliability, is the probability that the system works in a sequence of non-overlapping time intervals.

Keywords. semi-Markov processes, reliability, transient analysis

1 Discrete-time semi-Markov processes and reliability measures

Let us consider a random system with finite state space $E = \{1, \ldots, s\}$, $s < \infty$. Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and assume that the time evolution of the system is governed by a stochastic process $Z = (Z_k)_{k\in\mathbb{N}}$, defined on $(\Omega, \mathcal{A}, \mathbb{P})$ with values in E; that is to say that Z_k gives the state of the system at time k. Let $T = (T_n)_{n\in\mathbb{N}}$, defined on $(\Omega, \mathcal{A}, \mathbb{P})$ with values in $\mathbb{Z}$, be the successive time points when state changes in $(Z_k)_{k\in\mathbb{N}}$ occur (the jump times) and let also $J = (J_n)_{n\in\mathbb{N}}$, defined on $(\Omega, \mathcal{A}, \mathbb{P})$ with values in E, be the successively visited states at these time points. Set also $X = (X_n)_{n\in\mathbb{N}^*}$ for the

successive sojourn times in the visited states. Thus, $X_{n+1} = S_{n+1} - S_n$, $n \in \mathbb{N}$. The relation between the process Z and the process J of the successively visited states is given by $Z_k = J_{N(k)}$, or, equivalently, $J_n = Z_{T_n}, n, k \in \mathbb{N}$, where

$$N(k) := \max\{n \in \mathbb{N} \mid T_n \leq k\} \tag{1}$$

is the discrete-time counting process of the number of jumps in $[0, k] \subset \mathbb{N}$.

Definition 1 (semi-Markov chain SMC and Markov renewal chain MRC)**.** *If we have*

$$\mathbb{P}(J_{n+1} = j, T_{n+1} - T_n = k | J_n = i, J_{n-1}, \ldots, J_0, T_n, \ldots, T_0) = \mathbb{P}(J_{n+1} = j, T_{n+1} - T_n = k | J_n = i), \tag{2}$$

then $Z = (Z_k)_k$ is called a semi-Markov chain *(SMC) and $(J, T) = (J_n, T_n)_n$ is called a* Markov renewal chain *(MRC).*

All along this paper we assume that the MRC or SMC are homogeneous with respect to the time in the sense that Equation (2) is independent of n.

Note that if (J, T) is a MRC, then it can be proved that $J = (J_n)_{n\in\mathbb{N}}$ is a Markov chain with state space E, called the *embedded Markov chain* of the MRC (J, T) (or of the SMC Z).

Definition 2. *For a semi-Markov chain we define:*

- *the* semi-Markov kernel $(q_{ij}(k))_{i,j\in E,k\in\mathbb{N}}$, $q_{ij}(k) = \mathbb{P}(J_{n+1} = j, T_{n+1} - T_n = k | J_n = i)$;
- *the* initial distribution $(\mu_i)_{i\in E}$, $\mu_i = \mathbb{P}(J_0 = i) = \mathbb{P}(Z_0 = i)$;
- *the* transition matrix $(p_{ij})_{i,j\in E}$ *of the embedded Markov chain* $J = (J_n)_n$, $p_{ij} = \mathbb{P}(J_{n+1} = j | J_n = i)$;
- *the* conditional sojourn time distribution $(f_{ij}(k))_{i,j\in E,k\in\mathbb{N}}$, $f_{ij}(k) = \mathbb{P}(T_{n+1} - T_n = k | J_n = i, J_{n+1} = j)$;
- *the* sojourn time distribution in a state $(h_i(k))_{i\in E,k\in\mathbb{N}}$, $h_i(k) = \mathbb{P}(T_{n+1} - T_n = k | J_n = i)$.

Note that

$$q_{ij}(k) = p_{ij} f_{ij}(k).$$

Clearly, a semi-Markov chain is uniquely determined a.s. by an initial distribution $(\mu_i)_{i\in E}$ and a semi-Markov kernel $(q_{ij}(k))_{i,j\in E,k\in\mathbb{N}}$ or, equivalently, by an initial distribution $(\mu_i)_{i\in E}$, a Markov transition matrix $(p_{ij})_{i,j\in E}$ and conditional sojourn time distributions $(f_{ij}(k))_{i,j\in E,k\in\mathbb{N}}$.

Another assumptions we do are: (i) We do not allow transitions to the same state, i.e., $p_{ii} = 0$ for all $i \in E$, or equivalently $q_{ii}(k) = 0$, for all $i \in E,\ k \in \mathbb{N}$; (ii) We assume that there are not instantaneous transitions, that is $q_{ij}(0) \equiv 0$ or equivalently $f_{ij}(0) \equiv 0$ for all $i, j \in E$; note that this implies that T is a strictly increasing sequence.

For the conditional sojourn time distribution and sojourn time distribution in a state, one can consider the associated cumulative distribution functions defined by

$$\begin{aligned} F_{ij}(k) &:= \mathbb{P}(T_{n+1} - T_n \leq k | J_n = i, J_{n+1} = j) = \sum_{t=0}^{k} f_{ij}(t); \\ H_i(k) &:= \mathbb{P}(T_{n+1} - T_n \leq k | J_n = i) = \sum_{t=0}^{k} h_i(t). \end{aligned}$$

For any distribution function $F(\cdot)$ we can consider the associated survival/reliability function defined by

$$\overline{F}(k) := 1 - F(t).$$

Consequently we have

$$\begin{aligned} \overline{F}_{ij}(k) &:= \mathbb{P}(T_{n+1} - T_n > k | J_n = i, J_{n+1} = j) = 1 - \sum_{t=0}^{k} f_{ij}(t) = \sum_{t=k+1}^{\infty} f_{ij}(t); \\ \overline{H}_i(k) &:= \mathbb{P}(T_{n+1} - T_n > k | J_n = i) = 1 - \sum_{t=0}^{k} h_i(t) = \sum_{t=k+1}^{\infty} h_i(t). \end{aligned}$$

For investigating the reliability behaviour of a semi-Markov system, we split the space E into two subsets, U for the up-states and D for the down-states, $E = U \cup D$ and $E = U \cap D = \varnothing$. For simplicity, we consider $U = \{1, \ldots, s_1\}$ and $D = \{s_1 + 1, \ldots, s\}$.

Two important reliability measures of a system are the reliability (or survival) function at time $k \in \mathbb{N}$, denoted by $R(k)$, and the (instantaneous) availability function at time $k \in \mathbb{N}$, denoted by $A(k)$, defined respectively by

$$\begin{aligned} R(k) &:= \mathbb{P}(Z_0 \in U, \ldots, Z_k \in U), \\ A(k) &:= \mathbb{P}(Z_k \in U). \end{aligned}$$

If ever we condition on the initial state, we obtain the corresponding conditional reliability (or survival) function at time $k \in \mathbb{N}$ given that $\{Z_0 = i\}, i \in U$, denoted by $R_i(k)$, and the conditional (instantaneous) availability function at time $k \in \mathbb{N}$ given that $\{Z_0 = i\}, i \in E$, denoted by $A_i(k)$, defined respectively by

$$\begin{aligned} R(k) &:= \mathbb{P}(Z_0 \in U, \ldots, Z_k \in U \mid Z_0 = i), i \in U, \\ A(k) &:= \mathbb{P}(Z_k \in U \mid Z_0 = i), i \in E. \end{aligned}$$

Let us now define the interval reliability, introduced in Georgiadis et al. (2014) as the probability that the is in up-states during a time interval.

Definition 3 (interval reliability, conditional interval reliability, see Georgiadis et al. (2014))**.** *For $k, p \in \mathbb{N}$ and $i \in E$, the interval reliability $IR(k,p)$ and conditional interval reliability $IR_i(k,p)$ given the event $\{Z_0 = i\}$ are respectively defined by*

$$IR(k,p) := \mathbb{P}(Z_l \in U, l \in [k, k+p]); \tag{3}$$
$$IR_i(k,p) := \mathbb{P}(Z_l \in U, l \in [k, k+p] \mid Z_0 = i). \tag{4}$$

For $k, p \in \mathbb{N}$ and $i \in E$, it is clear that we have the following properties of the interval reliability and conditional interval reliability (cf. Proposition 1 and Remark 1 of [**?**]):

$$R(k+p) \leq IR(k,p) \leq A(k+p); \tag{5}$$
$$IR_i(0,p) = R_i(p); \tag{6}$$
$$IR_i(k,0) = A_i(k). \tag{7}$$

2 Sequential interval reliability

Let us consider a repairable system. In this section we introduce a new reliability measure, that we will call sequential interval reliability. This will generalize the notion of interval reliability presented before, in the sense that we are looking at the probability that the system is in working mode during two or several non-overlapping intervals.

More precisely, let us consider $\underline{t} := (t_i)_{i=1,\ldots,N}$ and $\underline{p} := (p_i)_{i=1,\ldots,N}$ two time sequences such that:

1. $t_1 > 0$, $t_i < t_{i+1}$ for all $i = 1, \ldots, N-1$;
2. $p_i \geq 0$ for all $i = 1, \ldots, N$;
3. $t_i + p_i < t_{i+1}$ for all $i = 1, \ldots, N-1$.

It is clear that, in this case, $\{[t_i, t_i + p_i]\}_{i=1,\ldots,N}$ is a sequence of non-overlapping real intervals.

For a sequence $\underline{t} := (t_i)_{i=1,\ldots,N}$ indexes $k_1, k_2 \in \mathbb{N}, k_1 \leq k_2$, we will also use the notation $t_{k_1:k_2} := (t_i)_{i=k_1,\ldots,k_2}$.

Definition 4 (sequential interval reliability, see Barbu et al. (2021))**.** *Let $(Z_k)_{k\in\mathbb{N}}$ be a discrete time semi-Markov system and let $\underline{t} := (t_i)_{i=1,\ldots,N}$ and $\underline{p} := (p_i)_{i=1,\ldots,N}$, $N \in \mathbb{N}^*$, be two time sequences such that $\{[t_i, t_i + p_i]\}_{i=1,\ldots,N}$ is a sequence of non-overlapping real intervals.*

1. *We define the sequential interval reliability,* $SIR^{(N)}(\underline{t},\underline{p})$*, as the probability that the system is in the up-states* U *during the time intervals* $\{[t_i, t_i+p_i]\}_{i=1,\ldots,N}$*, i.e.,*

$$SIR^{(N)}(\underline{t},\underline{p}) := \mathbb{P}(Z_l \in U, \text{ for all } l \in [t_i, t_i+p_i], i=1,\ldots,N); \tag{8}$$

2. *For* $v \in \mathbb{N}$ *and* $k \in E$*, we define the conditional sequential interval reliability,* $SIR_k^{(N)}(v;\underline{t},\underline{p})$*, as the conditional probability that the system is in the up-states* U *during the time intervals* $\{[t_i, t_i+p_i]\}_{i=1,\ldots,N}$*, given the event* $(k,v) := \{Z_0 = k, B_0 = v\} = \{J_{N(0)} = k, T_{N(0)} = -v\}$*, i.e.,*

$$\begin{aligned} SIR_k^{(N)}(v;\underline{t},\underline{p}) &:= \mathbb{P}(Z_l \in U, \text{ for all } l \in [t_i, t_i+p_i], i=1,\ldots,N \mid Z_0 = k, B_0 = v) \\ &= \mathbb{P}_{(k,v)}(Z_l \in U, \text{ for all } l \in [t_i, t_i+p_i], i=1,\ldots,N), \end{aligned} \tag{9}$$

where $B_t := t - T_{N(t)}$ *is the backward time process associated to the semi-Markov process.*

Note that we have the obvious relationship between the sequential interval reliability and the conditional sequential interval reliability:

$$SIR^{(N)}(\underline{t},\underline{p}) = \sum_{k\in E} \mu_k SIR_k^{(N)}(0;\underline{t},\underline{p}). \tag{10}$$

For notational convenience, we will set

$$SIR_k^{(N)}(\underline{t},\underline{p}) := SIR_k^{(N)}(0;\underline{t},\underline{p}) = \mathbb{P}(Z_l \in U, \text{ for all } l \in [t_i, t_i+p_i], i=1,\ldots,N \mid Z_0 = k).$$

Remark 1. *Under the previous notation, we have:*

1. *If* $t_i + p_i = t_{i+1} - 1$ *for all* $i = 1,\ldots,N-1$*, and* $v = 0$*, then* $SIR_k^{(N)}(0;\underline{t},\underline{p}) = IR(t_1, t_N + p_N - t_1)$*;*

2. *If* $t_1 = 0$*,* $t_i + p_i = t_{i+1} - 1$ *for all* $i = 1,\ldots,N-1$*,* $k \in U$ *and* $v = 0$*, then* $SIR_k^{(N)}(0;\underline{t},\underline{p}) = R_k(t_N + p_N)$*;*

3. *If* $p_i = 0$ *for all* $i = 1,\ldots,N$*, and* $v = 0$*, then* $SIR_k^{(N)}(0;\underline{t},\underline{p}) = \mathbb{P}_k(Z_{t_i} \in U, i = 1,\ldots,N) =: SA_k^{(N)}(\underline{t})$*; this function denoted by* $SA_k^{(N)}(\underline{t})$ *can be called sequential availability function;*

4. *If there exists a* $j \in \{1,\ldots,N\}$ *such that* $t_i = 0$ *for* $i < j$ *and* $t_h = t_j$ *for* $h \geq j$*,* $p_i = 0$ *for all* $i = 1,\ldots,N$*, and* $v = 0$*, then* $SIR_k^{(N)}(0;\underline{t},\underline{p}) = A(t_j)$*, the availability function computed in* t_j*.*

2.1 Transient analysis

We will now investigate the recursive formula for computing the sequential interval reliability of a discrete-time semi-Markov system.

Proposition 1. *Let $(Z_k)_{k\in\mathbb{N}}$ be a discrete time semi-Markov system and let $\underline{t} := (t_i)_{i=1,\dots,N}$ and $\underline{p} := (p_i)_{i=1,\dots,N}$, $N \in \mathbb{N}^*$, be two time sequences such that $\{[t_i, t_i + p_i]\}_{i=1,\dots,N}$ is a sequence of non-overlapping real intervals. Let $v \in \mathbb{N}$ be the value of the backward process at time $t = 0$ and $k \in E$ be the initial state. Then, the conditional sequential interval reliability, $SIR_k^{(N)}(v; \underline{t}, \underline{p})$, satisfies the following equation:*

$$SIR_k^{(N)}(v;\underline{t},\underline{p}) \;=\; g_k^{(N)}(v;\underline{t},\underline{p}) + \sum_{r\in E}\sum_{\theta=1}^{t_1} \frac{q_{kr}(v+\theta)}{\overline{H}_k(v)} SIR_r^{(N)}(0;\underline{t}-\theta\mathbf{1}_{1:N},\underline{p}), \tag{11}$$

where $\mathbf{1}_{1:N}$ is a vector of 1s of length N, and $g_k^{(N)}(v;\underline{t},\underline{p})$ is given by

$$\begin{aligned}
g_k^{(N)}(v;\underline{t},\underline{p}) := \mathbb{1}_{\{k\in U\}} \Bigg[& \frac{\overline{H}_k(t_N+p_N+v)}{\overline{H}_k(v)} \\
& + \sum_{\theta=t_1+1}^{t_1+p_1} \sum_{r\in E}\sum_{m\in U}\sum_{v'=0}^{t_1+p_1-\theta} \frac{q_{kr}(v+\theta)}{\overline{H}_k(v)} R_{rm}^b(v';t_1+p_1-\theta) SIR_m^{(N-1)}(v';t_{2:N}-\mathbf{1}_{2:N}(t_1+p_1),p_{2:N}) \\
& + \sum_{j=2}^{N}\sum_{\theta=t_j}^{t_j+p_j}\sum_{r\in E} \frac{q_{kr}(v+\theta)}{\overline{H}_k(v)} SIR_r^{(N-j+1)}\left(0;(\theta,t_{j+1:N}-\mathbf{1}_{j+1:N}\theta),(t_j+p_j-\theta,p_{j+1:N})\right) \\
& + \sum_{j=1}^{N-1}\sum_{\theta=t_j+p_j+1}^{t_{j+1}-1}\sum_{r\in E} \frac{q_{kr}(v+\theta)}{\overline{H}_k(v)} SIR_r^{(N-j)}(0;t_{j+1:N}-\mathbf{1}_{j+1:N}\theta,p_{j+1:N}) \Bigg],
\end{aligned} \tag{12}$$

where $\mathbb{1}_{\{k\in U\}}$ is the indicator function of the event $\{k \in U\}$ and $R_{ij}^b(v;k)$ is the reliability with final backward defined by

$$R_{ij}^b(v;k) := \mathbb{P}(Z_s \in U,\ \textit{for all } s \in \{0,\dots,k-v\}, Z_k = j, B_k = v \mid Z_0 = i, T_{N(0)} = 0). \tag{13}$$

References

Barbu, V. S., D'Amico, G., Gkelsinis, T. (2021), Sequential Interval Reliability for Discrete-Time Homogeneous Semi-Markov Repairable Systems, *Mathematics*, 9(16):1997. doi: 10.3390/math9161997

Georgiadis, S., Limnios, N. (2015), Nonparametric estimation of interval reliability for discrete-time semi-Markov systems, *Journal of Statistical Theory and Practice*, 10, 20–39. doi: 10.1080/15598608.2015.1064049

Georgiadis, S., Limnios, N. (2014), Interval reliability for semi-Markov systems in discrete time, *Journal de la Société Française de Statistique*, 153, 152–166.

Building wind loading estimation under uncertain inflow parameters by means of lattice Boltzmann simulation.

Jérôme Jacob [1], Lucie Merlier[2], and Pierre Sagaut[1]

[1]Laboratoire de Mécanique, Modélisation et Procédés Propres – Centre National de la Recherche Scientifique : UMR7340 / UMR6181, Ecole Centrale de Marseille, Aix Marseille Université – France
[2]Centre dÉnergétique et de Thermique de Lyon – Université Claude Bernard Lyon 1, Institut National des Sciences Appliquées de Lyon, Centre National de la Recherche Scientifique : UMR5008 – France

Résumé

Large eddy simulation is becoming lately an interesting tool for the simulation of flows inside complex urban areas for the estimation of building wind loading. To ensure the reliability of the predicted wind loading it becomes more and more important to analyze its response to inflow uncertainties. The present study combines the use of a lattice Boltzmann method(1) (LBM) for the simulation of building wind loading and the anchored ANOVA POD/Kriging(2) (c-APK) method for the uncertainty quantification in order to efficiently predict the response surfaces of pressure coefficients and aerodynamic forces on the different facades of a high rise building. For that purpose, numerical simulation have been carried out for the case of a simplified building located in the Shinjuku area in Tokyo using the same configuration described in Jacob and Sagaut(3) considering several inflow wind velocity and direction around a reference configuration considered as the deterministic reference. The obtained results highlighted differences between the physical quantities obtained from the reference sample and the ones obtained from the c-APK analysis. From these results it is shown that the variability of integrated forces on building walls in terms of magnitude and orientation is more important on the highest half of the building which is the area where these forces are the more important. The inflow wind direction is also highlighted to be the main parameter influencing wind loading in the top half of the building whereas inflow wind magnitude is of main importance in the bottom half, mainly in the built area. All the obtained results show the importance of uncertainty quantification to improve the understanding of wind loading sensitivity to the model setting in order to improve the reliability of the simulated wind loading in the framework of industrial applications.
(1) – J. Jacob, O. Malaspinas and P. Sagaut, A new hybrid recursive regularized Bhatnagar-Gros-Krook collision model for lattice Boltzmann method-based large eddy simulation. Journal of Turbulence, 2018

(2) – L. Margheri and P. Sagaut, A hybrid anchored-ANOVA POD/Kriging method for uncertainty quantification in unsteady high fidelity CFD simulations. Journal of Computational Physics, 2016
(3) – J. Jacob and P. Sagaut, Wind comfort assessment by means of large eddy simulation with lattice Boltzmann method in full scale city area. Building and Environment, 2018

Méta-modélisation multi-fidélité par processus gaussien en utilisant une base d'ondelettes

Baptiste Kerleguer[1,2] & Claire Cannamela [1] & Josselin Garnier [2]

[1] *CEA, DAM, DIF, F-91297, Arpajon, France baptiste.kerleguer@cea.fr* [2] *Centre de Mathématiques Appliquées, Ecole polytechnique, Institut Polytechnique de Paris, 91128 Palaiseau Cedex, France*

Résumé. La régression par processus gaussien est largement utilisée pour émuler la sortie d'un code coûteux. Nous nous intéressons à des codes multi-fidélité, qui peuvent être exécutés avec différents niveaux de précision et de coût et dont les sorties sont de grande dimension. Ici nous considérons des sorties sous la forme de séries temporelles. Nous étendons le modèle de co-krigeage auto-regressif proposé par Kennedy-O'Hagan aux sorties de grande dimension. Pour cela nous réalisons la régression sur les coefficients d'ondelettes de la sortie à l'aide d'une fonction de covariance spéciale. Nous montrons que le modèle permet de prédire efficacement en terme d'erreur de prédiction, mais aussi de quantification d'incertitudes.

Mots-clés. Processus Gaussiens, Transformée en ondelettes

Abstract. Gaussian process regression is widely used to emulate the output of an expensive code. We are interested in multi-fidelity codes, which can be run with different levels of accuracy and cost and whose outputs are high-dimensional. Here the code outputs are time series. We extend the auto-regressive co-kriging model proposed by Kennedy-O'Hagan to high-dimensional outputs. To do so, we perform the regression on the wavelet coefficients using a special covariance function. We show that the model allows for efficient prediction in terms of prediction error but also for the quantification of uncertainties.

Keywords. Gaussian process, Wavelet transform

1 Introduction

Il existe de nombreux exemples d'utilisation de régression par processus gaussien dans un cadre multi-fidélité dérivés de l'article de Kennedy et O'Hagan (2000). Dans cet article une relation linéaire est supposée entre la sortie du code haute et celle de basse fidélité. Cette méthode, basée sur la régression par processus gaussien, permet de quantifier les incertitudes de prédiction. Elle est également très efficace dans des contextes de données peu nombreuses, très fréquents lorsque le code est coûteux en temps de calcul. Ses principales limites sont le besoin d'hypothèses sur la régularité de la relation entre les sorties de différents niveaux de fidélité et surtout la grande difficulté de passer en grande dimension (en entrée et en sortie du modèle).

En parallèle, il existe différentes méthodes de réduction de la dimension efficaces qui permettent de se ramener à un problème de plus petite dimension. Dans une étude précédente dans Kerleguer (2021), nous avons utilisé des méthodes de réduction de la dimension qui dépendent des données. Or cela pose un problème pour la quantification de l'incertitude de prédiction, car la base de réduction de la dimension dépend des données et est donc incertaine. Pour cela il est intéressant d'utiliser une base qui ne dépend pas des données. Les bases d'ondelettes semblent alors très adaptées et en particulier les ondelettes de Daubechies.

Notre objectif est de combiner la régression multi-fidélité avec une décomposition sur une base d'ondelettes. On espère ainsi avoir un grand nombre de coefficients proches de zéro qu'on pourra négliger pour l'apprentissage, mais qui seront pris en compte dans l'incertitude de prédiction. Ainsi, il devient possible de faire de la régression par processus Gaussien avec une sortie de la forme d'une série temporelle.

2 Méthode

On note $f_L(\mathbf{x}, t)$ et $f_H(\mathbf{x}, t)$ les fonctions de basse et haute fidélité. On veut construire un métamodèle de f_H à partir de données issues d'appels des fonctions f_H et f_L. On utilise une approche par co-krigeage avec comme modèle le processus (F_L, F_H) :

$$F_L(\mathbf{x}, t) = B_L(\mathbf{x})\phi(t) + \sum_{(j,k)\in I} A_L(\mathbf{x}, j, k)\psi_{j,k}(t), \tag{1}$$

$$F_H(\mathbf{x}, t) = B_H(\mathbf{x})\phi(t) + \sum_{(j,k)\in I} A_H(\mathbf{x}, j, k)\psi_{j,k}(t), \tag{2}$$

où (A_L, A_H) et (B_L, B_H) sont deux processus gaussiens indépendants, ϕ est la fonction d'échelle associée à l'ondelette mère de Daubechies choisie et ψ l'ondelette de Daubechies de niveau j avec un décalage de k. I est l'ensemble des valeurs possibles du couple (j, k).

Pour construire une modèle de la fonction haute fidélité, on utilise un co-krigeage simple multi-fidélité à réaliser avec (B_L, B_H) indépendamment du reste. Pour le second terme des équations (1) et (2) le processus gaussien dépend de j et k ce qui nous oblige à construire une fonction de covariance associée. Pour cela on a supposé que l'évolution temporelle suivait un processus stationnaire de Matérn 5/2. Ainsi on peut construire une fonction de covariance en (j, k) appropriée, de la forme $C(j, k, j', k')$ avec $(j, k) \times (j', k') \in I^2$. La fonction de covariance en $(\mathbf{x}, j, k)$ est construite par tensorisation avec une covariance de Matérn en $\mathbf{x}$.

On dispose de N_F (avec F la fidélité L ou H) points $\mathbf{x}_i^F$ pour lesquels on dispose de N_t valeurs de t sur une grille $\{t_u, u = 1, \ldots, N_t\}$. On peut alors obtenir la loi a postériori de $F_H(\mathbf{x}, t)$ en un nouveau point $(\mathbf{x}, t)$ en conditionnant par $F_F(\mathbf{x}_i^F, t_u) = f_F(\mathbf{x}_i^F, t_u)$, $u = 1, \ldots, N_t$, $i = 1, \ldots, N_F$, $F \in \{L, H\}$. Une réduction de la dimension est nécessaire à cause de la complexité algorithmique du conditionnement. Il y a $N_L \times N_t$ point $(\mathbf{x}_i^L, j, k)$

lors du conditionnement basse fidélité. Pour cela on choisit les points $(\mathbf{x}_i^L, j, k)$ à prendre en compte en fonction du rapport entre la valeur observée $f_L(\mathbf{x}_i^L, t_u)$ et l'écart type a priori. On prend les points qui ont les rapports les plus grands. On réalise la prédiction avec tous les points $(\mathbf{x}, j, k)$ retenus.

3 Résultats

Nous présentons des résultats pour des codes à deux niveaux de fidélité ainsi que des sorties sous la forme de séries temporelles. Les codes que nous utilisons sont des fonctions analytiques dont la dimension d'entrée est de faible dimension. Ainsi nous devons aussi nous comparer à des modèles plus complexes ainsi que des méthodes de réduction de dimension simples. Les performances de notre approche semblent comparables à celles obtenues par d'autres méthodes. Les hyper-paramètres étant peu nombreux, la régression dans un contexte de nombre de données faible est possible.

Bibliographie

I. Daubechies (1988). Orthonormal bases of compactly supported wavelets, *Communications on pure and applied mathematics*, 41, pp. 909-996.
M. Kennedy et A. O'Hagan, (2000), Predicting the output from a complex computer code when fast approximations are available, *Biometrika*.
Kerleguer, B. (2021). Multi-Fidelity Surrogate Modeling for Time-Series Outputs. arXiv preprint arXiv:2109.11374.

SFdS - Jeunes Statisticiens

La santé mentale des jeunes chercheurs et chercheuses

Cette année, notre session spéciale portera sur la santé mentale des jeunes chercheurs et chercheuses. A cette occasion, nous espérons ouvrir la discussion sur ce sujet encore trop souvent tabou dans le monde de la recherche.

Lors de cet évènement, un créneau consacré à la parole et au témoignage de celles et ceux qui sont actuellement doctorant.e.s, stagiaires, ou jeunes docteur.e.s. Afin de recueillir cette parole, anonymement ou non, nous avons rédigé ce bref questionnaire que nous vous proposons de remplir (que vous comptiez participer ou non aux JdS). Votre témoignage pourrait être précieux pour la réussite de cet évènement, ainsi que pour ceux qui l'entendront.

Il s'agit d'une occasion que nous espérons importante de partager vos expériences, sans jugements et sans conséquences néfastes sur votre vie professionnelle. Nous espérons ainsi que certains vécus, aujourd'hui invisibilisés, puissent sortir du tabou, afin d'éviter une perpétuation des mêmes problèmes dans le futur.

Lien du questionnaire : https://framaforms.org/session-jeunes-jds-2022-la-sante-mentale-des-jeunes-chercheurs-et-chercheuses-1653321969

Agrégation d'experts

Modélisation de la charge et de l'occupation des véhicules électriques

Yvenn Amara-Ouali [1] & Bachir Hamrouche [2] & Yannig Goude [3] & Pascal Massart [4] & Jean-Michel Poggi [5]

[1] *Université Paris-Saclay, Inria Saclay, yvenn.amara-ouali@universite-paris-saclay.fr* [2] *EDF R&D et bachir.hamrouche@edf.fr* [3] *Université Paris-Saclay, EDF R&D, yannig.goude@edf.fr* [4] *Université Paris-Saclay, Inria Saclay, pascal.massart@universite-paris-saclay.fr* [5] *Université Paris-Saclay, Université de Paris, jean-michel.poggi@universite-paris-saclay.fr*

Résumé. Le développement des véhicules électriques est un levier majeur vers un transport bas carbone. Il s'accompagne d'un nombre croissant d'infrastructures de recharge qui peuvent être utilisées comme actifs flexibles de gestion du réseau. Pour permettre cette recharge intelligente (smart-charging), une prévision journalière efficace des comportements de charge est nécessaire. L'objectif de nos travaux est d'évaluer la performance de modèles de prévision des courbes de charge et d'occupation des points de charge sur 8 jeux de données ouverts. Nous étudions deux approches de modélisation : directe et bottom-up. L'approche directe consiste à prévoir la courbe de charge agrégée (resp. l'occupation des points de charge) d'une zone/station. L'approche bottom-up consiste à modéliser les sessions de recharge individuelles pour ensuite les agréger. Cette dernière est essentielle pour implémenter des stratégies de smart-charging. Nous montrons que les approches directes sont généralement plus performantes que les approches bottom-up. Le meilleur modèle peut néanmoins être amélioré en agrégeant les prévisions des approches directes et bottom-up à l'aide d'une stratégie d'agrégation adaptative.

Mots-clés. Apprentissage automatique, Modélisation statistique, Agrégation d'experts,

Abstract. The development of electric vehicles is a major lever towards low-carbon transport. It comes with a growing number of charging infrastructures that can be used as flexible assets for the grid. To enable this smart-charging, an effective daily forecast of charging behaviours is necessary. The purpose of our work is to evaluate the performance of models for predicting load curves and charging point occupancy on 8 open data sets. We study two modelling approaches: direct and bottom-up. The direct approach consists in forecasting the aggregate load curve (resp. the occupancy of the charging points) of an area/station. The bottom-up approach consists in modeling individual charging sessions and then aggregating them. The latter is essential for implementing smart-charging strategies. We show that direct approaches generally perform better than bottom-up approaches. The best model can nevertheless be improved by aggregating the predictions of the direct and bottom-up approaches using an adaptive aggregation strategy.

Keywords. Machine learning, Statistical modelling, Aggregation of experts

1 Introduction

Les véhicules électriques (VE) représentent un levier important pour une transition vers des transports bas carbone. En gérant soigneusement la recharge des VE, ils peuvent devenir un atout pour le réseau de diverses manières (e.g., équilibrage ou aplatissement de la courbe de charge). L'exploitation d'un système reposant sur la recharge intelligente ou "smart-charging" nécessite une compréhension approfondie des mécanismes qui sous-tendent les comportements de recharge. Par conséquent, au cours de la dernière décennie, de nombreux articles ont abordé la question des modèles de charge des VE, et en particulier, la prévision de la demande des VE.

Le marché des VE étant encore à ses débuts, la plupart des articles s'appuient sur des données simulées ou privées (souvent expérimentales) pour spécifier des modèles de recharge des VE. Le manque d'études utilisant des données ouvertes entrave la reproductibilité des recherches et l'établissement de modèles standards. Dans le même temps, plusieurs jeux de données ouverts ont été rendus disponibles mais sont encore sous-utilisés par la communauté [1]. Ce n'est que récemment que des études portent sur ces données ouvertes : [2] détaille un modèle graphique d'apprentissage profond du réseau pour prévoir la demande énergétique quotidienne des bornes de recharge publiques à Palo Alto (États-Unis). De même, [3] présente un réseau de neurones récurrent pour prévoir l'occupation des bornes de recharge publiques à Dundee (Royaume-Uni) à un horizon intra-journalier. Enfin, [4] présente une étude statistique centrée sur les données qui démontre l'importance des informations spatiales sur les comportements de tarification.

Pour reconstruire la courbe de charge, trois variables sont nécessaires pour chaque session de charge i : l'heure d'arrivée (a_i), la durée de charge (d_i) et l'énergie consommée (e_i). Avec ces trois variables, une courbe de charge élémentaire peut être reconstruite sous la forme d'une fonction en escalier :

$$l_i(t) = \begin{cases} \frac{e_i}{d_i} & \text{si } t \in [a_i, a_i + d_i] \\ 0 & \text{sinon.} \end{cases}$$

La courbe de charge globale est obtenue en sommant toutes les courbes de charge élémentaires : $L(t) = \sum_i l_i(t)$. Quant à la courbe d'occupation, le processus est encore plus simple. Seules deux variables de sessions de recharge sont requises: l'heure d'arrivée (a_i) et la durée de stationnement (p_i). Là encore, une courbe d'occupation élémentaire peut être reconstruite pour chaque session i :

$$o_i(t) = \begin{cases} 1 & \text{if } t \in [a_i, a_i + p_i] \\ 0 & \text{otherwise.} \end{cases}$$

et la courbe d'occupation globale est obtenue en sommant toutes les courbes d'occupation élémentaires : $O(t) = \sum_i o_i(t)$

2 Méthodologies

Une grande variété de modèles ont été présentés dans la littérature pour résoudre les problèmes de prévision liés aux VE [1]. On distingue deux approches majeures: une approche *directe* où la charge et/ou l'occupation sont prévues au niveau agrégé; une approche individuelle ou *bottom-up* où la charge et/ou l'occupation peuvent être déduites de sessions de recharge individuelles simulées par le modèle (voir Figure 1). Pour chaque session de recharge i, on suppose que a_i, d_i, p_i et e_i sont des réalisations des variables aléatoires A, D, P et E. Le vecteur cible pour la charge est (A, D, E). Le vecteur cible pour l'occupation est (A, P). Y est la variable cible qui peut être soit la charge, soit l'occupation.

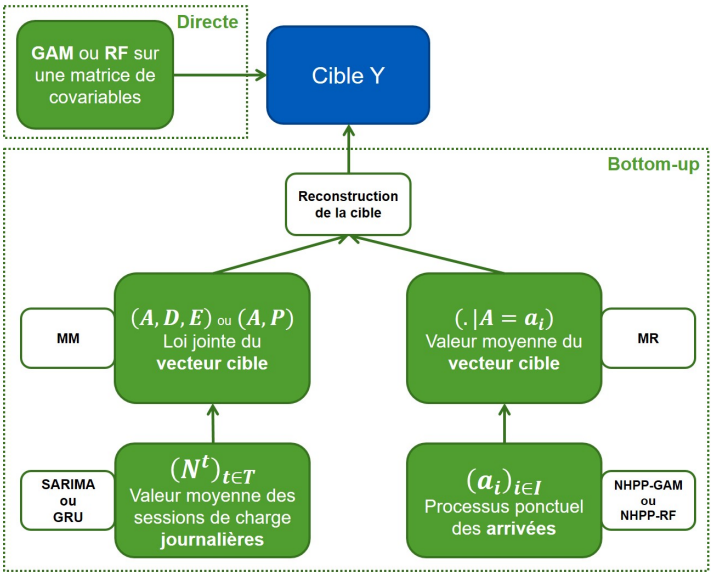


Figure 1: Approches directes et bottom-up

L'approche directe utilise la configuration traditionnelle de prévision de la charge électrique. Si le nombre de sessions est suffisamment élevé, l'agrégation de l'occupation ou de la charge de chaque session permettra de régulariser le signal en bénéficiant de la loi des grands nombres. Par conséquent, nous pouvons nous attendre à obtenir de bons résultats avec des méthodes similaires utilisées pour les données agrégées de consommation d'électricité. Ici, nous considérons en particulier les modèles additifs généralisés (GAM) ainsi que les forêts aléatoires (RF). De plus, un modèle de persistance utilisant la valeur observée de la cible 24 heures auparavant est considéré comme référence.

L'approche bottom-up est très pertinente pour évaluer les avantages des stratégies de recharge intelligentes. La prévision des sessions de recharge individuelles présente de nombreux avantages : elle permet de simuler naturellement différents schémas de recharge

et peut intégrer des informations individuelles (e.g., contraintes de recharge personnelles, trafic). Une session de recharge i est caractérisée par trois variables à estimer : l'heure d'arrivée (a_i), la durée de charge/stationnement (d_i ou p_i) et la demande d'énergie (e_i). Les sessions de recharge individuelles sont modélisées par les méthodes suivantes:

- Processus de Poisson non-homogène estimé avec GAM et mélange de régressions (NHPP-GAM-MR)
- Processus de Poisson non homogène estimé avec RF et mélange de régressions (NHPP-RF-MR)
- Gated Recurrent Units et modèle de mélange (GRU-MM)
- Seasonal Auto-Regressive Moving Average et modèle de mélange (SARIMA-MM)

Un modèle de correction d'erreur est également proposé pour ajuster chaque prévision. Il consiste en un modèle SARIMA sur les résidus. Les modèles de correction d'erreur sont écrits comme *Modèle*-err. Par exemple, GAM-err est la prévision produite par le modèle GAM combiné avec le modèle de correction d'erreur. Enfin, nous proposons un dernier modèle qui consiste en une agrégation des prévisions produites jusqu'ici (AGG). Nous utilisons l'algorithme ML-Poly proposé dans [5]. Cet algorithme suit le meilleur expert ou la meilleure agrégation convexe d'experts en donnant plus de poids à un expert qui générera un faible regret. Cela rend cet algorithme particulièrement intéressant car aucun réglage de paramètre n'est nécessaire.

3 Résultats

Nous considérons 8 jeux de données de sessions de recharge ouvertes qui sont présentés en détail dans [1]. Les comportements de recharge les plus courants sont couverts dans cette sélection: publique, au travail et résidentielle. Notez que les week-ends sont exclus de notre analyse en raison d'un faible nombre de sessions de recharge observées sur certains jeux de données et de comportements de recharge significativement différents par rapport aux jours ouvrés. Le profil des arrivées aux points de charge sont présentés en Figure 2 pour chaque jeux de données.

Les modèles considérés nécessitent une optimisation des hyperparamètres afin de les exploiter au mieux mais aussi d'éviter le sur-apprentissage. Par conséquent, nous avons défini une procédure de sélection pour chaque ensemble de données en raison de la grande variété de comportements de charge des véhicules électriques qu'ils représentent. Cette sélection est effectuée par BIC (Bayesian Information Criterion) pour les modèles statistiques et via une "grid search" pour les modèles d'apprentissage automatique. De même, une procédure de validation des modèles est également effectuée pour vérifier les hypothèses statistiques retenues pour les modèles statistiques concernés. Les modèles sont

validés par un test de Student pour les NHPP, un test de Ljung-Box pour les modèles SARIMA et une analyse empirique des résidus pour les modèles GAM.

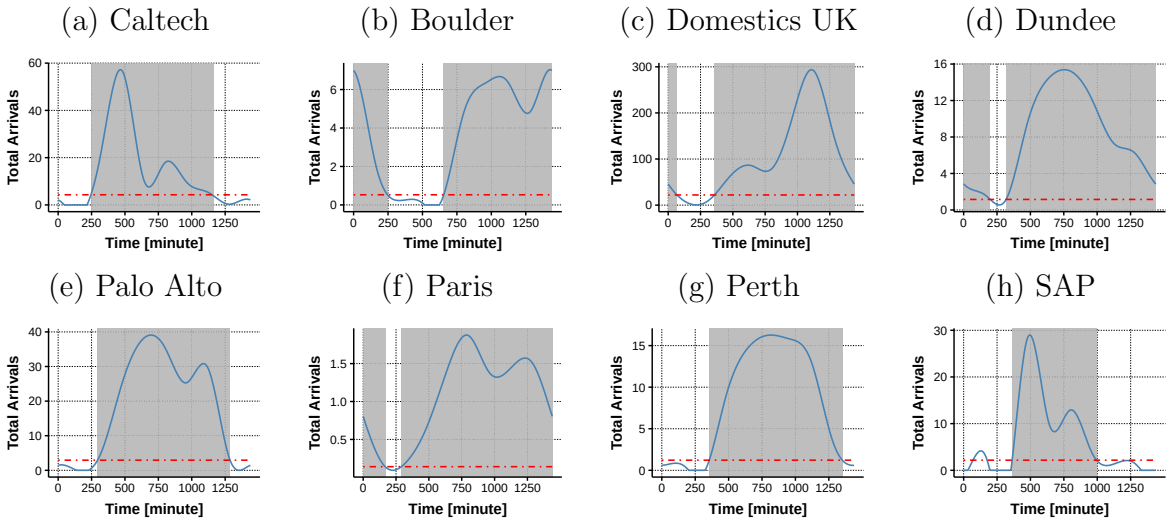


Figure 2: Somme des arrivées par minute de la journée au cours d'une période fixée pour les 8 jeux de données. L'aire grisée représente la fenêtre temporelle que nous gardons pour l'étude. La droite horizontale en rouge définit un seuil en-dessous duquel les instants associés ne sont pas conservés.

Nous définissons une métrique interprétable sur plusieurs jeux de données avec différentes échelles : le pourcentage d'amélioration de la prévision par rapport à la persistance (PIP). Formellement, nous pouvons écrire le PIP pour un signal de prévision donné $\{\hat{y}_i\}_{i\in\{1,\dots,n\}}$ comme suit: $\text{PIP} = 100 \times \left(\frac{\frac{1}{n}\sum_{i=1}^{n}|y_i - y_i^{pers}|}{\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|} - 1\right)$ avec $\{y_i\}_{i\in\{1,\dots,n\}}$ le signal observé, $\{y_i^{pers}\}_{i\in\{1,\dots,n\}}$ le signal prévu par la persistance et n la longueur du signal observé. Concrètement, le PIP est le rapport entre la norme L^1 des résidus de la persistance et celle du signal de prévision considéré auquel on soustrait 1. Les boxplots des erreurs de prévision en block-bootstrap [6] sont présentées sur la Figure 3. Les prévisions des modèles proposés sont en général nettement meilleures que la persistance et la meilleure prévision est produite par l'agrégation d'experts (AGG). Les approches directes ont obtenu de meilleurs résultats que les approches bottom-up. Comme prévu, l'approche directe bénéficie d'un effet "loi des grands nombres" où la somme des signaux individuels lisse le signal agrégé qui devient plus facile à prévoir. Néanmoins, compte tenu de la difficulté induite par le paradigme de l'approche bottom-up, on peut noter les performances relativement bonnes de celles basées sur le modèle de mélange (MM). La correction d'erreur basée sur les modèles ARIMA apporte toujours une amélioration significative. Les méthodes NHPP sont les moins bonnes et nous pensons que cela pourrait provenir de la partie mélange de régression (MR) qui relie l'arrivée aux autres composantes des vecteurs cibles.

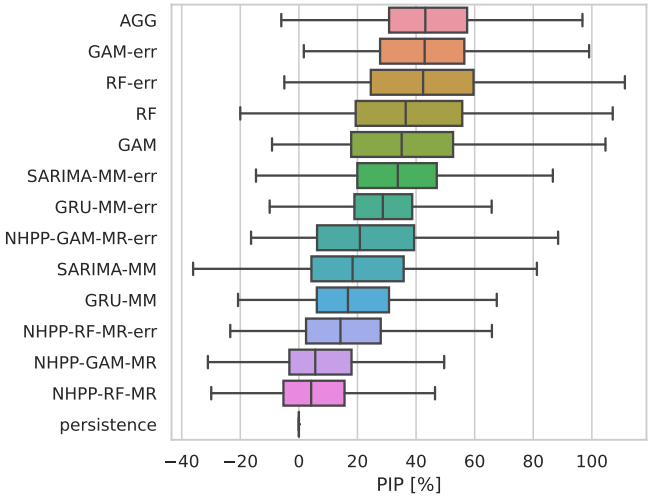


Figure 3: Boxplots des performances PIP en block-bootstrap sur tous les jeux de données et pour tous les modèles considérés

References

[1] Yvenn Amara-Ouali, Yannig Goude, Pascal Massart, Jean-Michel Poggi, and Hui Yan. A Review of Electric Vehicle Load Open Data and Models. *Energies*, 14(8):2233, January 2021. Number: 8 Publisher: Multidisciplinary Digital Publishing Institute.

[2] Frederik Boe Hüttel, Inon Peled, Filipe Rodrigues, and Francisco C. Pereira. Deep Spatio-Temporal Forecasting of Electrical Vehicle Charging Demand. *arXiv:2106.10940 [cs]*, June 2021. arXiv: 2106.10940.

[3] Tai-Yu Ma and Sébastien Faye. Multistep electric vehicle charging station occupancy prediction using hybrid LSTM neural networks. *Energy*, 244:123217, April 2022.

[4] Milan Straka, Rui Carvalho, Gijs Van Der Poel, and L'uboš Buzna. Analysis of Energy Consumption at Slow Charging Infrastructure for Electric Vehicles. *IEEE Access*, 9:53885–53901, 2021. Conference Name: IEEE Access.

[5] Pierre Gaillard, Gilles Stoltz, and Tim Van Erven. A second-order bound with excess losses. In *Conference on Learning Theory*, pages 176–196, 2014.

[6] Jakob W. Messner, Pierre Pinson, Jethro Browell, Mathias B. Bjerregård, and Irene Schicker. Evaluation of wind power forecasts—an up-to-date view. *Wind Energy*, 23(6):1461–1481, 2020.

Functional mixture-of-experts for classification

Nhat Thien Pham and Faïcel Chamroukhi

Normandie Univ, UNICAEN, CNRS, LMNO, 14000 Caen, France
nhat-thien.pham@unicaen.fr, *faicel.chamroukhi@unicaen.fr*

Abstract. We develop a mixtures-of-experts (ME) approach to the multiclass classification where the predictors are univariate functions. It consists of a ME model in which both the gating network and the experts network are constructed upon multinomial logistic activation functions with functional inputs. We perform a regularized maximum likelihood estimation in which the coefficient functions enjoy interpretable sparsity constraints on targeted derivatives. We develop an EM-Lasso like algorithm to compute the regularized MLE and evaluate the proposed approach on simulated and real data.

Keywords. Mixtures of experts, Functional predictors, EM algorithm, Sparsity

1 Introduction

Introduced in Jacobs et al. (1991), a mixtures of experts (ME) model can be defined as

$$f(y|x) = \sum_{k=1}^{K} \text{Gate}_k(x)\, \text{Expert}_k(y|x), \tag{1}$$

in which $f(y|x)$, the distribution of the response y given the covariate x, is modeled as a mixture distribution with covariate-dependent mixing proportions $\text{Gate}_k(x)$, referred to as gating functions, and conditional mixture components $\text{Expert}_k(y|x)$, referred to as experts functions, K being the number of experts. Some ME studies that may be mentioned here include ME for time series prediction (Zeevi et al., 1996; Yümlü et al., 2003), segmentation (Chamroukhi et al., 2013, 2009), ME for classification of gender and pose of human faces (Gutta et al., 2000), for social network data (Gormley and Murphy, 2010), among others. For an overview of practical and theoretical aspects of ME modeling, the reader is referred to Nguyen and Chamroukhi (2018). The study of ME for functional data analysis (FDA) (Ramsay and Silverman, 2005), is still however less investigated. In a recent study, we introduced in Chamroukhi et al. (2022) a functional ME (FME) framework for regression and clustering of observed pairs of scalar responses and univariate functional inputs.

2 Functional Mixture-of-Experts for classification

In this paper, we extend the FME framework for multiclass classification, derive adapted EM-like algorithms to obtain sparse and interpretable fit of the gating and experts network coefficients functions. Let $\{X_i(t), t \in \mathcal{T}; Y_i\}_{i=1}^n$, be a sample of n i.i.d. data pairs where $Y_i \in \{1, \ldots, G\}$ is the class label of a functional predictor $X_i(\cdot)$, G being the number

of classes. In this case of functional inputs, a natural choice to model the conditional distribution $\text{Expert}_k(y|x) = \mathbb{P}(Y = y|X_i(\cdot))$ in (1) is to use the functional multinomial logistic regression modeling, see e.g., Müller et al. (2005); James (2002), that is

$$P(y_i|X_i(\cdot);\boldsymbol{\beta}_k) = \prod_{g=1}^{G}\left[\frac{\exp\left\{\beta_{kg,0} + \int_{\mathcal{T}} X_i(t)\beta_{kg}(t)dt\right\}}{1+\sum_{g'=1}^{G-1}\exp\left\{\beta_{kg',0} + \int_{\mathcal{T}} X_i(t)\beta_{kg'}(t)dt\right\}}\right]^{y_{ig}}, \tag{2}$$

whre $\boldsymbol{\beta}_k$ represents the set of coefficient functions $\{\beta_{kg}(t), t \in \mathcal{T}\}$ and intercepts $\{\beta_{k,0}\}$ for $k \in [K] = \{1,\ldots,K\}$ and $g \in [G]$, and $y_{ig} = \mathbb{I}_{\{y_i=g\}}$. Similarly, a typical choice for the functional gating network $\text{Gate}_k(x) = \mathbb{P}(Z = k, X(\cdot))$ in (1), where $Z \in [K]$ is a hidden within-class clustering label, acting as weights for potential clusters $\{k\}$ in the heterogeneous functional inputs $X(\cdot)$ and which we denote as $\pi_k(X(\cdot))$, is to use a functional softmax function defined by

$$\pi_k(X_i(\cdot);\boldsymbol{\alpha}) = \frac{\exp\{\alpha_{k,0} + \int_{\mathcal{T}} X_i(t)\alpha_k(t)dt\}}{1+\sum_{k'=1}^{K-1}\exp\{\alpha_{k',0} + \int_{\mathcal{T}} X_i(t)\alpha_{k'}(t)dt\}}, \tag{3}$$

with $\boldsymbol{\alpha}$ is composed of the set of coefficient functions $\{\alpha_k(t), t \in \mathcal{T}\}$ and intercepts $\{\alpha_{k,0}\}$ for $k \in [K]$. Then, from (2) and (3) given $X_i(\cdot)$, the probability that $Y_i = y_i$, can be modeled by the following K-component FME model for classification

$$P(y_i|X_i(\cdot);\boldsymbol{\psi}) = \sum_{k=1}^{K}\pi_k(X_i(\cdot);\boldsymbol{\alpha})P(y_i|X_i(\cdot);\boldsymbol{\beta}_k), \quad \boldsymbol{\psi} = (\boldsymbol{\alpha},\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_K). \tag{4}$$

2.1 Smooth functional representation

In practice, $X_i(\cdot)$ is observed at a finite but large number of points on $\mathcal{T} \subset \mathbb{R}$. In the perspective of parameter estimation, this results in estimating a very large number of coefficients β and α. In order to handle this high-dimensional problem, we consider a usual approach that projects the predictors and coefficient functions onto a family of reduced number of basis functions. Let $\boldsymbol{b}_r(t) = [b_1(t),\ldots,b_r(t)]^\top$ be a r-dimensional basis (B-spline, Wavelet, ...). Then, with $r,p,q \in \mathbb{N}$ sufficiently large, one can approximate $X_i(\cdot)$, $\alpha_k(\cdot)$ and $\beta_{kg}(\cdot)$ respectively by

$$X_i(t) = \boldsymbol{x}_i^\top\boldsymbol{b}_r(t), \quad \alpha_k(t) = \boldsymbol{\zeta}_k^\top\boldsymbol{b}_p(t), \quad \beta_{kg}(t) = \boldsymbol{\eta}_{kg}^\top\boldsymbol{b}_q(t). \tag{5}$$

Here, $\boldsymbol{x}_i = (x_{i1},\ldots,x_{ir})^\top$, with $x_{ij} = \int_{\mathcal{T}} X_i(t)b_j(t)dt$ for $j \in [r]$, is the vector of coefficients of $X_i(\cdot)$ in the basis $\boldsymbol{b}_r(t)$, $\boldsymbol{\zeta}_k = (\zeta_{k,1},\ldots,\zeta_{k,p})^\top$, and $\boldsymbol{\eta}_{kg} = (\eta_{kg,1},\ldots,\eta_{kg,q})^\top$ are the unknown coefficient vectors associated with the gating coefficient function $\alpha_k(\cdot)$ and the expert coefficient function $\beta_{kg}(\cdot)$ in the corresponding basis. In our case, we used B-spline bases. Using the approximation of $X_i(\cdot)$ and $\alpha_k(\cdot)$ in (5), the functional softmax gating network (3) can be represented by

$$\pi_k(\mathbf{r}_i;\boldsymbol{\xi}) = \frac{\exp\{\alpha_{k,0} + \mathbf{r}_i^\top\boldsymbol{\zeta}_k\}}{1+\sum_{k'=1}^{K-1}\exp\{\alpha_{k',0} + \mathbf{r}_i^\top\boldsymbol{\zeta}_{k'}\}}, \tag{6}$$

where $\mathbf{r}_i = \left[\int_{\mathcal{T}} \boldsymbol{b}_r(t)\boldsymbol{b}_p(t)^\top dt\right]^\top \boldsymbol{x}_i$ is the design vector associated with the gating network and $\boldsymbol{\xi} = ((\alpha_{1,0}, \boldsymbol{\zeta}_1^\top), \ldots, (\alpha_{K-1,0}, \boldsymbol{\zeta}_{K-1}^\top)) \in \mathbb{R}^{(p+1)(K-1)}$ is the unknown parameter vector of the gating network, to be estimated. In the same manner, using the approximations of $X_i(\cdot)$ and $\beta_{kg}(\cdot)$ in (5), the expert conditional distribution (2) can be represented by

$$P(y_i|\mathbf{x}_i;\boldsymbol{\theta}_k) = \prod_{g=1}^{G}\left[\frac{\exp\left\{\beta_{kg,0}+\mathbf{x}_i^\top\boldsymbol{\eta}_{kg}\right\}}{1+\sum_{g'=1}^{G-1}\exp\left\{\beta_{kg',0}+\mathbf{x}_i^\top\boldsymbol{\eta}_{kg'}\right\}}\right]^{y_{ig}}, \tag{7}$$

where $\mathbf{x}_i = \left[\int_{\mathcal{T}} \boldsymbol{b}_r(t)\boldsymbol{b}_q(t)^\top dt\right]^\top \boldsymbol{x}_i$ is the design vector associated with the expert network, and $\boldsymbol{\theta}_k = (\boldsymbol{\theta}_{k1}^\top, \ldots, \boldsymbol{\theta}_{k,G-1}^\top)^\top$, with $\boldsymbol{\theta}_{kg} = (\beta_{kg,0}, \boldsymbol{\eta}_{kg}^\top)^\top \in \mathbb{R}^{q+1}$ for $g \in [G-1]$, is the unknown parameter vector to be estimated of the expert distribution k. Finally, combining (6) and (7), the conditional distribution $P(y_i|X_i(\cdot);\boldsymbol{\psi})$ in (4) can be rewritten as

$$P(y_i|X_i(\cdot);\boldsymbol{\Psi}) = \sum_{k=1}^{K}\pi_k(\mathbf{r}_i;\boldsymbol{\xi})P(y_i|\mathbf{x}_i;\boldsymbol{\theta}_k),$$

where $\boldsymbol{\Psi} = (\boldsymbol{\xi}^\top, \boldsymbol{\theta}_1^\top, \ldots, \boldsymbol{\theta}_K^\top)^\top$ is the unknown parameter vector of the model.
Parameter estimation: A maximum likelihood estimate (MLE) $\widehat{\boldsymbol{\Psi}}$ of $\boldsymbol{\Psi}$ can be obtained by using the EM algorithm for ME model for classification with vector data as in Chen et al. (1999). We will refer to this approach as FME-EM. To encourage sparsity in the model parameters $\boldsymbol{\Psi}$, one can perform penalized MLE by using the EM-Lasso algorithm as in Huynh and Chamroukhi (2019). We refer to this approach as FME-EM-Lasso.

2.2 An interpretable sparse estimation of FME for classification

Although fitting the FME model via EM-Lasso can accommodate sparsity in the parameters, it unfortunately does not ensure the reconstructed coefficient functions $\widehat{\alpha}_k(\cdot)$ and $\widehat{\beta}_{kg}(\cdot)$ are sparse and enjoy easy interpretable sparsity.To obtain interpretable and sparse fits for the coefficient functions, we simultaneously estimate the model parameters while constraining some targeted derivatives of the coefficient functions to be zero via (Chamroukhi et al., 2022). The construction of the interpretable FME model which we will fit with an adapted EM algorithm, is as follows. First, in order to calculate the derivative of the gating coefficient functions $\alpha_k(\cdot)$, let $\mathbf{A}_p$ be the matrix of approximate d_1th and d_2th derivative of $\boldsymbol{b}_p(t)$, defined as in James et al. (2009); Chamroukhi et al. (2022) by

$$\mathbf{A}_p = [\mathbf{A}_p^{[d_1]}\mathbf{A}_p^{[d_2]}]^\top = \left[D^{d_1}\boldsymbol{b}_p(t_1), \ldots, D^{d_1}\boldsymbol{b}_p(t_p), D^{d_2}\boldsymbol{b}_p(t_1), \ldots, D^{d_2}\boldsymbol{b}_p(t_p)\right]^\top,$$

where D^d is the dth finite difference operator. Here $\mathbf{A}_p^{[d_j]}$ is a square invertible matrix and $\mathbf{A}_p \in \mathbb{R}^{2p\times p}$. Similarly, to calculate the derivatives of the expert coefficient functions $\beta_{kg}(\cdot)$, let $\mathbf{A}_q = [\mathbf{A}_q^{[d_1]}\mathbf{A}_q^{[d_2]}]^\top \in \mathbb{R}^{2q\times q}$ be the corresponding matrix defined for the $\boldsymbol{b}_q(t)$'s. Now, if we define $\boldsymbol{\omega}_k = \mathbf{A}_p\boldsymbol{\zeta}_k$ and denote $\boldsymbol{\omega}_k = ({\boldsymbol{\omega}_k^{[d_1]}}^\top, {\boldsymbol{\omega}_k^{[d_2]}}^\top)^\top$, then $\boldsymbol{\omega}_k^{[d_1]}$ and $\boldsymbol{\omega}_k^{[d_2]}$

provide approximations to the d_1 and the d_2 derivatives of the coefficient function $\alpha_k(\cdot)$, respectively, which we denote as $\alpha_k^{(d_1)}(\cdot)$ and $\alpha_k^{(d_2)}(\cdot)$. Therefore, enforcing sparsity in $\boldsymbol{\omega}_k$ will constrain $\alpha_k^{(d_1)}(\cdot)$ and $\alpha_k^{(d_2)}(\cdot)$ to be zero at most of time points. Similarly, if we define $\boldsymbol{\gamma}_{kg} = \mathbf{A}_q\boldsymbol{\zeta}_k$ and denote by $\boldsymbol{\gamma}_{kg} = ({\boldsymbol{\gamma}_{kg}^{[d_1]}}^\top, {\boldsymbol{\gamma}_{kg}^{[d_2]}}^\top)^\top$, then we can derive the same regularization for the coefficient functions $\beta_{kg}(\cdot)$. From the definitions of $\boldsymbol{\omega}_k$ and $\boldsymbol{\gamma}_{kg}$ we can easily get the following relations:

$$\begin{cases} \boldsymbol{\zeta}_k = {\mathbf{A}_p^{[d_1]}}^{-1}\boldsymbol{\omega}_k^{[d_1]} \text{ and } \boldsymbol{\omega}_k^{[d_2]} = \mathbf{A}_p^{[d_2]}{\mathbf{A}_p^{[d_1]}}^{-1}\boldsymbol{\omega}_k^{[d_1]} & \text{(8a)} \\ \boldsymbol{\eta}_{kg} = {\mathbf{A}_q^{[d_1]}}^{-1}\boldsymbol{\gamma}_{kg}^{[d_1]} \text{ and } \boldsymbol{\gamma}_{kg}^{[d_2]} = \mathbf{A}_q^{[d_2]}{\mathbf{A}_q^{[d_1]}}^{-1}\boldsymbol{\gamma}_{kg}^{[d_1]}. & \text{(8b)} \end{cases}$$

Plugging the relation (8a) into (6) one gets the following new representation for $\pi_k(\mathbf{r}_i;\boldsymbol{\xi})$

$$\pi_k(\mathbf{s}_i;\mathbf{w}) = \frac{\exp\{\alpha_{k,0}+\mathbf{s}_i^\top\boldsymbol{\omega}_k^{[d_1]}\}}{1+\sum_{k'=1}^{K-1}\exp\{\alpha_{k',0}+\mathbf{s}_i^\top\boldsymbol{\omega}_{k'}^{[d_1]}\}}, \tag{9}$$

where $\mathbf{s}_i = ({\mathbf{A}_p^{[d_1]}}^{-1})^\top\mathbf{r}_i$ is now the new design vector and $\mathbf{w} = (\alpha_{1,0}, {\boldsymbol{\omega}_1^{[d_1]}}^\top, \ldots, \alpha_{K-1,0}, {\boldsymbol{\omega}_{K-1}^{[d_1]}}^\top)^\top$, with $(\alpha_{K,0}, {\boldsymbol{\omega}_K^{[d_1]}}^\top)^\top$ a null vector, is the unknown parameter vector of the gating network. Similarly, plugging (8b) into (7) one obtains the new representation fro $P(y_i|\mathbf{x}_i;\boldsymbol{\theta}_k)$:

$$P(y_i|\mathbf{v}_i;\boldsymbol{\Gamma}_k) = \prod_{g=1}^{G}\left[\frac{\exp\left\{\beta_{kg,0}+\mathbf{v}_i^\top\boldsymbol{\gamma}_{kg}^{[d_1]}\right\}}{1+\sum_{g'=1}^{G-1}\exp\left\{\beta_{kg',0}+\mathbf{v}_i^\top\boldsymbol{\gamma}_{k'g}^{[d_1]}\right\}}\right]^{y_{ig}}, \tag{10}$$

in which, $\mathbf{v}_i = ({\mathbf{A}_q^{[d_1]}}^{-1})^\top\mathbf{x}_i$ is now the new design vector and $\boldsymbol{\Gamma}_k = (\beta_{kg,0}, {\boldsymbol{\gamma}_{k'g}^{[d_1]}}^\top)^\top$ is the unknown parameter vector of the expert network. Finally, gathering the gating network (9) and the expert network (10), the iFME model for classification is given by $P(y_i|X_i(\cdot);\boldsymbol{\Upsilon}) = \sum_{k=1}^{K}\pi_k(\mathbf{s}_i;\mathbf{w})P(y_i|\mathbf{v}_i;\boldsymbol{\Gamma}_k)$, where $\boldsymbol{\Upsilon} = (\mathbf{w}^\top, \boldsymbol{\Gamma}_1^\top, \ldots, \boldsymbol{\Gamma}_K^\top)^\top$ is the unknown parameter vector to be estimated. We perform penalized MLE by penalizing the ML via a Lasso penalization on the derivative coefficients $\boldsymbol{\omega}_k$'s and $\boldsymbol{\gamma}_{kg}$'s of the form $\text{Pen}_{\chi,\lambda}(\boldsymbol{\Upsilon}) = \chi\sum_{k=1}^{K-1}\|\boldsymbol{\omega}_k\|_1 + \lambda\sum_{k=1}^{K}\sum_{g=1}^{G-1}\|\boldsymbol{\gamma}_{kg}\|_1$, with χ and λ regularization constants. The estimation is performed by using an adaptation to this classification context of the EM algorithm developed in Chamroukhi et al. (2022). The only difference resides in the maximization w.r.t. the expert network parameters $\boldsymbol{\Gamma}_k$.

3 Numerical results

We conducted experiments by considering a $G = 3$-class classification problem with a $K = 2$-component FME model. The simulation protocol will be detailed during the presentation due to lack of space here. The classification results obtained with the described algorithms FME-EM, FME-EM-Lasso and iFME-EM, as well as with functional multinomial logistic regression (FMLR), are given in Table 1 and show higher classification performance of the iFME-EM approach. We then applied the two algorithms allowing

Model	Correct Classification Rate	
	Noise level: $\sigma_\delta^2 = 1$	Noise level: $\sigma_\delta^2 = 5$
FME-EM	$.8560_{(.0199)}$	$.8474_{(.0196)}$
FME-EM-Lasso	$.9332_{(.0104)}$	$.9178_{(.0142)}$
iFME-EM	$\mathbf{.9346_{(.0108)}}$	$\mathbf{.9219_{(.0127)}}$
FMLR	$.7951_{(.0249)}$	$.7922_{(.0270)}$

Table 1: Correct classification rates obtained on testing data. The reported values are averages on 100 samples with standard errors in parentheses.

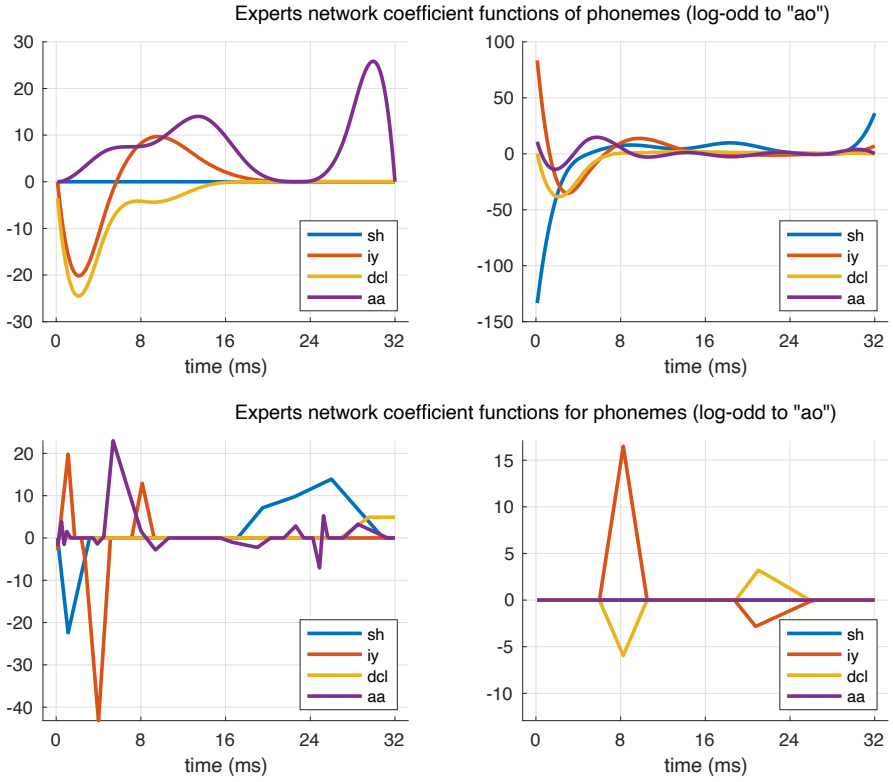


Figure 1: Results of (top) FME-EM-Lasso and (bottom) iFME-EM on phoneme data.

for sparsity (FME-EM-Lasso and iFME-EM) to the well-known phoneme data (Hastie et al., 1995). The data consists of $n = 1000$ log-periodogram recordings of length 256 each, used here as the univariate functional predictors, of five phonemes (the corresponding class labels). The obtained averaged correct classification rate for the two approaches are more than 0.94 in mean. Figure 1 shows the estimated coefficient functions for the expert network $\widehat{\beta}_{kg}(t)$ as functions of sampling time t, obtained by FME-EM-Lasso (top) and the iFME-EM (bottom); Here the iFME-EM is fitted with constraints on the zero and the second derivatives of the coefficients functions. The results show clearly sparse

and piece-wise-linear gating and experts functions when using the iFME-EM approach.

Acknowledgement: This research is supported by ANR SMILES ANR-18-CE40-0014.

References

Chamroukhi, F., Pham, T. N., Hoang, V. H., and McLachlan, G. J. (2022). Functional mixtures-of-experts. *ArXiv preprint arXiv:2202.02249*.

Chamroukhi, F., Samé, A., Govaert, G., and Aknin, P. (2009). Time series modeling by a regression approach based on a latent process. *Neural Networks*, 22(5-6):593–602.

Chamroukhi, F., Trabelsi, D., Mohammed, S., Oukhellou, L., and Amirat, Y. (2013). Joint segmentation of multivariate time series with hidden process regression for human activity recognition. *Neurocomputing*, 120:633–644.

Chen, K., Xu, L., and Chi, H. (1999). Improved learning algorithms for mixture of experts in multiclass classification. *Neural Networks*, 12(9):1229–1252.

Gormley, I. C. and Murphy, T. B. (2010). A mixture of experts latent position cluster model for social network data. *Statistical Methodology*, 7:385–405.

Gutta, S., Huang, J. R., Phillips, P. J., and Wechsler, H. (2000). Mixture of experts for classification of gender, ethnic origin, and pose of human faces. *IEEE transactions on neural networks*, 11 4:948–60.

Hastie, T., Buja, A., and Tibshirani, R. (1995). Penalized Discriminant Analysis. *Annals of Statistics*, 23:73–102.

Huynh, T. and Chamroukhi, F. (2019). Estimation and feature selection in mixtures of generalized linear experts models.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87.

James, G. M. (2002). Generalized linear models with functional predictor variables. *Journal of the Royal Statistical Society Series B*, 64:411–432.

James, G. M., Wang, J., and Zhu, J. (2009). Functional linear regression that's interpretable. *Annals of Statistics*, 37(5A):2083–2108.

Müller, H.-G., Stadtmüller, U., et al. (2005). Generalized functional linear models. *Annals of Statistics*, 33(2):774–805.

Nguyen, H. D. and Chamroukhi, F. (2018). Practical and theoretical aspects of mixture-of-experts modeling: An overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, pages e1246–n/a.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*. Springer Series in Statistics. Springer.

Yümlü, M. S., Gürgen, F. S., and Okay, N. (2003). Financial time series prediction using mixture of experts. In *ISCIS*.

Zeevi, A. J., Meir, R., and Adler, R. J. (1996). Time series prediction using mixtures of experts. In *Proceedings of NIPS'96*, page 309–315. MIT Press.

Ensemble Methods for Data Analytics

Camila Fernández [1,2] & Chung Shue Chen[2] & Pierre Gaillard[3] & Alonso Silva[4]

[1] *Sorbonne Université, camila.fernandez@nokia.com*
[2] *Nokia Bell Labs, chung_shue.chen@nokia-bell-labs.com*
[3] *INRIA, pierre.gaillard@inria.fr*
4 *Safran Tech, alonso.silva-allende@safrangroup.com*

Résumé. L'analyse de survie est une branche de la statistique qui a suscité l'intérêt de divers chercheurs en raison de son vaste champ d'application, par exemple, la maintenance prédictive, les soins de santé, la prédiction du taux de désabonnement, etc. Dans cette communication, nous présentons une façon simple d'agréger les méthodes d'analyse de survie qui surpasse en moyenne la performance de chaque prédicteur individuel.

Mots-clés. Analyse de survie, apprentissage automatique, statistique, agrégation.

Abstract. Time-to-event analysis is a branch of statistics that have been in many researchers' interest due to its wide application field, such as predictive maintenance, healthcare, customer churn prediction, among others. In this communication we present a simple way to aggregate time-to-event analysis methods that outperform on average the performance of every single predictor.

Keywords. Time-to-event, machine learning, statistics, aggregation.

1 Introduction

The objective of time-to-event analysis is to predict the time until a certain critical event occurs, such as the death of populations, the recurrence of a disease, the failure of a machine, among others. One of the things that difference time-to-event analysis from other classical regression problems is a phenomenon called right censoring. Right censoring raises from the fact that not all the individuals in the dataset have reached the critical event time by the end of the observation period, and so, the actual event time for some samples are unknown. It is essential to include this phenomenon in the methods to avoid bias and underestimation. In addition, it is important to remark that the censored time will be a lower bound for the event time.

Aggregating methods is a technique that has been increasingly used due to that on many occasions it can lead to an increment of accuracy and robustness of algorithms. We implemented a bunch of methods with diverse structures (parametric, semi-parametric and machine learning): Cox Proportional Hazards(1972), Gradient Boosting(2001), Aalen Additive(1989), Weibull AFT(2018), Random Survival Forest(2008) and DeepSurv(2018) from

three different python libraries: Lifelines(2020), Scikit-survival(2019) and Pysurvival(2019). In this paper, we propose a way to combine these methods by optimizing the Integrated Brier Score(1950) with a gradient descent algorithm. This aggregation outperforms on average the performance of every single predictor.

1.1 Mathematical Setting

Let us set $(x, y, \delta) \in \mathcal{X} \times \mathbb{R}_+ \times \{0, 1\}$ to be a point in the dataset where $\mathcal{X} \subset \mathbb{R}^d$ will be the features space, $y \in \mathbb{R}_+$ the observed time, and $\delta \in \{0, 1\}$ the censored indicator which is 1 if y corresponds to an event time and 0 if y is a censored time. We call T and C the non-negative random variables of the event and censored times. Then, we can define the observed time to be $Y = \min\{T, C\}$ and the censored indicator as $\delta = \mathbb{1}\{T \leq C\}$.

Supposing we have N samples, the objective will be to estimate the survival probability function:

$$S(t|x) = \mathbb{P}(T > t|X = x).$$

Many of the implemented methods estimate the survival probability function by assuming a particular shape in the hazard function, which is defined as following

$$h(t|x) = -\frac{\partial}{\partial t} log(S(t|x)),$$

that can also be written as the instantaneous risk

$$h(t|x) = \mathbb{P}(T = t|T \geq t, X = x).$$

1.2 Integrated Brier Score

The integrated Brier score (IBS), proposed by G. Brier (1950), provides an overall calculation of the method's performance at all the observed times. It is an estimator for the minimum square error (MSE). Given $\hat{S}$ an estimator of the survival probability function S, the MSE will be

$$MSE(\hat{S}) = \int_0^\tau \mathbb{E}[(\mathbb{1}\{T > t\} - \hat{S}(t|x))^2]dt,$$

where τ is a user selected horizon time. As we only observe y we cannot know T for all the samples and, therefore, we need to define an alternative score that considers the censorship. In order to do this, we need first to define the survival probability for the censored distribution $S_C(t|x) = \mathbb{P}(C > t|X = x)$ which will be computed by only considering the censored individuals. The integrated Brier score will be defined as follows

$$IBS(\hat{S}) = \frac{1}{\tau N} \sum_{i=1}^{N} \int_0^\tau W_i(t) \left(\mathbb{1}\{y_i > t\} - \hat{S}(t|x_i)\right)^2 dt$$

Algorithm 1 Exponential Gradient Descent

Input: m number of iteration, $\eta > 0$ learning rate
Initialization: $\lambda(0) = (1/n, \ldots, 1/n)$
for each iteration $i = 1, \ldots, m$ **do**
Define $Z_i = \sum_{j=1}^{n} \lambda_j(i) \exp(-\eta Df_j)$ where $Df_j = \frac{\partial IBS(\hat{S})}{\partial \lambda_j}$
Update $\lambda_k(i+1) = \frac{\lambda_k(i) \exp(-\eta Df_k)}{Z_i}$ for all $k = 1, \ldots, n$
end for

where

$$W_i(t) = \frac{\delta_i \mathbb{1}\{y_i \leq t\}}{\hat{S}_C(y_i|x_i)} + \frac{\mathbb{1}\{y_i > t\}}{\hat{S}(t|x_i)}.$$

This score is a consistent estimator of MSE (Gerds and Schummacher(2006)). We use in this paper, an implementation of the integrated Brier score provided by Chen (2019).

2 Ensemble Methods

Aggregation techniques can improve forecast accuracy by consider the advantages of many models at the same time. Let $\hat{S}_1, \ldots, \hat{S}_n$ to be estimators of S, particularly, in our case we are going to consider 6 predictors corresponding to the 6 implemented models mentioned previously in the introduction. Let set the aggregated model to be

$$\hat{S} = \sum_{i=1}^{n} w_i \hat{S}_i.$$

The weights will be optimized with an exponential gradient descent algorithm described in Algorithm 1 in order to minimize the integrated Brier score.

3 Results

In order to compare these methods, we implemented the different models and the aggregation model to three datasets, Primary Biliary Cirrhosis from Mayo Clinic (PBC, 2000), German Breast Cancer Study Group (GBCSG, 1994) and Kaggle Telecom Churn (TLCM, 2008). We define an overall score for each method in the following way

$$IBS_{cox} = \frac{IBS_{cox}(PBC) + IBS_{cox}(GBCSG) + IBS_{cox}(TLCM)}{3}.$$

Figure 1 shows the performance of the overall score for each method. We observe that the ensemble method proposed in the previous section outperforms the single predictors by 3.4%.

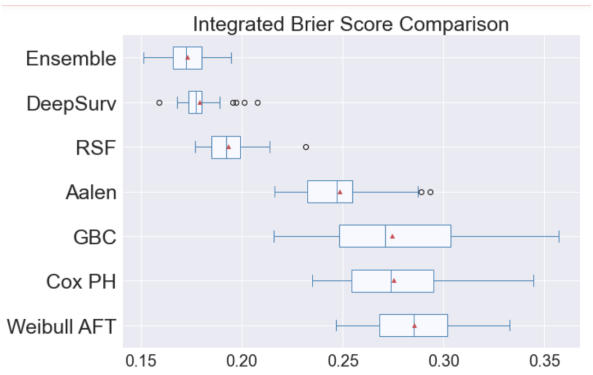


Figure 1: Overall score

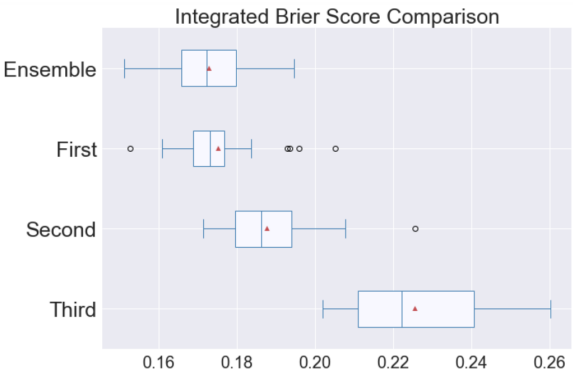


Figure 2: Ranking overall score

We define a second overall score based on the ranking position of the methods for each dataset. Given the score of each dataset's best predictor $IBS_{first}(PBC)$, $IBS_{first}(GBCSG), IBS_{first}(TLCM)$, we define the first overall score as

$$IBS_{first} = \frac{IBS_{first}(PBC) + IBS_{first}(GBCSG) + IBS_{first}(TLCM)}{3},$$

and similarly IBS_{second} and IBS_{third}. Figure 2 shows the results of this score where we observe that the ensemble method performs 1.14% better than the average score of the best method of the datasets.

In conclusion, both scores show an increment of performance by aggregating methods whose are, in addition, independent of the dataset; this adds robustness to our proposed model.

Bibliographie

Aalen, O. O. (1989). A linear regression model for the analysis of life times. *Statistics in Medicine*, vol. 8, pp. 907–925.

Brier, G. (1950), Verification of forecasts expressed in terms of probability. *Monthly weather review*, vol. 78, no. 1, pp. 1-3.

Chen, G. (2019). Nearest neighbor and kernel survival analysis: Nonasymptotic error bounds and strong consistency rates. arXiv preprint arXiv:1905.05285.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society*, Series B. 34 (2): 187–220.

Davidson-Pilon, C. (2019). Lifelines: Survival analysis in Python. *Journal of Open Source Software*, vol. 4, no. 40, pp. 1317.

Fotso, S. (2019). Pysurvival: open source package for survival analysis modeling. *https://www.pysurvival.io*

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics,* pp.1189-1232
Gerds, T., Schummacher, M. (2006). Consistent estimation of the expected Brier score in general survival models with right censored event times. *Biometrical Journal*, pp. 1029-1040.
IBM (2008). Kaggle Telco Customer Churn. *IBM Cognos Analytics*
Ishwaran, H., Kogalur, U. B., Blackstone, E. H., Lauer, M. S. (2008). Random survival forests. *Annals of Applied Statistics*, vol. 2, no. 3, pp. 841–860.
Katzman J., Shaham U., Cloninger A., Bates J., Jiang T., and Kluger Y. (2018). DeepSurv: Personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Medical Research Methodology.*
Brier, G. (1950), Verification of forecasts expressed in terms of probability. *Monthly weather review*, vol. 78, no. 1, pp. 1-3.
Pölsterl, S. (2019). Scikit-survival:v0.11, doi:10.5281/zenodo.3352342.
Schumacher, M., Basert, G., Bojar, H., Huebner, K., Olschewski, M., Sauerbrei, W., Schmoor, C., Beyerle, C., Neumann, R. L. and Rauschecker H. F., (1994), Randomized 2 x 2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. *Journal of Clinical Oncology*, 12(10), pp. 2086–2093.
Therneau, T., and Grambsch, P. (2000). Modeling Survival Data: Extending the Cox Model, Springer-Verlag, New York. ISBN: 0-387-98784-3.
Liu, E. (2018), Using Weibull accelerated failure time regression model to predict survival time and life expectancy. *BioRxiv.*

Consensual aggregation on randomly projected high-dimensional features of predictions for regression

Sothea Has

LPSM, Sorbonne Université, sothea.has@lpsm.paris

Résumé. Dans cet exposé, nous présentons une étude d'une agrégation consensuelle basée sur un noyau, mise en œuvre sur un vecteur de prédictions de grande dimension projetées de manière aléatoire pour la régression. Le schéma d'agrégation est composé de deux étapes : les prédictions données par un grand nombre d'estimateurs de régression sont projetées de manière aléatoire dans un sous-espace plus petit en utilisant le lemme de Johnson-Lindenstrauss (J-L) dans la première étape, puis une méthode d'agrégation consensuelle basée sur un noyau est mise en œuvre dans la deuxième étape. Nous montrons théoriquement que les performances du schéma d'agrégation sont proches de celles de la méthode d'agrégation mise en œuvre sur les prédictions originales de grande dimension, avec une grande probabilité. Ensuite, nous montrons numériquement que la méthode d'agrégation consensuelle conserve ses performances sur les prédictions fortement corrélées données par différents types de régresseurs, construits simplement sans sélection de modèle ni validation croisée. L'efficacité du schéma d'agrégation est illustrée sur plusieurs types de jeux de données synthétiques et réels.

Mots-clés. Agrégation consensuelle, projection aléatoire, régression.

Abstract. In this talk, we present a study of a kernel-based consensual aggregation on randomly projected high-dimensional features of predictions for regression. The aggregation scheme is composed of two steps: the high-dimensional features of predictions given by a large number of regression estimators are randomly projected into a smaller subspace using Johnson-Lindenstrauss Lemma (J-L) in the first step, then a kernel-based consensual aggregation is implemented in the second step. We theoretically show that the performance of the aggregation scheme is close to the performance of the aggregation on the original high-dimensional features, with high probability. Then, we numerically show that the consensual aggregation method upholds its good performance on highly correlated features of predictions given by different types of regression estimators, plainly constructed without model selection or cross-validation. The efficiency of the aggregation scheme is illustrated on several types of synthetic and real datasets.

Keywords. Consensual aggregation, random projection, regression.

1 Introduction

Consensual aggregation methods aim at combining a given number of predictors based on the consensus level or similarity between predictions of data points. The method was first introduced in classification framework by Mojirsheibani (1999), followed by its kernel-based settings, for instance, in Mojirsheibani (2000) and Mojirsheibani and Kong (2016). Later, the aggregation method of Mojirsheibani (1999) was configured to regression framework by Biau et al. (2016), followed by other related works such as a regular kernel-based setting by Has (2021), and consensual aggregation based on both inputs and predictions by Fischer and Mougeot (2019).

Let $\mathcal{D}_n = \{(X_1, Y_1), ..., (X_n, Y_n)\}$ be the training data of n iid copies of $\mathbb{R}^d \times \mathbb{R}$-valued generic couple (X, Y), and let $\|z\|$ be the Euclidean norm defined on the corresponding Euclidean space of z. Suppose that the individual regressors $r_1, ..., r_M$ are constructed independently of $\mathcal{D}_n$, then the aggregation g_n computed at any point $x \in \mathbb{R}^d$ takes the following form:

$$g_n(\mathbf{r}(x)) = \sum_{i=1}^{n} W_{n,i}(x) Y_i. \tag{1}$$

Here, $\mathbf{r}(x) = (r_1(x), ..., r_M(x))$ denotes the vector of predictions at point x, and the weights $W_{n,i}(x)$ are defined by

$$W_{n,i}(x) = \frac{K_h(\mathbf{r}(x) - \mathbf{r}(X_i))}{\sum_{j=1}^{n} K_h(\mathbf{r}(x) - \mathbf{r}(X_j))}, \ i = 1, ..., n, \tag{2}$$

for some kernel function K such that $K_h(z) = K(z/h)$ for some parameter $h > 0$ with the convention $0/0 = 0$. Note that the aggregation method of Biau et al. (2016) corresponds to the naive kernel function: $K(z) = \prod_{k=1}^{M} \mathbb{1}_{\{|z_k|<1\}}$ for any $z \in \mathbb{R}^M$.

It is shown in Biau et al. (2016) and Has (2021) that the aggregation methods inherit the consistency of the basic consistent estimators of the combination, which is also confirmed by several numerical experiments. Based on this result, one may be interested in the limitation of the method in terms of the number of basic estimators. In this study, we are interested in the behavior of the aggregation method on a large number of basic regression estimators. However, it is well-known that working in high-dimensional spaces is a challenging task due to highly computational cost and curse of dimensionality. To overcome such problems, the high-dimensional features of predictions are first randomly projected into a smaller subspace of dimension $m < M$ using Johnson-Lindenstrauss Lemma, then the projected features are aggregated. Johnson and Lindenstrauss showed that one can randomly project a finite set of high-dimensional Euclidean spaces into a smaller subspace, preserving pairwise distances of the original data points with high probability (see, for example, Johnson et al. (1986) and Dasgupta and Gupta (2003)). This method is suitable for our setting not only because of its theoretical property, but also because of its computational efficiency, which is as simple as generating a random matrix and performing a matrix multiplication.

1.1 Johnson-Lindenstrauss Lemma

In the sequel, we assume that M is large and the M-dimensional vector of predictions $\mathbf{r}(.)$ is a row vector. Let $G = (G_{i,j})_{1\leq i\leq M, 1\leq j\leq m}$ be an $M \times m$ random projection matrix, where the entries are independent centered Gaussian random variable with variance $1/m$ denoted by $G_{i,j} \sim \mathcal{N}(0, 1/m)$. Let $\tilde{\mathbf{r}}(x) = \mathbf{r}(x)G \in \mathbb{R}^m$, denote the projected vector of predictions of x. Thus, given all the machines $(r_i)_{1\leq i\leq M}$, the training input $(X_j)_{1\leq j\leq n}$ of size n, and any data point $x \in \mathbb{R}^d$, for any $\delta \in (0, 1)$, one has

$$\mathbb{P}\Big(\Big|\frac{\|\tilde{\mathbf{r}}(x) - \tilde{\mathbf{r}}(X_j)\|^2}{\|\mathbf{r}(x) - \mathbf{r}(X_j)\|^2} - 1\Big| \leq \delta, \forall j = 1, ..., n\Big) \geq 1 - 2n \exp\Big(-m(\delta^2/2 - \delta^3/3)/2\Big). \quad (3)$$

Note that the Euclidean norms in the left-hand side of (3) are not defined on the same space. As the original and the corresponding projected vector of predictions $\mathbf{r}(.)$ and $\tilde{\mathbf{r}}(.)$ are of dimension M and m respectively, thus the norms in the denominator and numerator of (3) are defined on $\mathbb{R}^M$ and $\mathbb{R}^m$ respectively.

1.2 The aggregation scheme

We are now in a position to formally describe the aggregation scheme. Given all the basic machines $(r_i)_{1\leq i\leq M}$, the training data $\mathcal{D}_n$, to predict the response value of any new observation $x \in \mathbb{R}^d$, we first compute the projected vector of x and the training data which is simply done as follows:

$$\begin{aligned}
\mathbf{r}(.)G &= \begin{pmatrix} r_1(x) & \dots & r_M(x) \\ r_1(X_1) & \dots & r_M(X_1) \\ \vdots & \ddots & \vdots \\ r_1(X_n) & \dots & r_M(X_n) \end{pmatrix} \times \begin{pmatrix} G_{11} & \dots & G_{1m} \\ G_{21} & \dots & G_{2m} \\ \vdots & \ddots & \vdots \\ G_{M1} & \dots & G_{Mm} \end{pmatrix} \\
&= \begin{pmatrix} \tilde{r}_1(x) & \tilde{r}_2(x) & \dots & \tilde{r}_m(x) \\ \tilde{r}_1(X_1) & \tilde{r}_2(X_1) & \dots & \tilde{r}_m(X_1) \\ \vdots & \vdots & \vdots & \vdots \\ \tilde{r}_1(X_n) & \tilde{r}_2(X_n) & \dots & \tilde{r}_m(X_n) \end{pmatrix}_{(n+1)\times m} \\
&= \tilde{\mathbf{r}}(.).
\end{aligned}$$

Then, the aggregation (1) is implemented on this projected features of predictions. More precisely, the prediction of the response value of x is defined explicitly from (1) and (2) by:

$$g_n(\tilde{\mathbf{r}}(x)) = \frac{\sum_{i=1}^n Y_i K_h(\|\tilde{\mathbf{r}}(x) - \tilde{\mathbf{r}}(X_i)\|)}{\sum_{j=1}^n K_h(\|\tilde{\mathbf{r}}(x) - \tilde{\mathbf{r}}(X_j)\|)}, \quad (4)$$

where in this case, K is a one dimensional kernel function with the same convention as described above.

2 Theoretical performance

In this part, we study the performance of the aggregation scheme compared to the full aggregation method implemented on the original features of predictions. In this case, we work with exponential kernel function $K(t) = \exp(-t^\alpha/\sigma)$ for some $\sigma > 0$ and $\alpha \geq 1$. One can show that the performance of the aggregation scheme on randomly projected features is close to the aggregation on the original features of predictions, with high probability. This result is given in the following theorem.

Theorem 1 *Assume that all the machines $r_1, r_2, ..., r_M$ and the response variable Y are bounded almost surely by R_0, thus for any $h, \varepsilon > 0, n \geq 1$, and for any $\delta \in (0, 1)$, with the choice of m satisfying:*

$$m \geq C_1 \frac{\log[2/(1 - \sqrt[n]{1-\delta})]}{h^{2\alpha}\varepsilon^2}, \text{ with } C_1 = 3(2+\alpha)^2(2R_0)^{2(1+\alpha)}/\sigma^2,$$

one has:

$$\mathbb{P}\Big(|g_n(\mathbf{r}(X)) - g_n(\tilde{\mathbf{r}}(X))| > \varepsilon\Big) \leq \delta.$$

The probability in Theorem 1 is computed under the laws of X, the training data $\mathcal{D}_n = \{(X_i, Y_i)_{i=1}^n\}$ and the random projection matrix G. This result is independent of M and can be viewed as a loss of projecting the features of predictions into a smaller subspace of dimension m. Note that in this result, the constant C_1 depends on R_0, which is in practice can be scaled to be, for example, less then 1. Therefore, the constant $C_1 \approx 12$ for Gaussian kernel, and the lower bound of m is roughly of order:

$$O\Big(\frac{\log(2n/\delta)}{\varepsilon^2 h^{2\alpha}}\Big),$$

for large n and small δ.

3 Numerical performance

In this part, we illustrate two important aspects of the aggregation method: one is about the performance of the aggregation method on a large (highly correlated) features of predictions, and the second one is to confirm the theoretical result stated in the previous part. In the numerical experiment, five types of predictive models are plainly constructed, where for each type, several models are trained by varying the value of the key parameters without any model selection or cross-validation. By construction, we expect that the high-dimensional features of predictions ($M = 1000$) are quite redundant. Despite being implemented on such highly correlated features, the numerical study shows that the aggregation method upholds its competitive performance (biases towards the best machine).

This is robust in a sense that it can flexibly merge different strongly redundant features of predictions without any model selection or validation methods. Secondly, several values of projected dimension $m = 2, 3, .., 9, 100, 200, ..., 900$ are studied, and the numerical experiment confirms that the performance of the aggregation scheme on much smaller projected subspaces are almost preserved. Moreover, with similar accuracy, the aggregation scheme is very efficient in computational speed (from 3 to 20 times faster according to the optimization parameters and the size of the data). Figure 1 highlights the performances of the aggregation method on some public datasets long with its computational times over 30 independent runs.

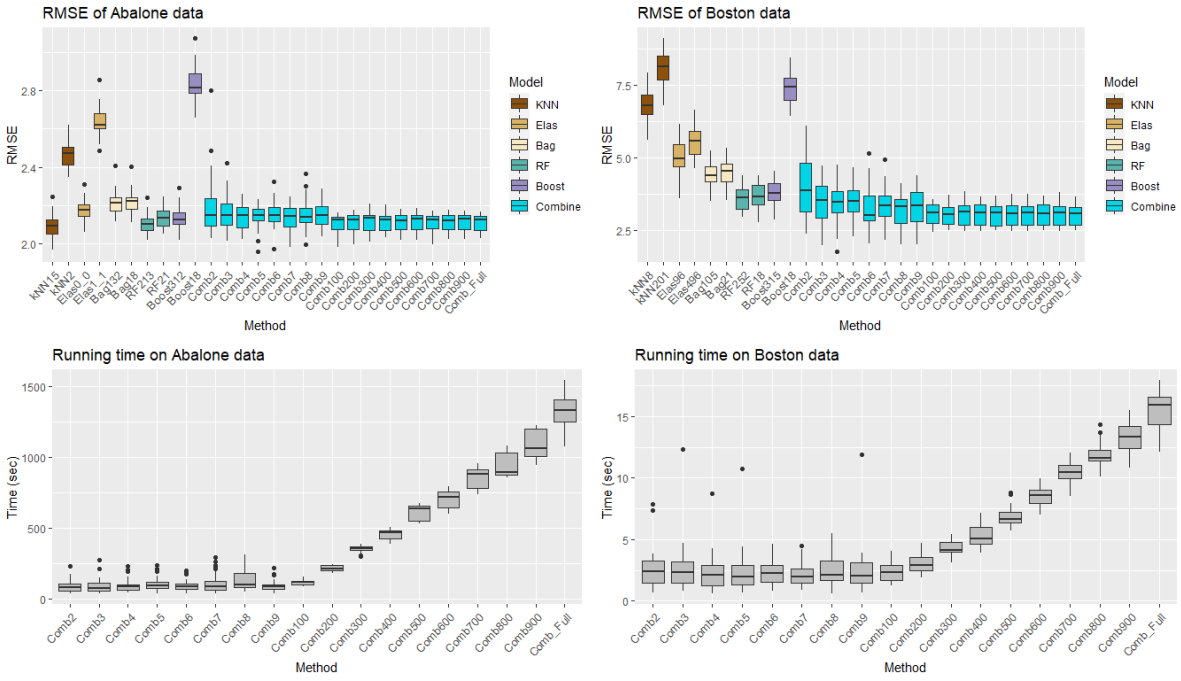


Figure 1: Boxplots of root mean square errors (RMSE) and the running times of the aggregation scheme. The methods are implemented on **Abalone** of Dua and Graff (2017) and **Boston** datasets from `MASS` library of `R` software (Brian et al. (2021)). In the upper part from left to right, the first ten boxplots are the best and the worst performance of k nearest neighbors (**kNN**), Elastic Net (**Elas**), Bagging (**Bag**), Random Forest (**RF**) and Gradient Boosting (**Boost**) regressors, respectively. The last eighteen boxplots represent the performances of the aggregation methods **Comb**m ($m = 2, 3, ..., 9, 100, 200, ..., 900$ and full method **Comb_Full**). The lower part contains boxplots of the corresponding running times of all the aggregation methods.

References

Gérard Biau, Aurélie Fischer, Benjamin Guedj, and James D. Malley. COBRA: a combined regression strategy. *Journal of Multivariate Analysis*, 146:18–28, 2016.

Ripley Brian, Venables Bill, M. Bates Douglas, Hornik Kurt, Gebhardt Albrecht, and Firth David. Mass: Support functions and datasets for venables and ripley's mass, 2021. URL `https://CRAN.R-project.org/package=MASS`.

Sanjoy Dasgupta and Anupam Gupta. An elementary proof of a theorem of johnson and lindenstrauss. *Random Structures and Algorithms*, 22(1):60–65, 2003. doi: https://doi.org/10.1002/rsa.10073.

Dheeru Dua and Casey Graff. UCI machine learning repository: Abalone data set, 2017.

A. Fischer and M. Mougeot. Aggregation using input-output trade-off. *Journal of Statistical Planning and Inference*, 200:1–19, May 2019.

Sothea Has. A Kernel-based Consensual Aggregation for Regression. working paper or preprint, April 2021. URL `https://hal.archives-ouvertes.fr/hal-02884333`.

William B. Johnson, Joram Lindenstrauss, and Gideon Schechtman. Extensions of lipschitz maps into banach spaces. *Israel Journal of Mathematics*, 54:129–138, 1986. ISSN 1565-8511. doi: https://doi.org/10.1007/BF02764938.

Majid Mojirsheibani. Combined classifiers via discretization. *Journal of the American Statistical Association*, 94(446):600–609, June 1999.

Majid Mojirsheibani. A kernel-based combined classification rule. *Journal of Statistics and Probability Letters*, 48(4):411–419, July 2000.

Majid Mojirsheibani and Jiajie Kong. An asymptotically optimal kernel combined classifier. *Journal of Statistics and Probability Letters*, 119:91–100, 2016.

Prédictions Conformelles Adaptatives pour les Séries Temporelles

Margaux Zaffran [1,2,3] & Aymeric Dieuleveut [3] & Olivier Féron [1,4] & Yannig Goude [1,5] & Julie Josse [2,6]

[1] *EDF R&D, Palaiseau, France, olivier-2.feron,yannig.goude@edf.fr*
[2] *INRIA Sophia-Antipolis, Montpellier, France, margaux.zaffran,julie.josse@inria.fr*
[3] *CMAP, Ecole Polytechnique, Palaiseau, France, aymeric.dieuleveut@polytechnique.edu*
[4] *FiME, Palaiseau, France*
[5] *LMO, Université Paris-Saclay, Orsay, France*
[6] *IDESP, Montpellier, France*

Résumé. La quantification d'incertitudes des modèles prédictifs est cruciale dans les problèmes de prise de décision. La prédiction conformelle est une approche générale et théoriquement solide. Cependant, elle nécessite des données échangeables, excluant les séries temporelles. Dans cet exposé, nous soutenons que l'Inférence Conformelle Adaptative (ACI, Gibbs et Candès, 2021), développée pour les séries temporelles avec des changements de distribution, est une bonne procédure pour les séries temporelles avec une dépendance générale. Nous analysons théoriquement l'impact du taux d'apprentissage sur son efficacité dans le cas échangeable et auto-régressif. Nous proposons une méthode sans paramètre, AgACI, reposant de manière adaptative sur ACI basée sur l'agrégation d'experts en ligne. Nous menons une comparaison avec des méthodes de la littérature conformelle sur de nombreuses simulations. Les résultats numériques plaident pour l'utilisation d'ACI pour les séries temporelles. Finalement, nous appliquons ces méthodes à la prévision des prix de l'électricité, où AgACI fournit des intervalles de prédiction efficaces.

Mots-clés. Séries temporelles, quantification d'incertitudes, prédiction conformelle, agrégation d'experts en ligne

Abstract. Uncertainty quantification of predictive models is crucial in decision making problems. Conformal prediction is a general and theoretically sound answer. However, it requires exchangeable data, excluding time series. In this talk, we argue that Adaptive Conformal Inference (ACI, Gibbs et Candès, 2021), developed for time series with distribution shift, is a good procedure for time series with general dependency. We theoretically analyze the impact of the learning rate on its efficiency in the exchangeable and autoregressive case. We propose a parameter-free method, AgACI, that adaptively builds upon ACI based on online expert aggregation. We conduct a comparison with methods from the conformal literature on many simulations. The numerical results argue for the use of ACI for time series. Finally, we apply these methods to electricity price forecasting, where AgACI provides efficient prediction intervals for day-ahead forecasting.

Keywords. Time series, uncertainty quantification, conformal prediction, online expert aggregation

1 Introduction

1.1 Prédiction conformelle

Dans de nombreux domaines tels que la santé ou l'énergie, l'absence de quantification d'incertitude des modèles prédictifs constitue un obstacle à l'adoption de puissantes méthodes d'apprentissage automatique. La prédiction conformelle (CP, Vovk, Gammerman et Saunders, 1999 ; Vovk, Gammerman et Shafer, 2005 ; Papadopoulos et al., 2002) permet de construire des intervalles prédictifs pour n'importe quel modèle prédictif qui sont *valides* – c'est-à-dire atteignant une couverture marginale nominale – en échantillon fini et sans aucune hypothèse de distribution, sauf que les données doivent être échangeables.

Ainsi, la CP a attiré l'attention récemment, favorisée par le développement de la *prédiction conformelle "split"* (SCP, Lei et al., 2018, reformulée de l'*inductive* CP, Papadopoulos et al., 2002). Supposons que nous disposons de n observations $(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$, $i \in [\![1, n]\!]$, réalisations de variables aléatoires $((X_i, Y_i))_{i=1}^n$, et que l'on cherche à prédire y_{n+1} en x_{n+1}. Étant donné un *taux d'erreur* $\alpha \in [0, 1]$ fixé par l'utilisateur (typiquement 0,1 ou 0,05), le but est de construire un intervalle de prédiction $\mathcal{C}_\alpha$ tel que :

$$\mathbb{P}\left\{Y_{n+1} \in \mathcal{C}_\alpha\left(X_{n+1}\right)\right\} \geq 1 - \alpha, \tag{1}$$

avec $\mathcal{C}_\alpha$ aussi petit que possible. Un intervalle est dit *valide* s'il vérifie (1) et *efficace* lorsqu'il est aussi petit que possible (Vovk, Gammerman et Shafer, 2005).

Pour ce faire, SCP divise d'abord les n points de l'ensemble d'apprentissage en deux ensembles $\mathrm{Tr}, \mathrm{Cal} \subset [\![1, \mathrm{n}]\!]$, pour créer un *"proper" ensemble d'apprentissage*, Tr, et un *ensemble de calibration*, Cal. Sur Tr, un modèle de régression $\hat{\mu}$ (choisi par l'utilisateur) est estimé, puis utilisé pour prédire sur Cal. Un *score de conformité* est appliqué pour évaluer la conformité entre les valeurs y_i de la calibration et les valeurs prédites, donnant $S_{\mathrm{Cal}} = \{(s_i)_{i \in \mathrm{Cal}}\}$. En régression, on utilise généralement la valeur absolue des résidus, c'est-à-dire $s_i = |\hat{\mu}(x_i) - y_i|$. Enfin, un quantile corrigé, le $(1 - \hat{\alpha})$-ème quantile de ces scores $\widehat{Q}_{1-\hat{\alpha}}(S_{\mathrm{Cal}})$, est évalué pour définir la taille de l'intervalle, qui, dans sa forme la plus simple, est centré sur la valeur prédite : $\mathcal{C}_\alpha\left(x_{n+1}\right) = \widehat{C}_{\hat{\alpha}}(x_{n+1}) := [\hat{\mu}(x_{n+1}) \pm \widehat{Q}_{1-\hat{\alpha}}(S_{\mathrm{Cal}})]$.

1.2 Contexte des séries temporelles

Cadre. Nous considérons T_0 observations $\left(x_1, y_1\right), \ldots, \left(x_{T_0}, y_{T_0}\right)$ dans $\mathbb{R}^d \times \mathbb{R}$. L'objectif est de prédire les réponses et de donner des intervalles de prédiction pour T_1 futures observations $x_{T_0+1}, \ldots, x_{T_0+T_1}$ séquentiellement : à chaque étape de prédiction $t \in [\![T_0 + 1, T_0 + T_1]\!]$, $y_{t-T_0}, \ldots, y_{t-1}$ ont été révélés. Ainsi, les données $\left(\left(x_{t-T_0}, y_{t-T_0}\right), \ldots, \left(x_{t-1}, y_{t-1}\right)\right)$ sont utilisées pour la construction de l'intervalle prédictif.

Littérature. La pierre angulaire des résultats de *validité* de SCP est l'hypothèse d'échangeabilité des données. Cependant, cette hypothèse n'est pas vérifiée par les séries tempo-

relles. Chernozhukov et al. (2018) étendent la théorie de la CP aux cas ergodiques, afin d'inclure les données dépendantes. Xu et Xie (2021) améliorent cette théorie et proposent un nouvel algorithme, EnbPI, adapté aux séries temporelles grâce à un aspect séquentiel.

Adaption de SCP pour les séries temporelles. De manière similaire à Wisniewski et al. (2020) et Kath et Ziel (2021), nous considérons une version en ligne de SCP en réajustant le modèle de régression sous-jacent et en recalibrant en utilisant les T_0 observations les plus récentes. De plus, pour tenir compte de manière appropriée de la structure temporelle des données, nous utilisons une *division séquentielle* : pour tout t, les indices temporels dans $\mathrm{Tr_t}$ sont plus petits que ceux de $\mathrm{Cal_t}$. Supprimer l'aléa vise à exclure les observations futures de $\mathrm{Tr_t}$, qui pourrait conduire à une sous-estimation des erreurs sur $\mathrm{Cal_t}$, donc à des intervalles plus petits avec sous-couverture. Cette procédure, appelée OSSCP (pour *Online Sequential Split Conformal Prediction*), est schématisée dans la Figure 1.

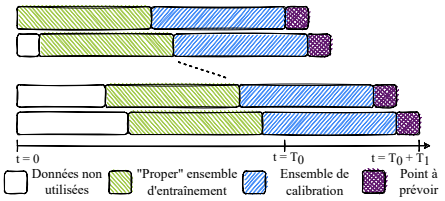


FIGURE 1 – Schéma d'OSSCP.

1.3 Inférence Conformelle Adaptative

Gibbs et Candès (2021) proposent l'inférence conformelle adaptative (ACI), dans le cas d'un nombre indéfini de changements de distribution de (X, Y). Elle est basée sur deux idées. D'une part, la totalité de la procédure (entraînement et calibration) est ré-effectuée à chaque instant en utilisant uniquement les dernières T_0 observations. D'autre part, un *niveau d'erreur effectif* α_t, mis à jour récursivement, est utilisé à la place du niveau cible α lors du calcul du quantile $\widehat{Q}_{1-\hat{\alpha}}(S_{\mathrm{Cal}})$. Soient $\alpha_1 = \alpha$, $\gamma \geq 0$, et pour $t \geq 1$:

$$\begin{cases} \widehat{C}_{\alpha_t}(x_t) & = [\hat{\mu}(x_t) \pm \widehat{Q}_{1-\alpha_t}(S_{\mathrm{Cal_t}})] \\ \alpha_{t+1} & = \alpha_t + \gamma \left(\alpha - \mathbb{1}\{y_t \notin \widehat{C}_{\alpha_t}(x_t)\}\right). \end{cases} \tag{2}$$

Un résultat asymptotique de *validité* pour n'importe quelle distribution de données est démontré dans Gibbs et Candès (2021). Toutefois, leur étude concerne uniquement les données temporelles avec des changements de distributions, telles que des ruptures dans les données financières. Notre exposé propose d'analyser cette méthode dans le cadre de données temporelles avec une dépendance générale, sans changement de distribution.

La première étape pour adapter une méthode aux séries temporelles dépendantes est de travailler en ligne, ce qui est le cas pour ACI. En outre, la mise à jour du niveau du quantile en fonction de l'erreur précédente implique qu'ACI pourrait s'adapter à un modèle ajusté qui n'a pas correctement saisi l'évolution temporelle, comme une tendance, une saisonnalité ou une dépendance au passé. Par conséquent, ACI est un candidat parfait pour la CP pour les séries temporelles avec une dépendance générale.

2 Analyse théorique de l'influence de γ sur ACI

L'analyse de distributions théoriques simples permet de construire une intuition sur le comportement d'ACI pour des données plus complexes. Nous faisons donc des hypothèses de modélisation supplémentaires sur les résidus, et nous ne considérons pas l'impact de l'ensemble de calibration : soit Q la fonction quantile des scores et supposons, pour tout $\hat{\alpha}$ et t, $\widehat{Q}_{1-\hat{\alpha}}(S_{\text{Calt}}) = Q(1-\hat{\alpha})$. Cela permet de se concentrer sur l'impact des mises à jour récursives dans (2) en décrivant leur comportement par la théorie des chaînes de Markov. Nous obtenons ainsi un résultat théorique dans le **cas échangeable** (Théorème 1) et une approximation numérique dans le **cas auto-régressif**.

Notons $L(\alpha_t) = 2Q(1-\alpha_t)$ la longueur de l'intervalle prédit par l'algorithme adaptatif en t, et $L_0 = 2Q(1-\alpha)$ celle de l'intervalle prédit lorsque $\gamma = 0$ (i.e., non-adaptatif).

Théorème 1 (cas échangeable) *Supposons que (i) $\alpha \in \mathbb{Q}$; (ii) les scores sont échangeables de fonction quantile Q ; (iii) la fonction quantile est parfaitement estimée à chaque instant ; (iv) la fonction quantile Q est bornée et $\mathcal{C}^4([0,1])$. Alors, pour tout $\gamma > 0$, $(\alpha_t)_{t>0}$ forme une chaîne de Markov, qui admet une distribution stationnaire π_γ, et*

$$\frac{1}{T}\sum_{t=1}^{T} L(\alpha_t) \underset{T\to+\infty}{\xrightarrow{p.s.}} \mathbb{E}_{\pi_\gamma}[L] \overset{not.}{=} \mathbb{E}_{\tilde{\alpha}\sim\pi_\gamma}[L(\tilde{\alpha})].$$

De plus, lorsque $\gamma \to 0$, $\mathbb{E}_{\pi_\gamma}[L] = L_0 + Q''(1-\alpha)\frac{\gamma}{2}\alpha(1-\alpha) + O(\gamma^{3/2})$.

Pour des distributions standard, $Q''(1-\alpha) > 0$. Alors, le Théorème 1 implique qu'ACI sur les scores échangeables *dégrade* l'*efficacité* linéairement avec γ par rapport à CP. Cela souligne que de tels algorithmes adaptatifs peuvent en fait nuire aux performances si les données n'ont pas de dépendance temporelle, et qu'un petit γ est préférable.

Dans le **cas de résidus auto-régressifs** de paramètre φ, nous prouvons la convergence de la longueur moyenne vers $\mathbb{E}_{\pi_{\gamma,\varphi}}[L]$, et nous estimons numériquement cette limite pour différents γ et φ. Cette analyse numérique met en évidence de nombreux éléments, dont le plus important est l'existence d'un $\gamma^* > 0$ minimisant la longueur, quand $\varphi > 0.6$.

Dans l'ensemble, ces résultats soulignent l'importance du choix de γ, car ne pas choisir γ^* peut conduire à des intervalles beaucoup plus grands. Cependant, le choix de γ dans des contextes pratiques plus complexes reste difficile : cela appelle des stratégies adaptatives.

3 AgACI

Nous proposons un nouvel algorithme, `AgACI`, basé sur l'agrégation d'experts en ligne (Cesa-Bianchi et Lugosi, 2006). L'utilisateur fournit une grille $\{\gamma_k, k \in [\![1, K]\!]\}$. K experts sont alors construits, l'expert k étant ACI avec γ_k. Deux agrégations indépendantes sont effectuées. L'une (resp. l'autre) pour la borne supérieure (inférieure) de l'intervalle utilisant uniquement les K bornes supérieures (inférieures) des experts ACI comme experts, avec pour fonction de perte la *pinball loss* de niveau $1-\alpha/2$ ($\alpha/2$).

4 Expériences synthétiques

Données et méthodologie. Nous générons des données selon le modèle suivant :

$$Y_t = 10 \sin(\pi X_{t,1} X_{t,2}) + 20 (X_{t,3} - 0.5)^2 + 10 X_{t,4} + 5 X_{t,5} + 0 X_{t,6} + \varepsilon_t, \tag{3}$$

où les X_t suivent une distribution multivariée uniforme sur $[0, 1]$, et $X_{t,6}$ représente une variable non-informative. Le bruit ε_t est généré selon un processus ARMA(1,1) de paramètres φ et θ, i.e. $\varepsilon_{t+1} = \varphi \varepsilon_t + \xi_{t+1} + \theta \xi_t$, avec ξ_t un bruit blanc. La dépendance temporelle est présente uniquement dans le bruit afin de contrôler sa force et son impact sur les performances des algorithmes de prédiction conformelle.

Pour évaluer l'impact de la structure temporelle, nous faisons varier φ et θ dans $\{0.1, 0.8, 0.9, 0.95, 0.99\}$. Pour chaque jeu de paramètres, nous générons $n = 500$ échantillons $(\varepsilon_t)_{t \in [\![1, T_0 + T_1]\!]}$ avec $T_0 = 200$ et $T_1 \in \{100, 200\}$ selon les expériences. Pour des raisons de simplicité, nous considérons $\varphi = \theta$.

Compte tenu de la structure non linéaire du processus de génération des données, nous utilisons une forêt aléatoire (RF) comme modèle prédictif. Nous comparons les méthodes suivantes : ACI avec 30 γ différents, AgACI, OSSCP et EnbPI.

Impact de la dépendance temporelle. Les résultats de ces expériences mettent notamment en évidence que :

- L'augmentation de la dépendance temporelle impacte la *validité*.
- **OSSCP.** Permet d'obtenir une couverture *valide* pour φ et θ inférieurs à 0.9, mais n'est pas robuste à la dépendance croissante.
- **EnbPI.** Sa *validité* dépend fortement de la distribution des données (elle est *valide* sur un bruit MA(1), pas sur un bruit AR(1) et ARMA(1,1)). Lorsque la méthode est *valide*, elle produit les intervalles les plus petits.
- **ACI.**
  - Atteint une couverture *valide* pour tous les paramètres de simulation avec un γ bien choisi, ou pour une dépendance telle que $\varphi < 0.95$.
  - Est robuste à l'augmentation de la force de la dépendance.
  - Plus la dépendance est forte, plus le choix de γ est critique.
- **AgACI.** Atteint une couverture *valide* pour tous les paramètres de simulation, avec une bonne *efficacité*.

5 Prévision des prix spot de l'électricité en France

Enfin, nous considérons un cas d'étude réel d'intérêt : la prévision des prix de l'électricité sur un marché court-terme français, les prix spot. L'utilisation croissante des énergies renouvelables intermittentes entraîne une dépendance et une volatilité accrues des marchés de l'énergie. Par conséquent, une prévision précise des prix de l'électricité est nécessaire

pour stabiliser la planification de la production d'énergie. Par ailleurs, le besoin en prévisions probabilistes a également augmenté, afin d'élaborer des stratégies fondées sur le risque.

Nous présentons les données en détails, puis nous comparons les méthodes précédentes. Cette application met en lumière qu'ACI et AgACI permettent d'améliorer l'*efficacité* par rapport aux méthodes de références de prévisions conformelles pour les séries temporelles. Ces dernières sont bien *valides* dans ce contexte. La Figure 2 représente des intervalles fournis par AgACI, afin d'illustrer la pertinence pratique d'une telle approche.

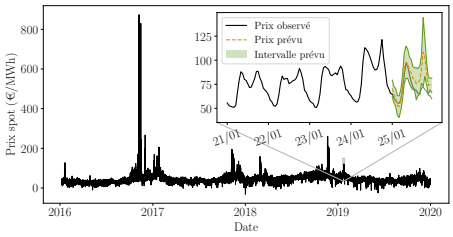


FIGURE 2 – Prix spot de l'électricité en France, 2016 à 2019. Prévision par AgACI pour le 25 janvier 2019.

Bibliographie

Cesa-Bianchi, N. et G. Lugosi (2006). *Prediction, learning, and games*. Cambridge University Press.

Chernozhukov, V. et al. (2018). Exact and Robust Conformal Inference Methods for Predictive Machine Learning with Dependent Data. *Conference On Learning Theory*. PMLR, p. 732-749.

Gibbs, I. et E. Candès (2021). Adaptive Conformal Inference Under Distribution Shift. *Advances in Neural Information Processing Systems*.

Kath, C. et F. Ziel (2021). Conformal prediction interval estimation and applications to day-ahead and intraday power markets. *International Journal of Forecasting* 37.2, p. 777-799.

Lei, J. et al. (2018). Distribution-Free Predictive Inference for Regression. *Journal of the American Statistical Association* 113.523, p. 1094-1111.

Papadopoulos, H. et al. (2002). Inductive Confidence Machines for Regression. *Machine Learning : ECML 2002*. Springer, p. 345-356.

Vovk, V., A. Gammerman et G. Shafer (2005). *Algorithmic Learning in a Random World*. Springer US.

Vovk, V., A. Gammerman et C. Saunders (1999). Machine-Learning Applications of Algorithmic Randomness. *Proceedings of the Sixteenth International Conference on Machine Learning*. Morgan Kaufmann Publishers Inc., p. 444-453.

Wisniewski, W. et al. (2020). Application of conformal prediction interval estimations to market makers' net positions. *Proceedings of the Ninth Symposium on Conformal and Probabilistic Prediction and Applications*. T. 128. PMLR, p. 285-301.

Xu, C. et Y. Xie (2021). Conformal prediction interval for dynamic time-series. *Proceedings of the 38th International Conference on Machine Learning*. T. 139. Proceedings of Machine Learning Research. PMLR, p. 11559-11569.

Model selection by penalization in mixture of experts models with a non-asymptotic approach

TrungTin Nguyen [1], Faicel Chamroukhi [2], Hien Duy Nguyen [3] & Florence Forbes [1]

[1] *Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France*
trung-tin.nguyen@inria.fr, florence.forbes@inria.fr
[2] *Normandie Univ, UNICAEN, CNRS, LMNO, 14000 Caen, France.*
faicel.chamroukhi@unicaen.fr
[3] *School of Mathematics and Physics, University of Queensland, St. Lucia, Australia.*
h.nguyen7@uq.edu.au

Résumé. Cette étude est consacrée au problème de la sélection de modèles parmi une collection de modèles de mélanges d'experts avec experts gaussiens et fonctions d'activations gaussiennes normalisées, caractérisés par le nombre de composantes du mélange et la complexité des experts moyens, dans un cadre d'estimation par maximum de vraisemblance pénalisée. En particulier, nous établissons des limites de risque non asymptotiques qui prennent la forme d'inégalités oracles faibles, sous une condition de limite inférieure pour la pénalité. Leur bon comportement empirique est ensuite démontré en simulation et sur des données réelles.

Mots-clés. Mélange d'experts, sélection de modèle, maximum de vraisemblance pénalisée.

Abstract. This study is devoted to the problem of model selection among a collection of Gaussian-gated localized mixtures of experts models characterized by the number of mixture components, and the complexity of Gaussian mean experts, in a penalized maximum likelihood estimation framework. In particular, we establish non-asymptotic risk bounds that take the form of weak oracle inequalities, provided that lower bounds of the penalties hold. Their good empirical behavior is then demonstrated on synthetic and real datasets.

Keywords. Mixture of experts, model selection, penalized maximum likelihood.

1 Introduction

Mixture of experts (MoE) models, originally introduced as neural network architectures in Jacobs et al. (1991), are flexible models that generalize the classical finite mixture and finite mixtures of regression models. The popularity of these conditional mixture density models arise largely due to their universal approximation properties, which have been studied in Nguyen et al. (2020b, 2021b), and which improve upon approximation

capabilities of unconditional finite mixture models, as studied in Nguyen et al. (2016), Ho et al. (2019), Nguyen et al. (2019, 2021a). Detailed reviews on the practical and theoretical aspects of MoE models can be found in Nguyen and Chamroukhi (2018), Nguyen (2021).

In this work, we study the class of MoE models with Gaussian experts and normalized Gaussian gating functions for clustering and regression, first introduced by Xu et al. (1995) and recently studied numerically in Chamroukhi et al. (2019). From hereon in, we refer to such models as *Gaussian-gated localized MoE* (GLoME) models. These models are useful to learn potentially nonlinear relationships between a multivariate output and a high-dimensional input issued from a heterogeneous population. This involves performing regression, clustering and model selection, simultaneously. While estimation can be performed using standard Expectation-Maximisation algorithms, it crucially depends and requires hyperparameter choices, including the number of mixture components (or clusters), and the degree of complexity of each Gaussian expert's mean function.

Traditional model selection criteria such as AIC, BIC, or ICL-BIC are based on asymptotic theory or Bayesian approaches. In contrast, the present contribution is to provide a finite-sample oracle inequality indicative of the quality of a data-driven selected GLoME model with respect to the true model. More specifically, we establish a non-asymptotic risk bound that takes the form of a weak oracle inequality, provided that a lower bound on the penalty holds true. Our non-asymptotic risk bound allows the number n of observations to be fixed while the dimensionality and cardinality of the models, characterized by the number of covariates and the dimension of the response, are allowed to grow with respect to n, and can be much larger than n. To the best of our knowledge, this is the most recent and advanced effort in literature to develop a finite-sample oracle inequality for the framework of MoE regression models.

From now on, we are interested in estimating the law of the random variable $\mathbf{Y}$, conditionally on $\mathbf{X}$, respectively of dimension L and D. Subsequently, $\left(\mathbf{X}_{[n]}, \mathbf{Y}_{[n]}\right) := (\mathbf{X}_i, \mathbf{Y}_i)_{i\in[n]}$, $[n] = \{1, \ldots, n\}, n \in \mathbb{N}^\star$, denotes a random sample, and $\mathbf{x}$ and $\mathbf{y}$ stands for the observed values of the random variables $\mathbf{X}$ and $\mathbf{Y}$, respectively. For a matrix $\mathbf{A}$, let $m(\mathbf{A})$ and $M(\mathbf{A})$ be, respectively, the modulus of the smallest and largest eigenvalues of $\mathbf{A}$.

2 Collection of GLoME models

We consider models with inverse conditional PDFs of the form (2.1). Such models have been considered in Xu et al. (1995), Deleforge et al. (2015) and are very useful in a high dimensional regression context, where typically $D \gg L$.

$$s_{\boldsymbol{\psi}_{K,d}}(\mathbf{x} \mid \mathbf{y}) = \sum_{k=1}^{K} g_k\left(\mathbf{y}; \boldsymbol{\omega}\right) \phi_D\left(\mathbf{x}; \boldsymbol{\upsilon}_{k,d}(\mathbf{y}), \boldsymbol{\Sigma}_k\right); \; g_k\left(\mathbf{y}; \boldsymbol{\omega}\right) = \frac{\pi_k \phi_L\left(\mathbf{y}; \mathbf{c}_k, \boldsymbol{\Gamma}_k\right)}{\sum_{j=1}^{K} \pi_j \phi_L\left(\mathbf{y}; \mathbf{c}_j, \boldsymbol{\Gamma}_j\right)}. \quad (2.1)$$

Here, $g_k(\cdot;\boldsymbol{\omega})$ and $\phi_D(\cdot;\boldsymbol{\upsilon}_{k,d}(\cdot),\boldsymbol{\Sigma}_k)$, $k \in [K]$, $K \in \mathbb{N}^\star$, $d \in \mathbb{N}^\star$, are called normalized Gaussian gating functions and Gaussian experts, respectively. Furthermore, we decompose the parameters of the model as follows: $\boldsymbol{\psi}_{K,d} = (\boldsymbol{\omega}, \boldsymbol{\upsilon}_d, \boldsymbol{\Sigma}) \in \boldsymbol{\Omega}_K \times \boldsymbol{\Upsilon}_{K,d} \times \mathbf{V}_K =: \boldsymbol{\Psi}_{K,d}$, $\boldsymbol{\omega} = (\boldsymbol{\pi}, \boldsymbol{c}, \boldsymbol{\Gamma}) \in (\boldsymbol{\Pi}_{K-1} \times \mathbf{C}_K \times \boldsymbol{V}'_K) =: \boldsymbol{\Omega}_K$, $\boldsymbol{\pi} = (\pi_k)_{k\in[K]}$, $\boldsymbol{c} = (\mathbf{c}_k)_{k\in[K]}$, $\boldsymbol{\Gamma} = (\boldsymbol{\Gamma}_k)_{k\in[K]}$, $\boldsymbol{\upsilon}_d = (\boldsymbol{\upsilon}_{k,d})_{k\in[K]} \in \boldsymbol{\Upsilon}_{K,d}$, and $\boldsymbol{\Sigma} = (\boldsymbol{\Sigma}_k)_{k\in[K]} \in \mathbf{V}_K$. Note that $\boldsymbol{\Pi}_{K-1} = \left\{(\pi_k)_{k\in[K]} \in (\mathbb{R}^+)^K, \sum_{k=1}^K \pi_k = 1\right\}$ is a $K-1$ dimensional probability simplex, $\mathbf{C}_K$ is a set of K-tuples of mean vectors of size $L \times 1$, $\boldsymbol{V}'_K$ is a set of K-tuples of elements in $\mathcal{S}_L^{++}$, where $\mathcal{S}_L^{++}$ denotes the collection of symmetric positive definite matrices on $\mathbb{R}^L$, $\boldsymbol{\Upsilon}_{K,d}$ is a set of K-tuples of mean functions from $\mathbb{R}^L$ to $\mathbb{R}^D$ depending on a degree d (*e.g.,* polynomial degree) and $\mathbf{V}_K$ is a set containing K-tuples from $\mathcal{S}_D^{++}$.

In order to establish our finite-sample oracle inequality, Theorem 3.1, we need to explicitly impose some classical boundedness conditions on the parameter space. Specifically, we assume that there exist deterministic positive constants $a_{\boldsymbol{\pi}}, A_{\boldsymbol{c}}, a_{\boldsymbol{\Gamma}}, A_{\boldsymbol{\Gamma}}$, and set

$$\widetilde{\boldsymbol{\Omega}}_K = \left\{\boldsymbol{\omega} \in \boldsymbol{\Omega}_K : \forall k \in [K], \|\mathbf{c}_k\|_\infty \leq A_{\boldsymbol{c}}, a_{\boldsymbol{\Gamma}} \leq m(\boldsymbol{\Gamma}_k) \leq M(\boldsymbol{\Gamma}_k) \leq A_{\boldsymbol{\Gamma}}, a_{\boldsymbol{\pi}} \leq \pi_k\right\}. \quad (2.2)$$

The set $\boldsymbol{\Upsilon}_{K,d}$ will be chosen as a tensor product of compact sets of moderate dimension (*e.g.,* a set of polynomials of degree smaller than d, whose coefficients are smaller in absolute values than $T_{\boldsymbol{\Upsilon}}$). In particular, we focus on the bounded $\mathcal{Y} = [0,1]^L$. In this case, $\varphi_{\boldsymbol{\Upsilon},i}$ can be chosen as monomials with maximum (non-negative) degree d: $\mathbf{y}^{\mathbf{r}} = \prod_{l=1}^L \mathbf{y}_l^{\mathbf{r}_l}$. Then, $\boldsymbol{\Upsilon}_{K,d} = \boldsymbol{\Upsilon}_{p,d}^K$, where

$$\boldsymbol{\Upsilon}_{p,d} = \left\{\mathbf{y} \mapsto \left(\sum_{|\mathbf{r}|=0}^{d} \boldsymbol{\alpha}_{\mathbf{r}}^{(j)} \mathbf{y}^{\mathbf{r}}\right)_{j\in[D]} =: (\boldsymbol{\upsilon}_{d,j}(\mathbf{y}))_{j\in[D]} : \|\boldsymbol{\alpha}\|_\infty \leq T_{\boldsymbol{\Upsilon}}\right\}. \quad (2.3)$$

For GLoME models, note that any covariance matrix $\boldsymbol{\Sigma}_k$ can be decomposed into the form $B_k \mathbf{P}_k \mathbf{A}_k \mathbf{P}_k^\top$ such that: $B_k = |\boldsymbol{\Sigma}_k|^{1/D}$ is a positive scalar corresponding to the volume, $\mathbf{P}_k$ is the matrix of eigenvectors of $\boldsymbol{\Sigma}_k$ and $\mathbf{A}_k$ the diagonal matrix of normalized eigenvalues of $\boldsymbol{\Sigma}_k$; $B_- \in \mathbb{R}^+, B_+ \in \mathbb{R}^+$, $\mathcal{A}(\lambda_-, \lambda_+)$ is a set of diagonal matrices $\mathbf{A}_k$, such that $|\mathbf{A}_k| = 1$ and $\forall i \in [D], \lambda_- \leq (\mathbf{A}_k)_{i,i} \leq \lambda_+$, where $\lambda_-, \lambda_+ \in \mathbb{R}$; and $SO(D)$ is the special orthogonal group of dimension D. In this way, we obtain the classical covariance matrix parameterization, described by Celeux and Govaert (1995) for Gaussian parsimonious clustering models, defined by

$$\mathbf{V}_K = \left\{\left(B_k \mathbf{P}_k \mathbf{A}_k \mathbf{P}_k^\top\right)_{k\in[K]} : B_- \leq B_k \leq B_+, \mathbf{P}_k \in SO(D), \mathbf{A}_k \in \mathcal{A}(\lambda_-, \lambda_+)\right\}. \quad (2.4)$$

For GLoME, we need to choose the degree of polynomials d and the number of components K among finite sets $\mathcal{D}_{\boldsymbol{\Upsilon}} = [d_{\max}]$ and $\mathcal{K} = [K_{\max}]$, respectively, where $d_{\max} \in \mathbb{N}^\star$ and $K_{\max} \in \mathbb{N}^\star$ may depend on the sample size n. We wish to estimate the unknown

true conditional density s_0 by conditional densities belonging to the following collection of models $(S_{\mathbf{m}})_{\mathbf{m}\in\mathcal{M}}$, $\mathcal{M} = \{(K,d) : K \in \mathcal{K}, d \in \mathcal{D}_{\boldsymbol{\Upsilon}}\}$,

$$S_{\mathbf{m}} = \left\{(\mathbf{x},\mathbf{y}) \mapsto s_{\boldsymbol{\psi}_{K,d}}(\mathbf{x} \mid \mathbf{y}) =: s_{\mathbf{m}}(\mathbf{x} \mid \mathbf{y}) : \boldsymbol{\psi}_{K,d} \in \widetilde{\boldsymbol{\Omega}}_K \times \boldsymbol{\Upsilon}_{K,d} \times \mathbf{V}_K\right\}, \tag{2.5}$$

where $\widetilde{\boldsymbol{\Omega}}_K$, $\boldsymbol{\Upsilon}_{K,d}$ and $\mathbf{V}_K$ are define previously in (2.2), (2.3) and (2.4), respectively.

3 Oracle inequality for collection of GLoME models

In the maximum likelihood approach, the Kullback–Leibler divergence is the most natural loss function. However, to take into account the structure of conditional densities and the random covariates $\left(\mathbf{Y}_{[n]}\right)$, we consider a *tensorized Kullback–Leibler divergence* $\mathrm{KL}^{\otimes \mathrm{n}}$, defined as:

$$\mathrm{KL}^{\otimes \mathrm{n}}(s,t) = \mathbb{E}_{\mathbf{Y}_{[n]}}\left[\frac{1}{n}\sum_{i=1}^{n}\mathrm{KL}\left(s\left(\cdot \mid \mathbf{Y}_i\right), t\left(\cdot \mid \mathbf{Y}_i\right)\right)\right], \tag{3.1}$$

if $s\ dy$ is absolutely continuous w.r.t. $t\ dy$, and $+\infty$ otherwise. We refer to our result as a *weak oracle inequality*, because its statement is based on a smaller divergence, when compared to $\mathrm{KL}^{\otimes \mathrm{n}}$, namely the *tensorized Jensen–Kullback–Leibler divergence* (Cohen and Le Pennec, 2011): given $\rho \in (0,1)$,

$$\mathrm{JKL}_{\rho}^{\otimes \mathrm{n}}(s,t) = \mathbb{E}_{\mathbf{Y}_{[n]}}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\rho}\mathrm{KL}\left(s\left(\cdot \mid \mathbf{Y}_i\right), (1-\rho)\, s\left(\cdot \mid \mathbf{Y}_i\right) + \rho t\left(\cdot \mid \mathbf{Y}_i\right)\right)\right].$$

In a penalized maximum likelihood estimation context (PMLE), by adding a suitable penalty $\mathrm{pen}(\mathbf{m})$, one hopes to create a trade-off between a good data fit and model complexity. For a given choice of $\mathrm{pen}(\mathbf{m})$, the *selected model* $S_{\widehat{\mathbf{m}}}$ is chosen as the one whose index is an η'-almost minimizer of the sum of the negative log-likelihood and this penalty. That is $S_{\widehat{\mathbf{m}}} = \widehat{s}_{\widehat{\mathbf{m}}}$, satisfying

$$\sum_{i=1}^{n} -\ln\left[\widehat{s}_{\widehat{\mathbf{m}}}\left(\mathbf{x}_i \mid \mathbf{y}_i\right)\right] + \mathrm{pen}\left(\widehat{\mathbf{m}}\right) \leq \inf_{\mathbf{m}\in\mathcal{M}}\left\{\sum_{i=1}^{n} -\ln\left[\widehat{s}_{\mathbf{m}}\left(\mathbf{x}_i \mid \mathbf{y}_i\right)\right] + \mathrm{pen}(\mathbf{m})\right\} + \eta', \tag{3.2}$$

$$\text{where } \sum_{i=1}^{n} -\ln\left[\widehat{s}_{\mathbf{m}}\left(\mathbf{x}_i \mid \mathbf{y}_i\right)\right] \leq \inf_{s_{\mathbf{m}}\in S_{\mathbf{m}}}\sum_{i=1}^{n} -\ln\left[s_{\mathbf{m}}\left(\mathbf{x}_i \mid \mathbf{y}_i\right)\right] + \eta. \tag{3.3}$$

Note that $\widehat{s}_{\widehat{\mathbf{m}}}$ is then called the η'-penalized likelihood estimate and depends on both the error terms η and η'. From hereon in, the term *selected model (estimate) or best data-driven model (estimate)* is used to indicate that it satisfies (3.2).

Theorem 3.1, proved in Nguyen et al. (2021d), provides a lower bound on the penalty function, $\mathrm{pen}(\mathbf{m})$, which guarantees that the PMLE selects a model in the collection that performs almost as well as the best model.

Theorem 3.1 (Finite-sample weak oracle inequality for GLoME models)**.** *Assume that we observe* $\left(\mathbf{x}_{[n]}, \mathbf{y}_{[n]}\right)$*, arising from an unknown conditional density* s_0*. Given a collection of GLoME models,* $(S_{\mathbf{m}})_{\mathbf{m}\in\mathcal{M}}$*, there is a constant* C *such that for any* $\rho \in (0,1)$*, for any* $\mathbf{m} \in \mathcal{M}$*,* $z_{\mathbf{m}} \in \mathbb{R}^+$*,* $\Xi = \sum_{\mathbf{m}\in\mathcal{M}} e^{-z_{\mathbf{m}}} < \infty$ *and any* $C_1 > 1$*, there is a constant* κ *depending only on* ρ *and* C_1*, such that if for every index* $\mathbf{m} \in \mathcal{M}$*,*

$$pen(\mathbf{m}) > \kappa\left[(C + \ln n)\dim\left(S_{\mathbf{m}}\right) + z_{\mathbf{m}}\right],$$

then the η'*-penalized likelihood estimate* $\widehat{s}_{\widehat{\mathbf{m}}}$*, defined in* (3.3) *and* (3.2)*, satisfies*

$$\mathbb{E}_{\mathbf{X}_{[n]},\mathbf{Y}_{[n]}}\left[\mathrm{JKL}_{\rho}^{\otimes \mathrm{n}}\left(s_0, \widehat{s}_{\widehat{\mathbf{m}}}\right)\right] \leq C_1 \inf_{\mathbf{m}\in\mathcal{M}}\left(\inf_{s_{\mathbf{m}}\in S_{\mathbf{m}}} \mathrm{KL}^{\otimes \mathrm{n}}\left(s_0, s_{\mathbf{m}}\right) + \frac{pen(\mathbf{m})}{n}\right) + \frac{\kappa C_1 \Xi + \eta + \eta'}{n}. \tag{3.4}$$

It is worth noting that Theorem 3.1 extends a corollary of (Montuelle et al., 2014, Theorem 1), which can be verified via Lemma 1 from Nguyen et al. (2021a), which makes explicit the relationship between softmax and Gaussian gating classes. In particular, for modeling a sample of high-dimensional regression data issued from a heterogeneous population with hidden graph-structured interaction between covariates, we refer readers to the works of Nguyen et al. (2021c) while Nguyen et al. (2020a) provides a stronger oracle inequality but slower convergence rate of the error upper bound. More precisely, in (3.4), we obtain a weak oracle inequality due to the different divergences on the left, $\mathrm{JKL}_{\rho}^{\otimes \mathrm{n}}$, and on the right $\mathrm{KL}^{\otimes \mathrm{n}}$, but with a faster convergence rate $\mathcal{O}(n^{-1})$ of the error upper bound compared to $\mathcal{O}(n^{-1/2})$ in Nguyen et al. (2020a, Theorem 3.2). A more detailed comparison can be found in Nguyen (2021, Section 1.2.12.3).

Numerical experiments are available in Nguyen et al. (2021d) and https://github.com/Trung-TinNGUYEN/NamsGLoME-Simulation, will be presented in the communication in order to investigate how well the empirical tensorized Kullback–Leibler divergence between the true model and the best data-driven model follows the finite-sample oracle inequality of Theorem 3.1, as well as the convergence rate of the error upper bound $\left(\kappa C_1 \Xi + \eta + \eta'\right)/n$.

References

Celeux, G. and Govaert, G. (1995). Gaussian parsimonious clustering models. *Pattern recognition*, 28(5):781–793. 3

Chamroukhi, F., Lecocq, F., and Nguyen, H. D. (2019). Regularized Estimation and Feature Selection in Mixtures of Gaussian-Gated Experts Models. In Nguyen, H., editor, *Statistics and Data Science*, pages 42–56, Singapore. Springer Singapore. 2

Cohen, S. and Le Pennec, E. (2011). Conditional density estimation by penalized likelihood model selection and applications. *Technical report, INRIA*. 4

Deleforge, A., Forbes, F., and Horaud, R. (2015). High-dimensional regression with gaussian mixtures and partially-latent response variables. *Statistics and Computing*, 25(5):893–911. 2

Ho, N., Yang, C.-Y., and Jordan, M. I. (2019). Convergence rates for gaussian mixtures of experts. *arXiv preprint arXiv:1907.04377*. 2

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87. 1

Montuelle, L., Le Pennec, E., et al. (2014). Mixture of gaussian regressions model with logistic weights, a penalized maximum likelihood approach. *Electronic Journal of Statistics*, 8(1):1661–1695. 5

Nguyen, H. D. and Chamroukhi, F. (2018). Practical and theoretical aspects of mixture-of-experts modeling: An overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1246. 2

Nguyen, H. D., Chamroukhi, F., and Forbes, F. (2019). Approximation results regarding the multiple-output Gaussian gated mixture of linear experts model. *Neurocomputing*, 366:208–214. 2

Nguyen, H. D., Lloyd-Jones, L. R., and McLachlan, G. J. (2016). A universal approximation theorem for mixture-of-experts models. *Neural computation*, 28(12):2585–2593. 2

Nguyen, H. D., Nguyen, T., Chamroukhi, F., and McLachlan, G. J. (2021a). Approximations of conditional probability density functions in Lebesgue spaces via mixture of experts models. *Journal of Statistical Distributions and Applications*, 8(1):13. 2, 5

Nguyen, T. (2021). *Model Selection and Approximation in High-dimensional Mixtures of Experts Models: from Theory to Practice*. PhD Thesis, Normandie Université. 2, 5

Nguyen, T., Chamroukhi, F., Nguyen, H., and McLachlan, G. (2021b). Approximation of probability density functions via location-scale finite mixtures in Lebesgue spaces. *Communications in Statistics - Theory and Methods*. 1

Nguyen, T., Chamroukhi, F., Nguyen, H. D., and Forbes, F. (2021c). Non-asymptotic model selection in block-diagonal mixture of polynomial experts models. *arXiv preprint arXiv:2104.08959*. 5

Nguyen, T., Nguyen, H. D., Chamroukhi, F., and Forbes, F. (2021d). A non-asymptotic penalization criterion for model selection in mixture of experts models. *arXiv preprint arXiv:2104.02640*. 4, 5

Nguyen, T., Nguyen, H. D., Chamroukhi, F., and McLachlan, G. J. (2020a). An l_1-oracle inequality for the Lasso in mixture-of-experts regression models. *arXiv preprint arXiv:2009.10622*. 5

Nguyen, T., Nguyen, H. D., Chamroukhi, F., and McLachlan, G. J. (2020b). Approximation by finite mixtures of continuous density functions that vanish at infinity. *Cogent Mathematics & Statistics*, 7(1):1750861. 1

Xu, L., Jordan, M., and Hinton, G. E. (1995). An Alternative Model for Mixtures of Experts. In Tesauro, G., Touretzky, D., and Leen, T., editors, *Advances in Neural Information Processing Systems*, volume 7. MIT Press. 2

Apprentissage non supervisé

Clustering Longitudinal Ordinal Data via Finite Mixture of Matrix-Variate Latent Gaussians

Francesco Amato [1] & Julien Jacques [1]

[1] *Univ Lyon, Univ Lyon 2, ERIC, Lyon.*
{*francesco.amato, julien.jacques*}*@univ-lyon2.fr*

Résumé. Dans les sciences sociales ou la médecine, les études sont souvent basées sur des questionnaires demandant aux participants d'exprimer des réponses ordonnées à plusieurs reprises au cours d'une période d'étude. Nous présentons un modèle pour effectuer un clustering temporel sur ces données. Le modèle repose sur un mélange de distributions normales à variation matricielle, en tenant compte simultanément des structures de dépendance temporelle internes et intermédiaires. Un algorithme MC-EM pour l'estimation du modèle est utilisé. Des applications sur les données synthétiques et réelles sont présentées.

Mots-clés. clustering. Données ordinal longitudinales. Tenseurs. Modèle de mélange. Loi gaussienne matricielle.

Abstract. In social sciences or medicine, studies are often based on questionnaires asking participants to express ordered responses several times over a study period. We present a model to perform temporal clustering on such data. The model relies on mixture of matrix-variate normal distributions, accounting for the within and between time-dependence structures simultaneously. A MC-EM algorithm for the model estimation is used. Applications on synthetic and real data are presented.

Keywords. Clustering. Ordinal longitudinal data. Three-way data. Mixture models. Matrix-variate Gaussians.

1 Context

In many areas of humanities and social sciences, the studies are based on questionnaires completed by participants several times over the study period. The researchers then analyse these questionnaires to determine typical behaviours within the studied population, being especially interested in their time evolution. Nonetheless, modelling temporal evolution is far from trivial. The most basic approach consists in performing analyses independently at each temporal phase, and then trying *a posteriori* to find links between these different analyses, by seeking from one phase to the other to find similar or different typical behaviours. An example is Selosse et al., 2019, clustering of ordinal data for an application in psychology. The ideal way to cluster these data would be to account for the

temporal evolution, modelling all the responses to the questionnaires at the same time. We propose a model-based clustering technique aiming at facilitate such temporal analysis, by grouping together the units behaving similarly in time. Over the decades, research has produced a vast number of different approaches to clustering. From our prospective, probabilistic (or model-based) clustering offers the advantage of clearly stating the assumptions behind the clustering algorithm, and allows cluster analysis to benefit from the inferential framework of statistics to address some of the practical questions arising when performing clustering: determine the number of clusters, detecting and treating outliers, assessing uncertainty in the clustering (Bouveyron et al., 2019).

1.1 Related works

An approach to clustering longitudinal data consists in arranging the data in a three-way format and modelling them through a matrix-variate mixture model. This approach offers the advantage of accounting for the overall time-behavior, grouping together the units that have a similar pattern across and within time. While not being new (Basford et al., 1985), matrix-variate distributions have recently gained attention, and mixtures of matrix-normals (MMN) have been developed and applied both in a frequentist framework in Viroli, 2011a and within a Bayesian one by Viroli, 2011b, where it was used to cluster Italian provinces based on a longitudinal crime-related score. From a frequentist point of view, these models represent a natural extension of the multivariate normal mixtures to account for temporal (or even spatial) dependencies, and have the advantage of being also relatively easy to estimate by means of EM algorithm (a nice short description of the EM application to MNN is provided in §2.1 of Wang et al., 2020). More recently, in Gallaugher et al., 2018 and Melnykov et al., 2018, 2019 extensions for non-normal skewed cases have been proposed and applied. However, matrix-variate models suffer from over-parametrization that leads to estimation issues. To overcome this issue a more parsimonious model (Sarkar et al., 2020) and a new R package (Zhu et al., 2021) has been proposed. Despite their efficacy, up to now these methods have only been applied to continous data. Our model expands the use to matrix-variate mixtures to ordinal data.

2 Model

Let denote by $y_{i,j,t}, i = 1, ..., N; j = 1, .., J$ and $t = t, ..., T$ the observation of the j-th variable for the i-th unit at time t, that is: imagine to observe N units and measuring J different ordinal variables T times throughout the course of the study. Let us reorganize our data in a random-matrix form such that $\mathbf{Y} = \{Y_i\}_{i=1}^{N}$ is a sample of $J \times T$-variate matrix observations (i.e. $Y_i = (y_{i,j,t}) \in \mathbb{R}^{J \times T}$). Then, we can assume that each variable $y_{i,j,t}$ is the manifestation of an underlying latent variable $z_{i,j,t}$ which follows a Gaussian distribution, as done in the clustMD model (McParland et al., 2016). At this point, we

can assume that each observed matrix Y_i is indeed the manifestation of a latent random matrix Z_i, which follows a matrix-normal distribution.

$$Y_i = \begin{pmatrix} y_{i,1,1} & \cdots & y_{i,1,t} & \cdots & y_{i,1,T} \\ \vdots & \ddots & \vdots & \cdots & \vdots \\ y_{i,j,1} & \cdots & y_{i,j,t} & \cdots & y_{i,j,T} \\ \vdots & \cdots & \vdots & \ddots & \vdots \\ y_{i,J,1} & \cdots & y_{i,J,t} & \cdots & y_{i,J,T} \end{pmatrix} \longleftarrow Z_i = \begin{pmatrix} z_{i,1,1} & \cdots & z_{i,1,t} & \cdots & z_{i,1,T} \\ \vdots & \ddots & \vdots & \cdots & \vdots \\ z_{i,j,1} & \cdots & z_{i,j,t} & \cdots & z_{i,j,T} \\ \vdots & \cdots & \vdots & \ddots & \vdots \\ z_{i,J,1} & \cdots & z_{i,J,t} & \cdots & z_{i,J,T} \end{pmatrix}$$

To map from Y_i to Z_i, let γ_j denote a C_{j+1}-dimensional vector of thresholds that partition the real line for the j-th ordinal variable that has C_j levels and let the threshold parameters be constrained such that $-\infty = \gamma_{j,0} \leq \gamma_{j,1} \leq ... \leq \gamma_{j,C_j} = \infty$. If the latent $z_{i,j,t}$ is such that $\gamma_{j,c-1} < z_{i,j,t} < \gamma_{j,c}$ then the observed ordinal response, $y_{i,j,t} = c$.

So, by assuming that each Z_i follows a matrix-normal distribution, we can then cluster our data by means of finite Mixture of Matrix-Normals (MNN) (Viroli, 2011a). Indeed, let $Z_i \sim \mathcal{MN}_{(J\times T)}(M, \Phi, \Sigma)$, where $M \in \mathbb{R}^{J\times T}$ is the matrix of means, $\Phi \in \mathbb{R}^{T\times T}$ is a covariance matrix containing the variances and covariances between the T occasions or times and $\Sigma \in \mathbb{R}^{J\times J}$ is the covariance matrix containing the variance and covariances of the J variables. The matrix-normal probability density function (pdf) is given by

$$f(Z|M, \Phi, \Sigma) = (2\pi)^{-\frac{TJ}{2}} |\Phi|^{-\frac{J}{2}} |\Sigma|^{-\frac{T}{2}} \exp\left\{-\frac{1}{2}\text{tr}[\Sigma^{-1}(Z - M)\Phi^{-1}(Z - M)^\intercal]\right\}. \quad (1)$$

The matrix-normal distribution represents a natural extension of the multivariate normal distribution, since if $Z \sim \mathcal{MN}_{(J\times T)}(M, \Phi, \Sigma)$, then $\text{vec}(Z) \sim \mathcal{MVN}_{JT}(\text{vec}(M), \Phi \otimes \Sigma)$, where vec(.) is the vectorization operator and $\otimes$ denotes the Kronecker product. Then, the mean and the variance of the matrix-normal distribution are:

$$\mathbb{E}(\text{vec}(Z)|M, \Phi, \Sigma) = \text{vec}(M) \quad \text{and} \quad \mathbb{V}(\text{vec}(Z)|M, \Phi, \Sigma) = \Phi \otimes \Sigma. \quad (2)$$

Being a special case of the multivariate normal distribution, the matrix-normal distribution shares the same various properties, like, for instance, closure under marginalization, conditioning and linear transformations (Gupta et al., 2000). The separability condition of the covariance matrix has the twofold advantage of allowing the modeling of the temporal pattern of interest directly on the covariance matrix Φ and of representing a more parsimonious solution than that of the unrestricted $\Phi \otimes \Sigma$.
The pdf of the MMN model is given by

$$f(Z|\boldsymbol{\pi}, \boldsymbol{\Theta}) = \sum_{k=1}^{K} \pi_k \phi^{(J\times T)}(Z|M_k, \Phi_k, \Sigma_k),$$

where $\phi^{(J\times T)}$ represents the density function of a $J\times T$-dimensional matrix-variate normal, K is the number of mixture components, $\boldsymbol{\pi} = \{\pi_k\}_{k=1}^K$ is the vector of mixing proportions, subject to constraint $\sum_{k=1}^K \pi_k = 1$ and $\boldsymbol{\Theta} = \{\Theta_k\}_{k=1}^K$ is the set of component-specific parameters with $\Theta_k = \{M_k, \Phi_k, \Sigma_k\}$.
In addition to Z, we introduce a latent binary variable that indicate whether the unit i belongs to the k-th cluster, $\boldsymbol{\ell}_i = (\ell_{i1}, ..., \ell_{iK})$, such that $\ell_{ik} = 1$ if the i-th unit belongs to the k-th cluster. Let the binary vector $\mathbf{Y}_i^R = (Y_{i1}^R, ..., Y_{iR}^R)$ of length R indicate the observed response patter for the i-th unit, such that if the r-th pattern is observed then $Y_{ir}^R = 1$ and any other entry in the vector equals zero. We can derive the joint density of $\mathbf{Z}_i, \mathbf{Y}_i^R, \boldsymbol{\ell}_i$ as:

$$f(\mathbf{Y}_i^R, \mathbf{Z}_i, \boldsymbol{\ell}_i) = f(\mathbf{Y}_i^R|\mathbf{Z}_i, \boldsymbol{\ell}_i) f(\mathbf{Z}_i|\boldsymbol{\ell}_i) f(\boldsymbol{\ell}_i).$$

Assuming that:

$$\begin{aligned}
\boldsymbol{\ell}_i &\sim \mathcal{M}(1, \boldsymbol{\pi}),\ \boldsymbol{\pi} = (\pi_1, ..., \pi_K) \\
\mathbf{Z}_i|\boldsymbol{\ell}_{ik} = 1 &\sim \mathcal{MN}_{(J\times T)}(\mathbf{Z}_i|\Theta_k),\ \Theta_k = \{M_k, \Phi_k, \Sigma_k\}, \\
\mathbf{Y}_i^R|\mathbf{Z}_i, \boldsymbol{\ell}_{ik} = 1 &\sim \mathcal{M}(1, \boldsymbol{\xi}_i^R),\ \boldsymbol{\xi}_i^R = (\mathbf{1}_{\Omega_1}(\mathbf{Z}_i), \ldots, \mathbf{1}_{\Omega_R}(\mathbf{Z}_i))
\end{aligned}$$

we get:

$$f(\boldsymbol{\ell}_i) = \prod_{k=1}^K \pi_k^{\ell_{ik}},\ f(\mathbf{Z}_i|\boldsymbol{\ell}_i) = \prod_{k=1}^K \left[\phi^{(J\times T)}(Z_i|\Theta_k)\right]^{\ell_{ik}},\ f(\mathbf{Y}_i^R|\mathbf{Z}_i, \boldsymbol{\ell}_i) = \prod_{r=1}^R \mathbf{1}_{\Omega_r}(Z_i)^{Y_{ir}^R},$$

where $\mathcal{M}$ indicate the multinomial distribution and $\mathbf{1}_{\Omega_r}(\mathbf{Z}_i)$ is the indicator function that equals 1 when the elements in $\mathbf{Z}_i$ have values that determine the r-th pattern: one can imagine Ω_r as a $J\times T$ matrix whose elements are the indicator functions of the thresholds linked to the r-th pattern. Of course, when Z_i is given, the the value of Y_i is no more random.

3 Estimation

A key point is of course the choice of the thresholds $\boldsymbol{\gamma} = \{\gamma_j\}_{j=1}^J$. In Corneli et al., 2020, thresholds are fixed in advance to avoid identifiability and computational complexity issues. While this can be a starting point in the estimation process, our model aims at treating them as parameters and estimating them, as in McParland et al., 2013, 2016.

If the thresholds are assumed to be fixed in advance, the estimation process clearly simplifies. To estimate the model, since we do not observe neither Z nor ℓ, we resort to the EM algorithm (Dempster et al., 1977). The complete log-likelihood can be then written

as

$$\log \mathcal{L}(\boldsymbol{\pi}, \boldsymbol{\Theta} | \mathbf{Y}^R, \mathbf{Z}, \boldsymbol{\ell}) \propto \sum_{i=1}^{N} \left\{ \sum_{r=1}^{R} Y_{ir} \mathbf{1}_{\Omega_r}(Z_i) + \sum_{k=1}^{K} \ell_{ik} \left[\log(\pi_k) - \frac{TJ}{2} \log(2\pi) - \frac{J}{2} \log(|\Phi_k|) - \frac{T}{2} \log(|\Sigma_k|) - \frac{1}{2} tr[\Sigma_k^{-1}(Z_i - M_k)\Phi_k^{-1}(Z_i - M_k)^\intercal] \right] \right\}$$

The expectation step (E-step) of the EM algorithm consists of computing the expectation of the complete log-likelihood with respect to the latent data Z and the cluster labels ℓ. In the first step we attempt to estimate the latent variables. For each response pattern r, we can approximate the value of $Z_i|\ell_i$ as the expected value of the truncated multivariate normal (using (2)), given the parameters Θ_k of the assigned cluster. The latent variable ℓ can be computed by means of Bayes'theorem as:

$$\mathbb{E}(\ell_{ik} | Y_i^R = r, \boldsymbol{\Theta}, \boldsymbol{\pi}) = \frac{\pi_k \int_{\Omega_r} f(Z|\Theta_k) dZ}{\sum_{k=1}^{K} \pi_k \int_{\Omega_r} f(Z|\Theta_k) dZ},$$

which requires Monte-Carlo approximation on the multivariate reparametrization. By taking the first derivatives of the log-likelihood, the M-step goes by:

$$\hat{\pi}_k = \frac{\sum_{i=1}^{N} \ell_{ik}}{N} \quad , \quad \hat{M}_k = \frac{\sum_{i=1}^{N} \ell_{ik} Z_i}{\sum_{i=1}^{N} \ell_{ik}}$$

$$\hat{\Phi}_k = \frac{\sum_{i=1}^{N} \ell_{ik} (Z_i - \hat{M}_k) \hat{\Sigma}_k^{-1} (Z_i - \hat{M}_k)^\intercal}{J \sum_{i=1}^{N} \ell_{ik}} \quad , \quad \hat{\Sigma}_k = \frac{(Z_i - \hat{M}_k)^\intercal \hat{\Phi}_k^{-1} (Z_i - \hat{M}_k)}{T \sum_{i=1}^{N} \ell_{ik}}$$

4 Conclusions

Mixture of matrix-variate normal distributions can be an efficient way to cluster longitudinal continuous data. Assuming that ordinal variables is a discretization of a latent continuous variable allows us to extend the use of these MMN to ordinal variables. Numerical study on synthetic data sets as well as real data application concerning diet choice during the pandemic (François-Lecompte et al., 2020) will be presented.

References

[1] Arthur P Dempster, Nan M Laird, and Donald B Rubin. "Maximum likelihood from incomplete data via the EM algorithm". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 39.1 (1977), pp. 1–22.

[2] Kaye E Basford and Geoffrey J McLachlan. "The mixture method of clustering applied to three-way data". In: *Journal of Classification* 2.1 (1985), pp. 109–125.

[3] A.K. Gupta and D.K. Nagar. *Matrix Variate Distributions*. Chapman and Hall/CRC, 2000.

[4] Cinzia Viroli. "Finite mixtures of matrix normal distributions for classifying three-way data". In: *Statistics and Computing* 21.4 (2011), pp. 511–522.

[5] Cinzia Viroli. "Model based clustering for three-way data structures". In: *Bayesian Analysis* 6.4 (2011), pp. 573–602.

[6] Damien McParland and Isobel Claire Gormley. "Clustering ordinal data via latent variable models". In: *Algorithms from and for Nature and Life*. Springer, 2013, pp. 127–135.

[7] Damien McParland and Isobel Claire Gormley. "Model based clustering for mixed data: clustMD". In: *Advances in Data Analysis and Classification* 10.2 (2016), pp. 155–169.

[8] Michael PB Gallaugher and Paul D McNicholas. "Finite mixtures of skewed matrix variate distributions". In: *Pattern Recognition* 80 (2018), pp. 83–93.

[9] Volodymyr Melnykov and Xuwen Zhu. "On model-based clustering of skewed matrix data". In: *Journal of Multivariate Analysis* 167 (2018), pp. 181–194.

[10] Charles Bouveyron et al. *Model-based Clustering and Classification for Data Science: with applications in R*. Cambridge University Press, 2019.

[11] Volodymyr Melnykov and Xuwen Zhu. "Studying crime trends in the USA over the years 2000–2012". In: *Advances in Data Analysis and Classification* 13.1 (2019), pp. 325–341.

[12] Margot Selosse et al. "Analysing a quality of life survey using a co-clustering model for ordinal data and some dynamic implications". In: *Journal of the Royal Statistical Society: Series C Applied Statistics* 68.5 (2019), pp. 1327–1349.

[13] Marco Corneli, Charles Bouveyron, and Pierre Latouche. "Co-clustering of ordinal data via latent continuous random variables and not missing at random entries". In: *Journal of Computational and Graphical Statistics* (2020).

[14] Agnès François-Lecompte et al. "Confinement et comportements alimentaires : Quelles évolutions en matière d'alimentation durable ?" In: *Revue Française de Gestion* 293 (2020).

[15] Shuchismita Sarkar et al. "On parsimonious models for modeling matrix data". In: *Computational Statistics & Data Analysis* 142 (2020), p. 106822.

[16] Yang Wang and Volodymyr Melnykov. "On variable selection in matrix mixture modelling". In: *Stat* 9.1 (2020), e278.

[17] Xuwen Zhu, Shuchismita Sarkar, and Volodymyr Melnykov. "MatTransMix: an R Package for Matrix Model-Based Clustering and Parsimonious Mixture Modeling". In: *Journal of Classification* (2021), pp. 1–24.

Process Mining : une nouvelle approche de l'évaluation des clusters à l'aide d'un Quantitative Model Checking[1]

Paolo Ballarini [1] & Benoît Barbot [2] & Emmanuelle Claeys [3] & Pierre Cry [2 3]

[1] *CentraleSupélec, Université Paris Saclay, France*
[2] *Université Paris Est Créteil, LACL, F-94010 Créteil, France*
[3] *Unviversité Paul Sabatier, Irit Toulouse, France*

Résumé. Le *Process Mining* regroupe différentes méthodes pour construire un modèle formel à partir d'un ensemble de traces émises par un système réel. Dans ce cadre, les traces sont données sous la forme d'un *event log*, c'est-à-dire un jeu de données enregistrant un ensemble d'événements temporels. L'un des objectifs finaux du *Process mining* est de produire une représentation (sous forme de réseaux de Petri) et d'extrapoler des temps de parcours dans différentes situations. Malheureusement, la qualité des techniques classiques d'extraction de réseaux de Petri dépend fortement de l'ensemble des données, pour lequel des parcours possibles trop grands ou trop complexes rendent les réseaux de Petri produits inutilisables pour faire des prédictions. Par conséquent, il est parfois nécessaire de discriminer ou de regrouper cet *event log* afin de générer des processus plus lisibles et exploitables pour l'utilisateur. Ces groupes de parcours similaires sont appelés *clusters*. De nombreux algorithmes permettant de *clusteriser* les parcours sont apparus ces dernières années, cependant, il est difficile d'évaluer la qualité des clusters obtenus, le choix des métriques existantes étant arbitraire. Nous proposons dans cet article une nouvelle façon d'évaluer la qualité du clustering en convertissant ces clusters en réseaux de Petri et en utilisant un outil de model checking statistique basé sur une logique temporelle HASL. Nous expérimentons notre technique dans un benchmark classique de process mining avec des résultats concluants.

Mots-clés. Process Mining , Statistical Model Checking, Réseaux de Pétri.

Abstract. *Process Mining* brings together different methods for building a formal model from a set of traces provided by a real system. In this paper, the traces are given in the form of an event log, i.e. a large data set recording a set of temporal events. One of the final objectives of process mining is to produce a representation (in the form of Petri nets) and extrapolate path times in different situations. Unfortunately, the quality of classical Petri net mining techniques is highly dependent on the dataset, for which too large or too complex possible paths make the produced Petri nets unusable for prediction. Therefore, it is sometimes necessary to discriminate or group this *event log* in order to generate more readable and exploitable processes for the user. These groups of similar traces are called clusters. Many algorithms for clustering traces have appeared in recent years, but it is difficult to assess the quality of the clusters obtained, as the choice of existing metrics is arbitrary. In this paper, we propose a new way to evaluate the quality

of clustering by converting these clusters into Petri nets and using a statistical model checking tool based on a temporal HASL logic. We experiment our technique in a classical process mining benchmark with conclusive results.

Keywords. Process Mining, Statistical Model Checking, Petri Nets.

1 Introduction

L'exploration de processus implique des méthodes d'analyse de *process* représentés par des modèles à partir de *event logs* (c'est-à-dire les données réelles émises, supposées suivre le processus observé). Son objectif principal est la construction d'un modèle à partir de données (journal d'événements), où le modèle extrait doit être capable de reproduire, aussi fidèlement que possible, les comportements observés dans les données. En effet, un autre aspect pertinent de l'exploration de processus est la vérification de la conformité, dont l'objectif est d'évaluer dans quelle mesure le modèle obtenu est conforme à l'*event logs*.

Chaque événement de l'*event logs* est composé de trois informations : un identifiant de cas, un horodatage et une activité. Grâce à ces informations, il est possible de savoir quel individu est concerné par l'événement (cas), quand l'événement se produit (horodatage) et l'action réalisée (activité). Nous distinguons ici la notion de *parcours* et de *trace*. Un parcours est la séquence d'événements avec un identifiant de cas commun, chaque parcours résume essentiellement les activités d'un individu qui a pris part au *process*. Une trace, au contraire, est une séquence d'événements qui correspond à un parcours unique et qui est associée à une valeur de fréquence (exprimant le nombre d'individus qui ont présenté cette trace). Dans un *event log*, plusieurs individus ont un même parcours associé à une unique trace. À partir de cet *event log*, l'utilisateur peut extraire un modèle sous forme de *workflow* à l'aide de différentes méthodes comme Alpha miner [1] ou encore Heuristic miner [7].

L'évaluation d'un modèle se fait généralement au moyen de quatre métriques quantitatives : la *fitness*, la *précision*, la *généralisation* et la *simplicité*[5, 2]. Par conséquent, l'exploration de processus est essentiellement un problème d'optimisation multi-objectifs, et un problème difficile étant donné la nature conflictuelle de ces métriques. Une autre métrique courante est le score F1 (rapport entre l'adéquation et la précision) qui peut être calculé de manière simple en comptant le nombre de traces dans l'*event log* qui peuvent être complètement reproduites par le modèle, sur le nombre total de traces présentes dans l'*event log* [2].

Dans la littérature, les modèles extraits (*workflow*) sont souvent transformés en réseaux de Petri pour faire des analyses statistiques sur le modèle. Bien qu'important pour évaluer la conformité d'un modèle, les métriques décrites ci-dessus ne tiennent pas compte des aspects du processus axé sur la performance. Par exemple, si l'utilisateur souhaite répondre

à des questions liées à des propriétés statistiques telles que "Quel est le temps moyen de traitement d'un client ?", non seulement les mesures standards mais aussi le *type qualitatif* de modèle classiquement obtenu à partir de l'*event logs* ne sont pas suffisamment expressifs. En effet, une analyse statistique brute des traces dans l'*event log* inclut des chemins qui ne sont pas pertinents (qui peuvent être considérés ici comme du "bruit"). Elle ne permet pas non plus d'analyser un type de parcours plutôt qu'un autre.

Nous présentons ici un nouveau cadre permettant d'aborder le problème de l'évaluation de l'exploration de processus en fonction d'indicateurs de performance spécifiques (tel que la capacité à prédire le temps de parcours d'un individu pour achever le processus). Dans ce but, nous étendons la phase de découverte du modèle de sorte qu'à partir d'un *workflow*, nous dérivons un réseau de Petri stochastique visant à reproduire les statistiques pertinentes dans *l'event log*. Il s'agit essentiellement de dériver, pour chaque paire d'événements successifs, une transition temporelle stochastique correspondante qui reproduit adéquatement la chronologie des événements. En appliquant une approche statistique expressive de vérification de modèle, nous évaluons ensuite la conformité (par rapport aux statistiques pertinentes) entre l'*event log* et le modèle de réseau de Petri stochastique correspondant.

La phase de découverte de processus pouvant être très difficile en raison de la complexité des données contenues dans l'*event log*, nous proposons d'appliquer un **clustering de traces** pour résoudre ce problème. Il existe plusieurs algorithmes de regroupement de traces, le plus couramment utilisé étant l'Active Trace Clustering (ActiTraC) [6] qui calcule les regroupements par fréquence. ActiTraC étant fortement dépendant de la taille des traces sur le temps de calcul. Hybrid Trace Clustering [3] est une alternative à ActiTraC utilisant le même système avec des métriques différentes et ajoutant un clustering préliminaire avec des algorithmes de partitionnement comme k-means.

Cependant, il n'existe à notre connaissance, aucun moyen de savoir automatiquement quelle technique de clustering est adaptée ou quel paramètre choisir (comme le nombre de clusters ou le seuil de métrique) en fonction de l'*event log*. Ces paramètres sont généralement choisis manuellement par un expert.

Nous présentons ici une méthode pour guider l'utilisateur dans le choix de la méthode de clustering et dans le nombre de clusters à définir. Notre méthode repose sur la capacité de prédire le comportement de l'*event logs* en comparant le chemin temporel prédit avec le modèle de processus résultant et le temps de parcours réel pour chaque individu. Pour cela, nous utilisons un outil de vérification de modèle quantitatif, appelé Cosmos, basé sur la logique HASL [4]. La section 2 présente notre contribution et la section 3 illustre notre expérimentation sur des ensembles de données et les résultats obtenus. Quelques conclusions sont données dans la section 4.

2 Contribution : Une nouvelle méthode d'évaluation du clustering

Dans notre approche, nous avons cherché à évaluer la qualité des clusters produits par un algorithme en fonction de leur capacité à obtenir des propriétés statistiques conformes à la réalité. Nous sélectionnons dans un premier temps un dataset de données pour l'entraînement et un dataset de test. À partir de l'ensemble de données d'entraînement, nous générons un processus global décrivant tous les événements. Ensuite, nous séparons les traces en différents clusters en choisisant parmi deux méthodes de clustering possibles : ActiTraC ou Hybrid Trace clustering. Chacun des clusters est converti en réseau de Petri stochastique pour faire des prédictions sur les traces, tout en incluant le caractère aléatoire de la transition d'une activité à l'autre (et donc en reflétant mieux la réalité). Une fois les clusters et leurs réseaux appris, nous prenons chaque individu dans l'ensemble de données de test et l'associons au cluster le plus proche (en utilisant comme métrique de comparaison le score f1). Nous simulons leur trace dans le réseau et observons le temps de traitement moyen prédit (à l'aide de l'outil de modèle checking COSMOS couplé à la logique HASL) pour chaque individu à tester. Enfin, nous comparons les différentes configurations (méthode et nombre de clusters) associées aux meilleures performances de prédiction et au score F1. Nous concluons sur le nombre de clusters et le type de méthode à choisir.

3 Matériel et cadre expérimental

Pour évaluer la performance de notre méthode, nous comparons la prédiction de plusieurs clusters à une prédiction globale (sans clustering). Les principaux critères de cette évaluation sont la racine de l'erreur quadratique moyenne (Root Mean Square Error) entre le temps de parcours total réel et prédit ainsi que le score F1. Pour chaque jeu de données utilisé, un échantillon aléatoire de 80 % de l'ensemble de données est utilisé pour l'ensemble d'apprentissage. Cet ensemble de données est utilisé pour le clustering. Les 20% des données non utilisées pour l'entraînement sont utilisées comme base de test pour l'évaluation des clusters extraits. Les expériences sont menées sur deux ensembles de données: Event Log of Hospital qui est un ensemble de données publiques artificielles et BPI Challenge 2013 qui est un ensemble de données publiques.

Paramétrage du clustering Tous les algorithmes de clustering sont dérivés de l'algorithme standard de clustering k-centroïde, qui nécessite le réglage du paramètre k. Le paramétrage des distributions statistiques dans Cosmos ne sera pas abordé ici. Pour évaluer l'impact de ce paramètre sur les résultats, nous avons réalisé des expériences avec différentes valeurs de k (de 2 clusters à 4 clusters, trouvés généralement dans la littérature).

Pour limiter le temps CPU consommé par les algorithmes de clustering et le nombre d'expériences à réaliser, nous fixons le seuil de fitness / f1-score cible et la taille minimale des clusters aux mêmes valeurs pour tous les algorithmes. De cette façon, l'expérimentation se concentre uniquement sur le choix des algorithmes et du nombre de clusters. Les clusters sont ensuite transformés en réseaux de Petri avec un script R et sont utilisés par Cosmos pour le choix de la simulation.

Étape de prédiction L'étape de prédiction de notre méthode consiste à prédire pour tous les patients/incidents non vus, un temps de traitement en utilisant notre *model checking* statistique Cosmos sur les réseaux de Petri associés à leurs clusters. Nous supposons ici que les patients/incidents non vus sont associés aux bons clusters, la sensibilité aux parcours partiels n'étant pas abordée dans cet article. Nous nous intéressons au temps moyen que prend un jeton (c'est-à-dire un individu/incident) pour traverser l'ensemble du *process*. À partir du réseau de Petri, le marquage initial de chaque réseau ne contient qu'un seul jeton. Nous simulons du début à la fin du parcours 1000 fois pour obtenir la moyenne et l'intervalle de confiance de ce temps de parcours.

Nous calculons la *RMSE* du test set et comparons les différentes méthodes/paramètres selon ce critère. Nous détaillons dans le Tab 1 également le RMSE pour chaque cluster afin d'identifier les clusters qui sont de "bons prédicteurs". Le critère de fitness étant le critère habituel d'évaluation des processus extraits, nous le présentons également dans le Tab 1. Les meilleurs résultats sont affichés en bleu/gras et renvoi l'utilisateur vers la meilleure configuration à choisir en fonction de son critère d'optimisation (RMSE ou F1 score).

	Plain	2 Clusters		3 Clusters		4 Clusters	
		ActiTrac	Hybrid	ActiTrac	Hybrid	ActiTrac	Hybrid
RMSE							
Hospital	8.17	7.69	6.53	**3.78**	6.8	5.68	5.12
BPI	180	74.7	202.56	**28.79**	188.13	419.07	131.33
F1-score							
Hospital	0.232	0.56	0.28	0.65	0.25	**0.7**	0.3
BPI	0.92	0.93	0.81	**0.94**	0.84	**0.94**	0.84

Table 1: RMSE et F1 scores obtenus pour les ensembles de données Hospital et BPI

Nous remarquons que le choix d'une configuration en fonction de la RMSE est différent de celui guidé par la F1 score. Si l'objectif de l'utilisateur est de prédire le temps de parcours d'un individu, c'est la RMSE qui sera davantage privilégiée. Ces résultats et d'autres complémentaires sont disponibles sur un gitlab public contenant les scripts R et les commandes Cosmos utilisés dans l'expérimentation : git.lacl.fr/barbot/processminingcosmos.git

4 Conclusion

Nous proposons une nouvelle méthode pour évaluer la configuration d'un clustering en utilisant une simulation et des métriques courantes en *machine learning*. La RMSE guide l'utilisateur vers le bon nombre de clusters et le meilleur algorithme de clustering. C'est une amélioration par rapport aux méthodes traditionnelles nécessitant un expert pour choisir les paramètres. Le RMSE apporte une évaluation orientée sur la capacité des processus finaux à reproduire les événements de l'*event log* au lieu de se concentrer sur la qualité des processus indépendamment de leur capacité prédictive. Dans la partie expérimentale, nous montrons des résultats intéressants, démontrant l'intérêt de RMSE par rapport au F1 score, puisque nous avons pu sélectionner avec plus de justification une configuration appropriée. La partie centrale de la partie expérimentale est basée sur l'utilisation du vérificateur de modèle statistique COSMOS et de la logique HASL. Pour nos travaux futurs, la logique HASL peut être utilisée pour exprimer des métriques beaucoup plus complexes, exprimant des propriétés statistiques spécifiques aux clusters, par exemple des lois de distribution entre deux activités non Gaussiennes.

References

[1] van der Aalst, W., Weijters, T., Maruster, L.: Workflow mining: discovering process models from event logs. IEEE Transactions on Knowledge and Data Engineering **16**(9), 1128–1142 (2004)

[2] van der Aalst, W.: Data Science in Action, pp. 3–23. Springer Berlin Heidelberg, Berlin, Heidelberg (2016). https://doi.org/10.1007/978-3-662-49851-4˙1https://doi.org/\detokenize{10.1007/978-3-662-49851-4_1}

[3] Abdollahi, M.: Hybrid Trace Clustering. Master's thesis, Utrech University (2021)

[4] Ballarini, P., Djafri, H., Duflot, M., Haddad, S., PEKERGIN, N.: Hasl: An expressive language for statistical verification of stochastic models. In: VALUETOOLS'11 (2011)

[5] de Medeiros, A.K.A., Guzzo, A., Greco, G., van der Aalst, W.M.P., Weijters, A.J.M.M., van Dongen, B.F., Saccà, D.: Process mining based on clustering: A quest for precision. pp. 17–29. Springer Berlin Heidelberg (2008)

[6] Weerdt, J., vanden Broucke, S., Vanthienen, J., Baesens, B.: Active trace clustering for improved process discovery. IEEE Trans. Knowl. Data Eng. **25**, 2708–2720 (12 2013)

[7] Weijters, A., Aalst, W., Medeiros, A.: Process Mining with the Heuristics Miner-algorithm, vol. 166 (01 2006)

Model-Based Clustering by Greedy Maximization of the Integrated Classification Likelihood

Nicolas Jouvin [1] & Etienne Côme [2] & Pierre Latouche [3] & Charles Bouveyron [4]

[1] *INRAE, MIA Paris-Saclay, UMR 518, 16 rue Claude Bernard, 75005 Paris, nicolas.jouvin@inrae.fr*
[2] *COSYS/GRETTIA, Univ. Gustave Eiffel, F-77454 Marne-la-Vallée, etienne.come@univ-eiffel.fr*
[3] *Université de Paris, MAP5, UMR CNRS 8145, 45 rue des Saints-Pères, 75006 Paris, pierre.latouche@math.cnrs.fr*
[4] *Université Côte d'Azur, Inria, Cnrs, Laboratoire LJAD, UMR CNRS 7135, Maasai, charles.bouveyron@univ-cotedazur.fr*

Résumé. Nous présentons une nouvelle approche pour le clustering probabiliste. S'appuyant sur une maximisation directe de la vraisemblance classifiante intégrée par rapport à la partition, elle permet de faire conjointement clustering et sélection du nombre de groupes. Ce problème d'optimisation hautement combinatoire est résolu grâce à un algorithme génétique, et une étape finale de clustering hiérarchique permet d'explorer les résultats à différentes échelles ainsi que d'extraire un ordre sur les clusters, utile pour la visualisation. Applicable à une grande variété de modèles, cette méthode convient au clustering de différents types de données, ainsi qu'a des données hétérogènes. La présentation s'appuiera notamment sur le paquet R **greed** qui implémente la méthode pour les principaux modèles utilisés dans la littérature.

Mots-clés. Classification non-supervisée, Clustering, Vraissemblance Classifiante Intégrée, Sélection de modèle, Hiérarchique, Algorithmes génétiques, Paquet R ...

Abstract. We introduce a new, flexible method for model-based clustering. Based on the direct maximization of the exact Integrated Classification Likelihood with respect to the partition, it allows to jointly perform clustering and selection of the number of groups. This combinatorial problem is handled through an efficient genetic algorithm, while a final hierarchical step allows to access coarser partitions and extract an ordering of the clusters. This methodology is applicable in a wide variety of latent variable models and, hence, can handle various data types as well as combination of such yielding heterogeneous data. The presentation will focus on the **greed** R package implementing this method for classical clustering models from Gaussian or Multinomial Mixtures to the Stochastic Block Model.

Keywords. Model-based clustering, Integrated Classification Likelihood, Model selection, Hierarchical, Genetic algorithms, R package ...

1 Introduction

Clustering consists in the unsupervised task of grouping a set of objects into distinct groups or *clusters*. Allowing to unveil relevant structure in datasets, it holds an important part in modern data analysis, with a wide range of applications involving data of different nature. Grounded on a statistical approach, model-based clustering provides a flexible method capable of handling the variety and complexity of modern data such as continuous, count, graphs or mixed data in a common framework using finite mixtures or stochastic block models (Bouveyron et al. 2019).

Consider a set of observations $\boldsymbol{X}$ related to n objects. The latter may be a collection of n vectors in $\mathbb{R}^p$ in the case of mixture modeling, or an $n \times n$ adjacency matrix in the case of graph clustering. The unobserved partition of $\{1, \dots, n\}$ will be denoted as $\boldsymbol{Z} = \{\boldsymbol{z}_1, \dots, \boldsymbol{z}_n\}$ where $\boldsymbol{z}_i$ is a binary vector of size K indicating cluster membership of object i. Model-based clustering can be decomposed in a two-stage generative process. First, the partitions $\boldsymbol{Z}$ is drawn from a product of multinomial distributions with parameter $\boldsymbol{\pi}$, the latter quantifying the a priori probability to belong to each of the K groups. Second, observations are drawn conditionally on the partition, according to some parametric distribution with parameters $\boldsymbol{\theta}$ depending on the clusters assignments. Depending on the context, observations may be continuous or discrete vectors in dimension p, or edges in a graph. Thus, there are a variety of *observational models* that can be handled which such an approach. The discrete latent variable models (DLVMs) framework proposed in Côme et al. (2021) assumes *conditional independence* of the observation given the partition

$$p(\boldsymbol{X}, \boldsymbol{Z} \mid \boldsymbol{\pi}, \boldsymbol{\theta}) = \prod_{\boldsymbol{z} \in \boldsymbol{Z}} p(\boldsymbol{z} \mid \boldsymbol{\pi}) \underbrace{\prod_{\boldsymbol{x} \in \boldsymbol{X}} p(\boldsymbol{x} \mid \boldsymbol{Z}, \boldsymbol{\theta})}_{\text{conditional independence}} , \tag{1}$$

Table 1: A non-exhaustive list of standard observational models handled by the **greed** package: Gaussian mixture model (GMM), Mixture of multinomial (MoM) and stochastic block models (SBM) for (possibly weighted) graphs.

Model name	Observation type	Observational model: $\boldsymbol{X} \mid \boldsymbol{Z}$
GMM	Multivariate continuous: $\boldsymbol{X} \in \mathbb{R}^{n \times p}$	$\boldsymbol{x}_i \mid z_{ik} = 1, \boldsymbol{\theta}_k \sim \mathcal{N}_p(\boldsymbol{x}_i \mid \boldsymbol{m}_k, \boldsymbol{S}_k)$
MoM	Multivariate discrete: $\boldsymbol{X} \in \mathbb{N}^{n \times p}$	$\boldsymbol{x}_i \mid z_{ik} = 1, \boldsymbol{\theta}_k \sim \mathcal{M}_p(\boldsymbol{x}_i \mid \boldsymbol{\theta}_k)$
SBM	Graphs: $\boldsymbol{X} \in \mathbb{R}^{n \times n}$	$x_{ij} \mid z_{ik} z_{jl} = 1, \boldsymbol{\theta}_{kl} \sim p(x_{ij} \mid \boldsymbol{\theta}_{kl})$

Standard frequentist approaches casts clustering as a parameter estimation problem, yielding values $\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\theta}}$, and traditionally dealt with a maximum-likelihood approach solved

with the help of some (variational) EM procedure in most cases. The clustering comes as a byproduct since a partition can then be obtained via the posterior distribution of $\boldsymbol{Z} \mid \boldsymbol{X}, \hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\theta}}$. The choice of the number of clusters K is usually delayed post-inference as a *model selection* problem (Fruhwirth-Schnatter et al. 2019, Chapter 7). There exists a large variety of model selection criteria, generally involving penalized likelihoods (Akaike 1974; Schwarz 1978). In this framework, performing clustering and model selection requires to estimate the parameters for a grid of models indexed by K, which can be quite cumbersome as the computational complexity typically grows with K.

In this work, we focus on an *exact*[1] version of the integrated classification likelihood (ICL, Biernacki et al. 2000), a widely used model selection criterion in the clustering context. Adopting a Bayesian point-of-view on the model parameters $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$, it consists in integrating the latter out hence naturally penalizing for complex models. Introducing a factorized prior $p(\boldsymbol{\theta}, \boldsymbol{\pi}) = p(\boldsymbol{\theta} \mid \boldsymbol{\beta})p(\boldsymbol{\pi} \mid \boldsymbol{\alpha})$ with respective hyper-parameters $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$, the ICL_{ex} writes as:

$$\mathrm{ICL}_{ex}(\boldsymbol{Z}; \boldsymbol{\alpha}, \boldsymbol{\beta}) = \log p(\boldsymbol{X} \mid \boldsymbol{Z}, \boldsymbol{\beta}) + \log p(\boldsymbol{Z} \mid \boldsymbol{\alpha}). \tag{2}$$

The first term of the right hand side is model-dependent, and must be computed on a case-by-case basis with a suitable choice of conjugate prior for the observational model parameters $\boldsymbol{\theta}$. However, putting a symmetric Dirichlet prior $p(\boldsymbol{\pi}) = \mathrm{Dir}(\boldsymbol{\pi} \mid \boldsymbol{\alpha} = (\alpha, \ldots, \alpha))$ over the group proportions, the second term is universal to any DLVM and can be derived exactly thanks to Dirichlet-Multinomial conjugacy. Analytical expressions and detailed computations of ICL_{ex} for the models in Table 1 can be found in the Supplementary Materials of Côme et al. (2021).

2 Genetic algorithm and hierarchy construction

In this presentation, we will describe the two-step methodology proposed by Côme et al. (2021), based on a greedy maximization of ICL_{ex}. The latter first finds a clustering $\boldsymbol{Z}^{(K^\star)}$ with $K^\star$ clusters, and then extracts a set of nested partitions with K varying from $K^\star$ to 1.

2.1 Greedy maximization of exact ICL with a genetic algorithm

We are interested in solving the following highly combinatorial problem

$$\boldsymbol{Z}^{(K^\star)} \in \underset{K, \boldsymbol{Z}}{\arg\max}\, \mathrm{ICL}_{ex}(\boldsymbol{Z}, K), \tag{3}$$

[1] We use the term exact to emphasize the need for an analytic expression of Equation (2), in opposition to the common versions of the ICL which use Laplace approximation to derive a penalized likelihood criterion *à la* BIC

jointly performing clustering and model selection.

Recent approaches for the SBM (Côme and Latouche 2015), or GMMs (Bertoletti et al. 2015), showed that greedy hill-climbing heuristics based on swap and merge moves lead to computationally efficient algorithms, capable of achieving state-of-the-art results while susceptible to spurious local maxima Fruhwirth-Schnatter et al. (2019, p. 137). Côme et al. (2021) build on these idea and proposed a principled approach for any DLVM admitting and tractable ICL_{ex} expression, based on a hybrid Genetic algorithm (GAs Eiben and Smith 2004).

Standard GAs evolve a population of solutions by selecting some of the most promising ones, crossing them, and possibly mutating them until a specified number of generations or some stopping criterion is reached. Their capacity to efficiently explore huge space of solutions makes them ideal candidate for combinatorial, discrete optimization problems such as in Equation (3), and increasingly popular in the context of clustering (Tessier et al. 2006; Hruschka et al. 2009; Andrews and McNicholas 2013). Moreover, in order to improve the exploitation capacity of GAs around local optima, one may hybrid them with efficient local search algorithms (see Eiben and Smith 2004, Chap. 10) such as greedy heuristics based on swap and merge moves.

This is the essence of the hybrid GA proposed by Côme et al. (2021) for solving Equation (3). Since the ICL_{ex} function is invariant under permutations of the cluster indexes (label switching), the integer encoding representation in $\boldsymbol{Z}$ is redundant for our optimization problem. This fact has an important impact on the design of the recombination (crossover) operator for GA. Indeed, simple recombination operators based on crossover points will not consider this particularity and will completely break the structure of the solution (Hruschka et al. 2009). One solution to circumvent this issue is to define the crossover and mutation operators directly over the space of partitions $\{1, \dots, n\}$ into K clusters with a variable K, denoted as $\mathcal{P} = \{\boldsymbol{C}_1, ..., \boldsymbol{C}_K\}$. Such operators will not suffer from the label switching problems and will therefore not breaks the structures already found. The cross-over operator, recombining two "parent" solutions into a new "child" one, is based on the *cross partition* operator. The cross partition of two partitions is simply the partition built by considering all the possible intersections between the elements of the two partitions being crossed. More formally:

$$\mathcal{P}^1 \times \mathcal{P}^2 := \{\mathbf{C}_i^1 \cap \mathbf{C}_j^2 \,,\, \forall i \in \{1, ..., |\mathcal{P}^1|\}, j \in \{1, ..., |\mathcal{P}^2|\}\} \setminus \emptyset,$$

with $\mathcal{P}^1 = \{\mathbf{C}_1^1, ..., \mathbf{C}_{K_1}^1\}$ and $\mathcal{P}^2 = \{\mathbf{C}_1^2, ..., \mathbf{C}_{K_2}^2\}$ two partitions of $\{1, ..., n\}$. This operator produces a new solution which is a refinement of both solutions being crossed, with at most $|\mathcal{P}^1|.|\mathcal{P}^2|$ clusters. It also defines the coarser clustering which can lead to either $\mathcal{P}^1$ or $\mathcal{P}^2$ using merge operations, *i.e.* their first common ancestor in the partition lattice. While this crossing may create superfluous clusters when done near local maxima of ICL_{ex}, the greedy local search based on merges used in the hybrid GA efficiently removes these extra clusters.The remaining aspects to set up regard the selection procedures and the mutation operators. The choices made for these operators are more classical : a classical

rank-based selection policy (see Eiben and Smith 2004, p. 81-82) and a random split mutation operator (split a cluster in two at random).

2.2 Building a hierarchy of ICL-dominant solutions

Once a solution with a number K^* of clusters has been found by one of the previous algorithms, Côme et al. (2021) also provide a hierarchical approach to build a set of ICL dominant solutions for all the value of K between 1 and K^*. This is achieved in the same Bayesian paradigm with the ICL_{ex} objective, now considering the Dirichlet prior parameter α as a regularization controlling the granularity of the clustering and unlocking access to simpler, coarser, solutions as it decreases towards 0. The method is based on a log-linear approximation of ICL_{ex} when α is small, which makes it computationally efficient, and produces a hierarchy of nested partitions along with the sequence of the regularization parameters which enabled the fusions $(\boldsymbol{Z}^{(k)}, \alpha^{(k)})_{k=K^\star,\dots,1}$. Each of these partitions is dominant in term of ICL when α is in the value range between $\alpha^{(k-1)}$ and $\alpha^{(k)}$. Such a hierarchical processing allows for the exploration of the clustering at coarser scales, together with an optimal ordering of the clusters which is a powerful tool both for visual representation of the results and their analysis. This final ordering is computed in an optimal fashion in order to minimize the sum of merge costs between adjacent clusters (Côme et al. 2021, Section 4) thanks to the dynamic programming algorithm of Bar-Joseph et al. (2001). It is particularly useful for the dendrogram representation of the hierarchy and other visualization tools described in Section 3.

3 Examples and numerical results

As emphasized above, this general framework instantiates in any DLVM with tractable ICL_{ex}. The **greed** R package[2] implements the algorithm of Section 2 for various popular models such as GMMS, MoM, SBM, or Latent class analysis, with detailed vignettes for each case. Here, we illustrate the interest of the proposed method on a toy example for graph data clustering with the SBM. We simulate a random SBM graph with $n = 1500$ nodes and a hierarchical cluster structure with 3 clusters each composed of 5 small clusters. The small clusters have an intra-connectivity probability of 0.1 and a probability of connecting a node from the same cluster of 0.025. Otherwise, two random nodes may be connected with a probability of 0.001. Figure 1 illustrates the interest of the hybrid GA, recovering the fine-grained clustering structure with an Adjusted Rand Index (ARI) of 1. Moreover, the evolution of $-\log(\alpha)$ with the number of clusters clearly highlights the coarser structure present at $K = 3$.

[2]Avalaible on the CRAN or at https://comeetie.github.io/greed/

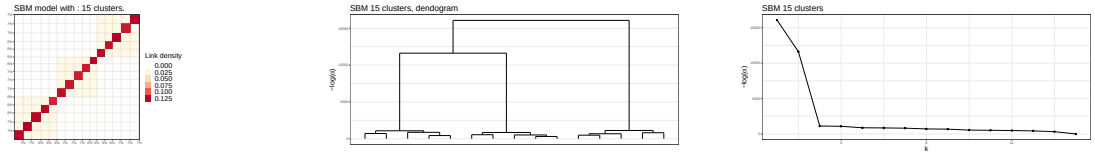


Figure 1: From left to right: aggregated adjacency of the graph according to $\boldsymbol{Z}^{(K^\star)}$ found by the hybrid GA, dendrogram of extracted hierarchy with the level of regularization $-\log(\alpha)$ unlocking fusions, and the evolution of $-\log(\alpha)$ with respect to K in the hierarchy. The optimal ordering of the cluster enhances the adjacency matrix and the dendrogram visualization.

References

Akaike, Hirotugu (1974). "A new look at the statistical model identification". In: *IEEE transactions on automatic control* 19.6, pp. 716–723 (cit. on p. 3).

Andrews, Jeffrey L and Paul D McNicholas (2013). "Using evolutionary algorithms for model-based clustering". In: *Pattern Recognition Letters* 34.9, pp. 987–992 (cit. on p. 4).

Bar-Joseph, Ziv, David K. Gifford, and Tommi S. Jaakkola (June 2001). "Fast optimal leaf ordering for hierarchical clustering ". In: *Bioinformatics* 17.1, S22–S29 (cit. on p. 5).

Bertoletti, Marco, Nial Friel, and Riccardo Rastelli (2015). "Choosing the number of clusters in a finite mixture model using an exact integrated completed likelihood criterion". In: *METRON* 73.2, pp. 177–199 (cit. on p. 4).

Biernacki, Christophe, Gilles Celeux, and Gerard Govaert (2000). "Assessing a mixture model for clustering with the integrated completed likelihood". In: *IEEE Transaction on Pattern Analysis and Machine Intelligence* 7, pp. 719–725 (cit. on p. 3).

Bouveyron, Charles et al. (2019). *Model-Based Clustering and Classification for Data Science: With Applications in R*. Vol. 50. Cambridge University Press (cit. on p. 2).

Côme, Etienne and Pierre Latouche (2015). "Model selection and clustering in stochastic block models based on the exact integrated complete data likelihood". In: *Statistical Modelling* 15.6, pp. 564–589 (cit. on p. 4).

Côme, Etienne et al. (2021). "Hierarchical clustering with discrete latent variable models and the integrated classification likelihood". In: *Advances in Data Analysis and Classification*, pp. 1–30 (cit. on pp. 2–5).

Eiben, A. E. and J. E. Smith (2004). *Introduction to Evolutionary Computing, 2nd Edition*. Springer-Verlag (cit. on pp. 4, 5).

Fruhwirth-Schnatter, Sylvia, Gilles Celeux, and Christian P Robert (2019). *Handbook of mixture analysis*. Chapman and Hall/CRC (cit. on pp. 3, 4).

Hruschka, E. R. et al. (2009). "A Survey of Evolutionary Algorithms for Clustering". In: *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 39.2, pp. 133–155 (cit. on p. 4).

Schwarz, Gideon (1978). "Estimating the dimension of a model". In: *The annals of statistics* 6.2, pp. 461–464 (cit. on p. 3).

Tessier, Damien et al. (2006). *Evolutionary latent class clustering of qualitative data*. Tech. rep. (cit. on p. 4).

A Dynamic Latent Block Model for Co-clustering of Zero-Inflated Count Data Streams

Giulia Marchello[1], Marco Corneli[2] & Charles Bouveyron[1]

[1] *Université Côte d'Azur, Inria, CNRS, Laboratoire J.A.Dieudonné, Maasai team, Nice, France.*

[2] *Université Côte d'Azur, Inria, Maison de la Modélisation des Simulations et des Interactions (MSI), Maasai team, Nice, France.*

Résumé. Le clustering simultané des observations et des variables explicatives des jeux de données (connu sous le nom de co-clustering) est récemment apparu comme une application centrale de l'apprentissage automatique pour résumer des jeux de données massifs. Ce travail introduit un nouveau modèle à blocs latents pour le co-clustering dynamique de flux de données de comptage avec une forte sparsité. Pour modéliser correctement ce type de données, nous supposons que les observations suivent un mélange de lois de Poisson "zéro inflated" dépendantes du temps et des blocs. Afin de modéliser et de détecter les changements brusques dans la dynamique de l'appartenance aux clusters et de la dispersion des données, les proportions de mélange et de sparsité sont modélisées par des systèmes d'équations différentielles ordinaires. L'inférence du modèle est basée sur une procédure variationnelle originelle dont l'étape de maximisation entraîne des réseaux neuronaux récurrents afin de résoudre les systèmes dynamiques. Des expériences numériques sur des ensembles de données simulées démontrent l'efficacité de la méthodologie proposée.

Mots-clés. Co-clustering, modèle à blocs latents, distribution de Poisson zéro inflated, systèmes dynamiques, algorithme VEM.

Abstract. The simultaneous clustering of observations and features of data sets (known as co-clustering) has recently emerged as a central machine learning application to summarize large data sets. This work introduces a novel latent block model for the dynamic co-clustering of count data streams with high sparsity. To properly model this type of data, we assume that the observations follow a time and block dependent mixture of zero-inflated Poisson distributions. To model and detect abrupt changes in the dynamics of both clusters memberships and data sparsity, the mixing and sparsity proportions are modeled through systems of ordinary differential equations. The model inference relies on an original variational procedure whose maximization step trains recurrent neural networks in order to solve the dynamical systems. Numerical experiments on simulated data sets demonstrate the effectiveness of the proposed methodology.

Keywords. Co-clustering, latent block model, ZIP distribution, dynamic systems, VEM algorithm.

1 Introduction

In data science there is an increasing need to group observations and features of high dimensional sparse matrices simultaneously. Machine learning models that allow to perform this kind of operations are called co-clustering methods (Govaert and Nadif, 2003) and are used in a wide range of applications (e.g. natural language processing, recommending systems, biomedical data, etc.). However, in an increasing number of applications it happens to deal with highly sparse data, this may make the co-clustering task challenging in terms of detection of signals. In this paper, we focus on sparse counting datasets and we assume that the observations follow a time and block dependent mixture of zero-inflated Poisson distributions (Lambert, 1992), thus combining two independent processes: a dynamic mixture of Poisson distributions and a time-dependent sparsity process. Performing co-clustering of temporal sparse count matrices is make amenable by the introduction of two evolving multinomial latent variables that model the group memberships of observations and features, respectively. Moreover, the parameters of the two latent variables and the sparsity proportions arise from three systems of ordinary differential equations that define their evolution over time. The proposed model is able to detect abrupt changes in the progression of interactions between observations and features as each element of the matrix is allowed to change its cluster membership in the considered time period. Thus, the interpretation of the results is highly intuitive: when there is a change in the affiliation of the elements to the clusters it means that a breaking point has been detected, giving space for further analysis to verify the reason.

2 Stream-dLBM

The aim of the proposed model is to perform co-clustering of dynamic sparse counting matrices in such a way to capture the possible evolution in the behaviors of both observations and features. We therefore aim at simultaneously clustering the rows and columns of an evolving counting data table $X(t) = \{X_{ij}(t)\}_{i\in 1,\dots,M, j\in 1,\dots,P}$, with M rows and P columns over the partitioned time interval $[0, T]$:

$$0 < \cdots < t < t+1 < \cdots < T-1 < T$$

The first assumption of the model concerns the time-evolving sparsity. We rely on a Bernoulli time-dependent latent matrix, A, of dimension $M \times P$, where:

$$A \sim \mathcal{B}(\pi(t)), \qquad \text{such that:} \qquad A_{ij}(t) = \begin{cases} 1 & \text{if } X_{ij}(t) = 0 \\ 0 & \text{if } X_{ij}(t) > 0 \end{cases} \tag{1}$$

Then, the main idea is that the observations and features in $X(t)$ are respectively clustered into Q and L groups, whose numbers do not change in time. However, each row and column is allowed to switch cluster, in $[0, T]$. More formally, the latent structure of

rows and columns of X is identified by two evolving multinomial distributions, respectively parametrized by $\alpha(t)$ and $\beta(t)$:

$$Z_i(t) \sim \mathcal{M}(1, \alpha(t) := (\alpha_1(t), \dots, \alpha_Q(t))), \tag{2}$$

where $\mathcal{M}(1, \cdot)$ denotes the multinomial probability mass function with parameters $(1, \cdot)$ and $\alpha_q(t) = \mathbb{P}\{z_{iq}(t) = 1\}$, with $\sum_{q=1}^{Q} \alpha_q(t) = 1$. Thus $Z(t) := \{z_{iq}(t)\}_{i \in 1,\dots,M, q \in 1,\dots,Q}$ represents the clustering of the rows into Q groups at a given time t. In a similar fashion, for the column clusters, we assume:

$$W_j(t) \sim \mathcal{M}(1, \beta(t) := (\beta_1(t), \dots, \beta_L(t))), \tag{3}$$

where $\beta_\ell(t) = \mathbb{P}\{w_{j\ell}(t) = 1\}$, with $\sum_{\ell=1}^{L} \beta_\ell(t) = 1$.

A third assumption directly concerns the observations. As mentioned above, the data are assumed to be modeled as a mixture of Zero-Inflated Poisson (ZIP) distributions,namely:

$$X_{ij}(t) \sim ZIP(\Lambda_{Z_i(t), W_j(t)}, \pi(t)). \tag{4}$$

Being a mixture of a Poisson process and a binary process, the ZIP distribution is used to account for excess of zeros in the data. That is:

$$\begin{cases} X_{ij}(t) \sim 0 & \text{with probability } \pi(t) \\ X_{ij}(t) \sim \mathcal{P}(\Lambda_{Z_i(t), W_j(t)}) & \text{with probability } 1 - \pi(t) \end{cases} \tag{5}$$

Finally, t to model the clusters proportions and the sparsity evolution over time, we assume that the mixing parameters, $\alpha(t)$ and $\beta(t)$, and the sparsity proportions, $\pi(t)$, are respectively related with a system of ordinary differential equations (ODEs). In this way, we are able to both capture the evolution of the clusters composition at each time, and take into account the evolution in the number of missing data. Thus, the three systems of discretized ODEs are:

$$\alpha_q(t) = \text{softmax}(a_q(t)) = \frac{e^{a_q(t)}}{\sum_{q=1}^{Q} e_q^a(t)}, \tag{6}$$

$$\beta_\ell(t) = \text{softmax}(b_\ell(t)) = \frac{e^{b_\ell(t)}}{\sum_{\ell=1}^{L} e_\ell^b(t)}, \tag{7}$$

$$\pi(t) = \text{softmax}(p(t)) = \frac{e^{p(t)}}{e^{p(t)} + e^{(1-p(t))}}. \tag{8}$$

with

$$a(t) = a(t-1) + f_Z(a(t-1)), \qquad \forall t \in \{1, \dots, T\}, \tag{9}$$

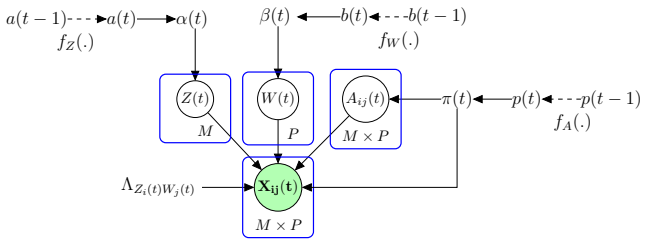


Figure 1: Graphical representation of stream-dLBM.

$$b(t) = b(t-1) + f_W(b(t-1)), \qquad \forall t \in \{1, \dots, T\}, \tag{10}$$

$$p(t) = p(t-1) + f_A(p(t-1)), \qquad \forall t \in \{1, \dots, T\}. \tag{11}$$

where $f_Z(\cdot)$, $f_W(\cdot)$ and $f_A(\cdot)$ are three one-layer fully connected neural networks. Figure 1 shows the graphical representation of stream-dLBM.

2.1 The joint distribution

The model described so far has a set of parameter denoted by $\theta = (\Lambda, a(t), b(t), p(t))$ and a set of latent variables: $Z(t)$, $W(t)$, $A(t)$. As a first move, we can compute the likelihood of the complete data:

$$p(X, Z, W, A|\theta) = p(X|Z, W, A, \Lambda, p)p(A \mid p)p(Z|a)p(W|b) \tag{12}$$

where:

$$p(X|A, Z, W, \Lambda, p) = \prod_{i=1}^{M}\prod_{j=1}^{P}\prod_{t=1}^{T}\left\{\text{soft}(p(t))^{A_{ij}(t)}\left((1-\text{soft}(p(t))\frac{\Lambda_{Z_i(t)W_j(t)}^{X_{ij}(t)}}{X_{ij}(t)!}\exp(-\Lambda_{Z_i(t)W_j(t)})\right)^{(1-A_{ij}(t)}\right\}, \tag{13}$$

$$p(A|p) = \prod_{i=1}^{M}\prod_{j=1}^{P}\prod_{t=1}^{T}\text{soft}(p(t))^{A_{ij}(t)}\Big(1-\text{soft}(p(t))\Big)^{(1-A_{ij}(t))}, \tag{14}$$

$$p(Z|a) = \prod_{i=1}^{M}\prod_{q=1}^{Q}\prod_{t=1}^{T}\text{soft}(a_q(t))^{Z_{iq}(t)}, \tag{15}$$

$$p(W|b) = \prod_{j=1}^{P}\prod_{\ell=1}^{L}\prod_{t=1}^{T}\text{soft}(b_\ell(t))^{W_{j\ell}(t)}. \tag{16}$$

3 The inference algorithm

Regarding the parameters estimation, the traditional procedure for such a model would be to maximize the log-likelihood $p(X|\theta)$. However, for Z and W, neither the direct maximization nor the classical EM algorithm (Dempster et al., 1977; Bishop, 2006a) are feasible in our case because of the impossibility to compute the posterior distribution of the latent variables, $p(Z, W|X, \theta)$, due to their double missing structure. This is why we rely on the Variational-EM algorithm. On the other hand, it is possible to compute the posterior distribution of A, hence for its estimation we can rely on the classical E-step of the EM algorithm. Regarding the M-step, we obtain an explicit formula only for the estimation of Λ, while for $a(t)$, $b(t)$ and $p(t)$ it is not possible due to their association with the respective systems of ordinary differential equations. Therefore, in order to optimize $a(t)$, $b(t)$ and $p(t)$ we employed a stochastic gradient descend algorithm for numerical optimization. Figure 2 shows the complete pseudo-code of the proposed inference algorithm.

Algorithm 1 VEM-SGD Algorithm

Require: $X, Q, L, max_iter, H, K, V$, (where H, K, V are the hidden layer sizes.)
Initialization of τ and η with kmeans.
Initialization of $\delta(t)$ sampling from $\mathcal{B}(\pi(t))$
Initialization of $\alpha_q(t) = 1/Q$, $\beta_\ell(t) = 1/L$, $\pi(t) = 1/2$
M-Step:
$\hat{\Lambda}_{q\ell} = \frac{\sum_{i,j,t}\left\{\tau_{iq}(t)\eta_{j\ell}(t)\left(X_{ij}(t)-\delta_{ij}(t)X_{ij}(t)\right)\right\}}{\sum_{i,j,t}\left\{\tau_{iq}(t)\eta_{j\ell}(t)\left(1-\delta_{ij}(t)\right)\right\}}$.
for epoch = 1 to Epochs **do**
  Update $\hat{\alpha}(t), \hat{\beta}(t), \hat{\pi}(t)$:
  Loss Evaluation;
  Algorithm backpropagation;
  Numerical optimization with SGD.
end for
E-Step:
Updatings of $\delta_{ij}(t)$:

$$\begin{cases} \text{if } X_{ij}(t) = 0 \Longrightarrow \delta_{ij}(t) & = \frac{p(t)}{p(t)+(1-p(t))\exp(-\Lambda_{Z_i(t)W_j(t)})} \\ \text{if } X_{ij}(t) > 0 \Longrightarrow \delta_{ij}(t) & = 0 \end{cases}$$

VE-Step:
Updatings of $\tau(t)$ and $\eta(t)$ through:

$$\tau_{iq}(t) = \frac{1}{N_q}\exp\left(\sum_{j,\ell}\left\{\eta_{j\ell}(t)X_{ij}(t)\log(\Lambda_{q\ell}) - \eta_{j\ell}(t)\Lambda_{q\ell} - \delta_{ij}(t)\eta_{j\ell}(t)X_{ij}(t)\log(\Lambda_{q\ell}) + \delta_{ij}(t)\eta_{j\ell}(t)\Lambda_{q\ell}\right\} + a_q(t)\right).$$

$$\eta_{j\ell}(t) = \frac{1}{N_\ell}\exp\left(\sum_{i,q}\left\{\tau_{iq}(t)X_{ij}(t)\log(\Lambda_{q\ell}) - \tau_{iq}(t)\Lambda_{q\ell} - \delta_{ij}(t)\tau_{iq}(t)X_{ij}(t)\log(\Lambda_{q\ell}) + \delta_{ij}(t)\tau_{iq}(t)\Lambda_{q\ell}\right\} + b_\ell(t)\right).$$

Figure 2: Pseudo-code of the VEM-SGD algorithm.

4 Experiment on simulated data

The main purpose of this section is to highlight the main features of the proposed model over a simulated data set. We aim at demonstrating the validity of the inference algorithm presented on the previous sections. A simulated data set was used to perform this experiment.Specifically, the simulated dynamics of $\alpha(t)$, $\beta(t)$and $\pi(t)$ can be seen

in the left-hand side of Figure 3. Based on these mixture proportions, the values of the latent variables were then simulated through their distributions. Next, based on the sparsity proportions, $\pi(t)$ and on the intensity function Λ, the three-dimensional tensor X was simulated as a Zero-Inflated Poisson variable. We fitted stream-dLBM to simulated dataset with the actual values for $Q = 3$ and $L = 2$ to show the ability of the model to fully recover the model parameters. To fully understand the scope of stream-dLBM one can look at Figure 3, where the evolution of $\hat{\alpha(t)}$, $\hat{\beta(t)}$ and $\pi(t)$ over time is depicted. In particular, in the left-hand side the true mixture proportions are shown while in the right-hand side there are the estimated ones. Finally, we make use of the adjusted Rand index (ARI) (Rand, 1971) as a measure to evaluate the performance of the model in identifying the correct row and column partitions for all the time instants. The closer the index is to 1, the more two label vectors are similar to each other. We compared the original partitions with the estimated ones. Figure 4 shows that we obtained an ARI index of 1 for rows and columns partitions at each time instant t. This result is totally coherent with the other outcomes, since the dynamic composition of each cluster is well retrieved by the process.

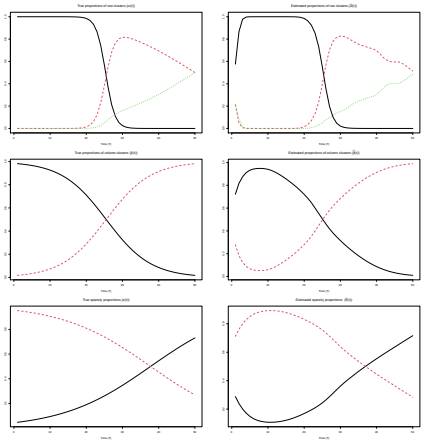


Figure 3: Comparison between the evolution of the true parameters (left side) and the estimated ones (right side).

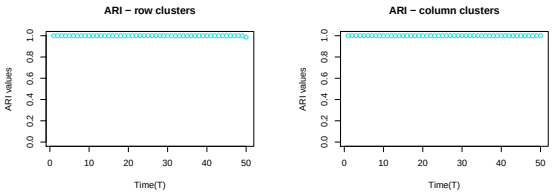


Figure 4: ARI values.

References

Bishop, C. M. (2006a). Approximate inference. In Bishop (2006b), pages 461–517.

Bishop, C. M. (2006b). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

Govaert, G. and Nadif, M. (2003). Clustering with block mixture models. *Pattern Recognition*, 36(2):463–473.

Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1–14.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850.

False clustering rate control in mixture models

Ariane Marandon[1], Tabea Rebafka[1], Etienne Roquain[1], Nataliya Sokolovska[2]

[1] *LPSM, Sorbonne Université, ariane.marandon-carlhian@sorbonne-universite.fr, tabea.rebafka@sorbonne-universite.fr, etienne.roquain@sorbonne-universite.fr*
[2] *NutriOmics, Sorbonne Université, nataliya.sokolovska@sorbonne-universite.fr*

Résumé. Le clustering consiste à partitioner les membres d'un échantillon en groupes. Cependant, pour la plupart des jeux de données, certains individus sont ambigus et intrinséquement difficiles à classer. De plus, une mauvaise classification des individus peut se révéler désastreuse dans certains cas pratiques. Pour surmonter cette difficulté, l'idée suivie ici est de ne classer qu'une partie de l'échantillon afin de maintenir un taux de fausses classifications assez bas. Cette approche est bien connue dans le cadre supervisée et est appelée classification avec abstention. L'objectif de cet article est de revisiter cette approche dans le cadre non supervisée d'un modèle de mélange. Le problème est formalisé en termes de contrôle du taux moyen de faux classés en dessous d'un niveau préscrit α, tout en maximisant le nombre d'éléments classés. Nous introduisons à cet effet de nouvelles procédures, qui montre un comportement proche de l'optimum à la fois théoriquement et sur des expériences numériques. Une application à un jeu de données sur le cancer du sein illustre les avantages de la nouvelle approche d'un point de vue pratique.

Mots-clés. Classification avec abstention, clustering, FDR, modèles de mélange.

Abstract. The clustering task consists in delivering labels to the members of a sample. However, for most data sets, some individuals are ambiguous and intrinsically difficult to attribute to one or another cluster. Moreover, in practical applications, misclassifying individuals is potentially disastrous. To overcome this difficulty, the idea followed here is to classify only *a part* of the sample in order to maintain a low misclassification rate. This approach is well known in the supervised setting, and referred to as classification with an abstention option. The purpose of this paper is to revisit this approach in an unsupervised mixture-model framework. The problem is formalized in terms of controlling the false clustering rate (FCR) below a prescribed level α, while maximizing the number of clustered items. New procedures are introduced and their behavior is shown to be close to the optimal one by establishing theoretical results and conducting numerical experiments. An application to breast cancer data illustrates the benefits of the new approach from a practical viewpoint.

Keywords. Classification with abstention, clustering, FDR, mixture models.

1 Introduction

Clustering is a standard statistical task that aims at grouping individuals with similar features. However, it is common that data sets include a number of ambiguous individuals that are inherently difficult to classify, which makes the clustering result potentially unreliable. A standard solution is to avoid classification for such ambiguous individuals, and to adopt a wiser "abstention decision", that leaves the door open for further analysis. In a supervised setting, classification with rejection is a long-standing paradigm (**??**). The goal of the present work is to propose a labelling guarantee on the classified items in the more challenging unsupervised setting, where no training set is available and data are assumed to be generated from a finite mixture model. This is achieved by the possibility to refuse to cluster ambiguous individuals. Hence, in our context, a procedure is based both on a clustering and on a selection of the individuals to be classified. Our labelling guarantee is given in terms of the false clustering rate (FCR), defined as the average proportion of misclassifications among the classified data points. Thus, our objective is to build a procedure that keeps the FCR below some nominal level α, while classifiying a maximum number of data points. In a nutshell, our method thresholds the (estimated) posterior probabilities of the cluster labels with a data-driven choice of the threshold. Let us finally mention that the oracle part of our work is similar to the result of **?**. We did a number of new contributions, as we describe below.

Contributions We introduce three procedures: the plug-in procedure and two bootstrap procedures (parametric and non-parametric). We prove that the plug-in procedure controls the FCR at the desired level α, up to a reminder term which becomes small when the sample size grows. In addition, this procedure is shown to satisfy the following optimality property: any other procedure that provides an FCR control necessarily classifies less items than the plug-in procedure, up to a small remainder term. We also put forward specific convergence rates depending on the parameter estimation accuracy. Numerical experiments establish that the bootstrap procedures improve the plug-in procedure in terms of non-asymptotic FCR control, and thus are more reliable for practical use, where the sample size may be moderate.

2 Setting

Model Let $\mathbf{X} = (X_1, \ldots, X_n)$ be an observed random sample of size n. Each X_i is an i.i.d. copy of a d-dimensional real random vector, which is assumed to follow the standard mixture model:

$$\begin{aligned} Z &\sim \mathcal{M}(\pi_1, \ldots, \pi_Q), \\ X|Z = q &\sim F_{\phi_q}, \quad 1 \leq q \leq Q, \end{aligned}$$

where $\{F_u, u \in \mathcal{U}\}$ is a collection of distributions on $\mathbb{R}^d$. The overall parameter is thus $\theta = (\pi, \phi)$. The latent vector $\mathbf{Z} = (Z_1, \ldots, Z_n)$ encodes a partition of the n observations into Q classes given by $\{1 \leq i \leq n \ : \ Z_i = q\}$, $1 \leq q \leq Q$. The number of classes Q is assumed to be known. In the sequel, we let θ^* denote the true model parameter.

False clustering rate Our approach starts with a given clustering rule $\widehat{\mathbf{Z}} = (\hat{Z}_i)_{1 \leq i \leq n} \in \{1, \ldots, Q\}$, that aims at recovering the true latent clustering. The classification error of $\hat{Z}$ is thus given by $\varepsilon(\widehat{\mathbf{Z}}, \mathbf{Z}) = \sum_{i=1}^{n} \mathbb{1}\{Z_i \neq \hat{Z}_i\}$. Obviously, since switching the labels in $\mathbf{Z}$ does not change the true latent partition, we should allow for switching the labels when making the decision $\widehat{\mathbf{Z}}$ from the data. The clustering risk of $\widehat{\mathbf{Z}}$ is thus defined by

$$R(\widehat{\mathbf{Z}}) = \mathbb{E}_{\theta^*}\left(\min_{\sigma \in [Q]} \mathbb{E}_{\theta^*}\left(n^{-1}\varepsilon(\sigma(\widehat{\mathbf{Z}}), \mathbf{Z}) \mid \mathbf{X}\right)\right),$$

where $[Q]$ denotes the set of all permutations on $\{1, \ldots, Q\}$. Classically, we aim to find a clustering rule $\widehat{\mathbf{Z}}$ such that the clustering risk is "small". However, as mentioned above, whether this is possible or not depends on the intrinsic difficulty of the clustering problem and thus of the true parameter θ^*. Therefore, the idea is to provide a selection rule, that is, a (measurable) function of the observation $\mathbf{X}$ returning a subset $S \subset \{1, \ldots, n\}$, such that the clustering risk *with restriction to* S is "small". In our framework, a *procedure* refers to a couple $\mathcal{C} = (\widehat{\mathbf{Z}}, S)$, where $\widehat{\mathbf{Z}}$ is a clustering rule and S is a selection rule. The false clustering rate (FCR) of a procedure $\mathcal{C} = (\widehat{\mathbf{Z}}, S)$ is given by

$$\mathrm{FCR}_{\theta^*}(\mathcal{C}) = \mathbb{E}_{\theta^*}\left(\min_{\sigma \in [Q]} \mathbb{E}_{\theta^*}\left(\frac{\varepsilon_S(\sigma(\widehat{\mathbf{Z}}), \mathbf{Z})}{\max(|S|, 1)} \,\middle|\, \mathbf{X}\right)\right),$$

where $\varepsilon_S(\widehat{\mathbf{Z}}, \mathbf{Z}) = \sum_{i \in S} \mathbb{1}\{Z_i \neq \hat{Z}_i\}$ denotes the mis-classification error restricted to the subset S. In this work, the aim is to find a procedure $\mathcal{C}$ such that the false clustering rate is controlled at a nominal level α, that is, $\mathrm{FCR}_{\theta^*}(\mathcal{C}) \leq \alpha$, while classifying as much individuals as possible, that is, with $\mathbb{E}_{\theta^*}|S|$ as large as possible.

The FCR involves an expectation of a ratio, which is more difficult to handle that a ratio of expectations. Hence, we also consider the *marginal* false clustering rate (mFCR) of a procedure $\mathcal{C} = (\widehat{\mathbf{Z}}, S)$, given by

$$\mathrm{mFCR}_{\theta^*}(\mathcal{C}) = \frac{\mathbb{E}_{\theta^*}\left(\min_{\sigma \in [Q]} \mathbb{E}_{\theta^*}\left(\varepsilon_S(\sigma(\widehat{\mathbf{Z}}), \mathbf{Z}) \,\middle|\, \mathbf{X}\right)\right)}{\mathbb{E}_{\theta^*}(|S|)},$$

with the convention $0/0 = 0$.

3 Methods

In this section, we introduce new methods for controlling the FCR. We start by identifying an oracle method, in the sense that it uses the true value of the parameter θ^*. Substituting in that oracle the unknown parameter θ^* by an estimator provides our first method, called the plug-in procedure. We then define a refined version of the plug-in procedure, that accounts for the variability of the estimator and is based on a bootstrap approach.

Oracle procedure We start by noting that the Maximum a Posteriori (MAP) clustering $\widehat{\mathbf{Z}}^* = (\text{argmax}_q \mathbb{P}_{\theta^*}(Z_i = q|\mathbf{X}))_{1\leq i\leq n}$ is a minimizer of the clustering risk. Therefore, a starting point is to consider the MAP clustering output, and we aim to construct a selection rule that will classify, or not, data points according to this output. In a mixture model, confidence in the clustering is expressed through the posterior probability that a data point belongs to a given cluster. Then, if one wishes to classify a subset of data points so that the averaged proportion of misclassifications is below α, a natural selection rule (referred to as the fixed threshold rule in the sequel) is to only keep those points for which the confidence in the clustering label is higher than $1-\alpha$. While this does yield a valid procedure, i.e. with an FCR below α, we will see that this is, however, sub-optimal. To derive a procedure that is both valid and more powerful, we make use of the following observation. Let T_i^* be the posterior mis-classification probabilities of the MAP clustering, given by

$$T_i^* = \mathbb{P}_{\theta^*}(Z_i \neq \hat{Z}_i^*|X_i) = 1 - \max_q \mathbb{P}_{\theta^*}(Z_i = q|X_i), \;\; i \in \{1,\dots,n\}.$$

A high value of T_i^* indicates that i is difficult to classify, that is, the posterior class membership probabilities are all close to each other. If we work with the MAP clustering, for a selection rule S, the FCR of $\mathcal{C} = (\widehat{\mathbf{Z}}^*, S)$ is given by

$$\text{FCR}_{\theta^*}(\mathcal{C}) = \mathbb{E}_{\theta^*}\left(\frac{\sum_{i\in S} T_i^*}{\max(|S|,1)}\right).$$

Hence, choosing any selection rule $S \subset \{1,\dots,n\}$ satisfying $\sum_{i\in S} T_i^* \leq \alpha|S|$ controls the FCR at level α. Maximizing the latter in the selection number leads to the *oracle* procedure. It uniformly improves the procedure with fixed threshold and will in general lead to a (much) broader selection.

Plug-in procedure The oracle procedure cannot be used in practice since the true parameter θ^* of the mixture model is generally unknown. A natural idea is then to approach θ^* by an estimator $\hat{\theta}$ and then to plug this estimate into the oracle procedure. We establish that the plug-in procedure has suitable properties: when the sample size tends to infinity, provided that the chosen estimator $\hat{\theta}$ behaves well and under mild regularity assumptions on the model, the FCR of the plugin procedure close to the level α while it is close to be optimal in terms of selection number.

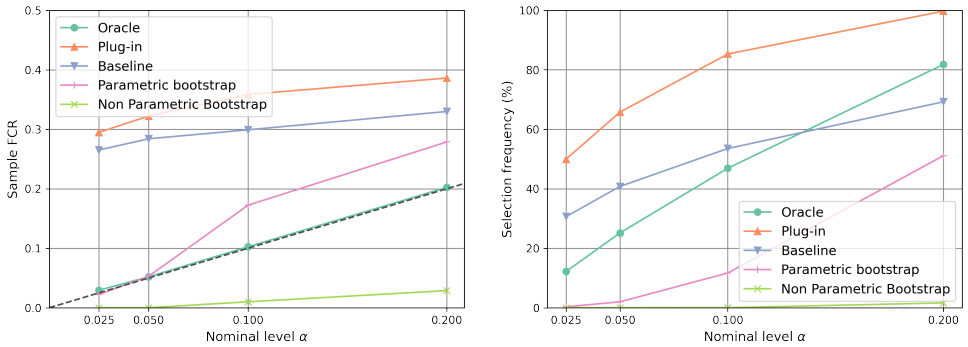
Bootstrap procedures Despite its favorable theoretical properties, the plug-in procedure achieves an FCR that can exceed α in some situations, as we will see in our numerical experiments. This is in particular the case when the estimator $\hat{\theta}$ is too rough. Indeed, the uncertainty of $\hat{\theta}$ near θ^* is ignored by the plug-in procedure. To take into account this effect, we propose to use a bootstrap approach. The general idea is as follows: since the FCR of the plug-in can exceed α, we choose α' as large as possible such that $\widehat{\mathrm{FCR}}_{\alpha'} \leq \alpha$, for which $\widehat{\mathrm{FCR}}_{\alpha'}$ is a bootstrap approximation of the FCR of the plug-in procedure applied at level α'. For the bootstrap approximation, we generate bootstrap samples and compute for each, an unbiased point estimate of the FCR; and then take the average. This point estimate is computed by replacing the true distribution by the estimator $\hat{\theta}$. The bootstrap samples are generated using either the parametric bootstrap or the non-parametric bootstrap.

4 Experiments

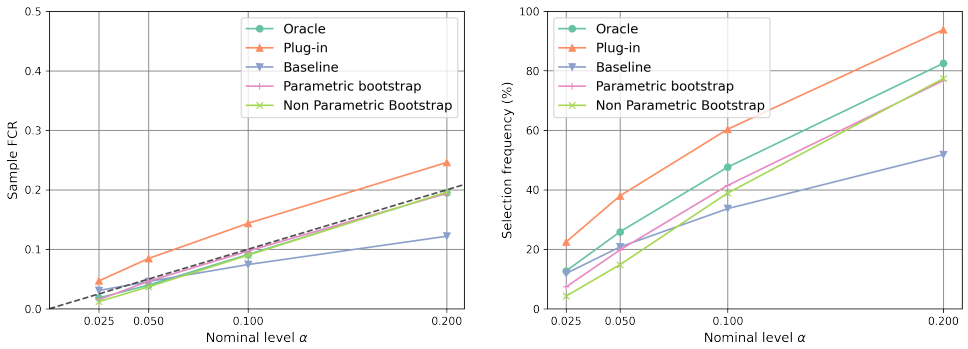
In this section, we evaluate the behavior of the new procedures: plug-in, parametric bootstrap and non parametric bootstrap. For this, we use both synthetic and real data. As a baseline, we consider the fixed threshold procedure in which one selects data points that have a maximum posterior group membership probability that exceeds $1 - \alpha$.

Synthetic data set We consider gaussian mixture models with varying restrictions on the parameter space. Figure 1 displays the results for a three-component model with $d = 4$. It shows that the plug-in procedure is anti-conservative. This comes from the parameter estimation, which is difficult in that case. The bootstrap procedures provide a correction. Specifically, the parametric bootstrap procedure has a FCR that is close to α for a large sample size. For small n the non-parametric bootstrap provides stricter FDR control, although it is very conservative in this case. Overall, we note that for all our procedures (including the plug-in), the gap between the FCR and the nominal level is roughly constant with α: this illustrates the adaptive aspect of our procedures. This is in contrast with the baseline procedure, for which this gap highly depends on α, and which may be either anti-conservative or sub-optimal depending on the α value.

Real data set We consider the Wisconsin Breast Cancer Diagnosis (WCBD) dataset from the UCI ML repository. We model the data as a mixture of Students as proposed in Peel and McLachlan (2000). Compared to Gaussian components, Students are less concentrated and thus produce estimates of the posterior probabilities of class memberships that are less extreme. We set the degree of freedom of each component to 4, and no constraints are put on the rest of the parameters. In what follows, we restrict the analysis to two variables of the dataset, the mean radius and the mean texture of the images. For illustration, Figure 2a displays the data. Different colors indicate the ground truth labels.

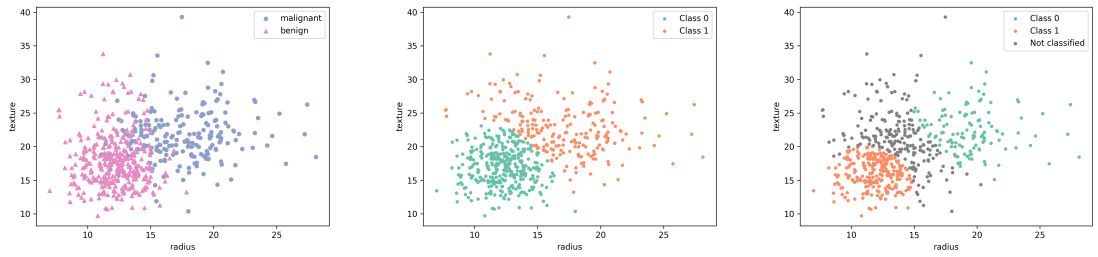


(a) $n = 200$



(b) $n = 1000$

Figure 1: FCR (left panel) and selection frequency (right panel) as a function of the nominal level α. Here $Q = 3$, $d = 4$, $\pi_1 = \pi_2 = \pi_3 = 1/3$, $\Sigma_1 = \Sigma_2 = \Sigma_3 = I_4$, the mean separation is $\|\mu_1 - \mu_2\|_2 = \|\mu_1 - \mu_3\|_2 = \|\mu_2 - \mu_3\|_2 = 2$, and $n \in \{200, 1000\}$.



(a) Ground truth labels

(b) MAP clustering

(c) Clustering with bootstrap selection for $\alpha = 5\%$

Figure 2: Comparison of the clustering result with t-mixture modelling with ground truth labels on the WCBD dataset, restricted to the variables *mean radius* and *mean texture*, with and without selection. With the parametric bootstrap procedure applied at $\alpha = 5\%$, the FCR w.r.t. the ground truth labels is of 3% versus 14% without selection.

One can see that the Student approximation is fairly good for each of the groups, and there is some overlap between them. Figure 2b displays the MAP clustering result for the t-mixture model without any selection. The FCR is computed with respect to the ground truth labels and amounts to 14%. Finally, Figure 2c provides the result of our parametric bootstrap procedure with nominal level $\alpha = 5\%$. The procedure rejects points that are at the intersection of the clusters and the FCR equals 3%, which is below the targeted level.

References

Peter L. Bartlett and Marten H. Wegkamp. Classification with a reject option using a hinge loss. *Journal of Machine Learning Research*, 9(59):1823–1840, 2008. URL `http://jmlr.org/papers/v9/bartlett08a.html`.

Christophe Denis and Mohamed Hebiri. Consistency of plug-in confidence sets for classification in semi-supervised learning. *Journal of Nonparametric Statistics*, 32(1):42–72, 2020. doi: 10.1080/10485252.2019.1689241. URL `https://doi.org/10.1080/10485252.2019.1689241`.

Tristan Mary-Huard, Vittorio Perduca, Gilles Blanchard, and Martin-Magniette Marie-Laure. Error rate control for classification rules in multiclass mixture models, 2021.

Transport optimal

Transport optimal pour l'appariement automatique de données métabolomiques non ciblées

Marie Breeur[1], George Stepaniants[2], Philippe Rigollet[3], Vivian Viallon[4]

[1] *International Agency for research on Cancer, Nutrition and metabolism Branch*
150 Cours Albert Thomas, 69372 Lyon cedex 08
breeurm@iarc.fr

[2] *Massachusetts Institute of Technology, Department of Mathematics and Institute for Data Systems and Society (IDSS)*
77 Massachusetts Ave, Cambridge MA 02139
gstepan@mit.edu

[3] *Massachusetts Institute of Technology, Department of Mathematics*
77 Massachusetts Ave, Cambridge MA 02139
rigollet@math.mit.edu

[4] *International Agency for research on Cancer, Nutrition and metabolism Branch*
150 Cours Albert Thomas, 69372 Lyon cedex 08
viallonv@iarc.fr

Résumé.

Le profilage métabolomique non ciblé permet de mesurer une large gamme de métabolites présents dans un échantillon biologique. Cependant, le nombre d'échantillons mesurés dans les études métabolomiques non-ciblées est généralement faible. Une solution pour augmenter le pouvoir statistique des études métabolomiques consisterait à combiner ou méta-analyser les données issues de plusieurs études. Or, un métabolite mesuré dans le cadre d'une approche non-ciblée est uniquement défini par son rapport masse sur charge (m/z) et son temps de rétention (RT)et n'est donc pas immédiatement identifiable. En outre, les m/z et RT mesurés dans des conditions différentes sont soumis à des variations potentiellement importantes. Par conséquent, les métabolites communs à deux études différentes ne peuvent pas être directement identifiés, ce qui complique la combinaison ou méta-analyse des données de ces études.

Nous développons ici une méthode non supervisée destinée à détecter et apparier les métabolites communs à deux études métabolomiques non ciblées en utilisant des informations sur les m/z, RT et intensités de signal. Notre approche repose sur la distance de Gromov-Wasserstein (GW), une extension du transport optimal conçue pour coupler des ensembles en tirant profit de leur structure. Dans notre approche, une contrainte additionnelle nous permet de restreindre ce couplage aux paires ayant des m/z proches. Finalement, la déviation des RT entre les deux études est estimée afin de ne retenir que les paires présentant des RT compatibles.

Nous évaluons les performances empiriques de notre méthode et la comparons à une autre approche récente qui utilise les mêmes informations (m/z, RT et intensités du signal). Les résultats obtenus sont prometteurs tant sur données simulées que sur données réelles. Notre méthode pourrait avoir de nombreuses applications en métabolomique, de la comparaison de protocoles d'acquisition à la combinaison ou méta-analyse de données issues de différentes études, permettant aux données métabolomiques non-ciblée d'être utilisées à leur plein potentiel, par exemple en épidémiologie.

Mots-clés. Données métabolomiques non-ciblées, appariement automatique, transport optimal, distance de Gromov-Wasserstein

Abstract.

Untargeted metabolomic profiling through LC-MS allows the measurement of a wide range of metabolites present in a biological sample. However, the number of samples measured in non-targeted metabolomics studies is generally small. One way to increase the statistical power of metabolomic studies would be to combine or meta-analyse data from several studies. However, a metabolite measured in an untargeted approach is only defined by its mass-to-charge ratio (m/z) and retention time (RT) and is therefore not immediately identifiable. Furthermore, m/z and RT measured under different conditions are subject to potentially large variations. Therefore, metabolites common to two different studies cannot be directly identified, making it difficult to combine or meta-analyse data from these studies.

Here we develop an unsupervised method to align features from two LC-MS untargeted datasets using their mass-to-charge ratios (m/z), retention times (RT) and signal intensities. Our approach is based on the Gromov-Wasserstein (GW) distance, an extension of optimal transport designed to couple sets by taking advantage of their structure. In our approach, an additional constraint allows us to restrict this coupling to pairs with close m/z. Finally, the deviation of the RTs between the two studies is estimated in order to retain only those pairs with compatible RTs.

We evaluate the empirical performance of our method, comparing it to another recent approach that uses the same information (m/z, RT and signal intensities). The results obtained are promising on both simulated and real data. If these good performances are confirmed, our method could have many applications in metabolomics, from the comparison of acquisition protocols to the combination or meta-analysis of data from different studies. This would allow non-targeted metabolomic data to be used to their full potential, for example in epidemiology.

Keywords. Untargeted metabolomics, Features alignment, Optimal transport, Gromov-Wasserstein distance

1. Introduction

La métabolomique est l'étude de l'ensemble des métabolites (composés organiques tels que des sucres, des acides gras ou encore des acides aminés) présents dans un tissu. Décrivant les niveaux de concentration de centaines voire milliers de métabolites, par exemple dans un échantillon sanguin, les données métabolomiques reflètent ainsi la santé métabolique d'un individu. Elles sont de plus en plus utilisées en épidémiologie, notamment du cancer, pour caractériser les mécanismes biologiques impliqués dans la carcinogénèse.

Les données métabolomiques sont dites ciblées lorsqu'elles concernent une liste préétablie de métabolites. Les données métabolomiques non-ciblées portent quant à elles sur un nombre beaucoup plus élevé de métabolites, virtuellement tous ceux présents dans un échantillon. Les données non ciblées permettent ainsi une description beaucoup plus fine de l'état de santé d'un individu. Leur analyse soulève cependant certains défis spécifiques. En particulier, un métabolite mesuré dans le cadre d'une approche non-ciblée n'est pas immédiatement identifiable. Plus précisément, il n'est caractérisé que par deux de ses propriétés chimiques : son rapport masse sur charge (m/z) et son temps de rétention (RT), respectivement mesurés par spectrométrie de masse et par chromatographie liquide. Ces mesures sont soumises à des sources de variabilité qui compliquent considérablement l'annotation des variables d'un jeu de données métabolomiques non ciblées.

Par ailleurs, le nombre d'échantillons mesurés dans les études métabolomiques non-ciblées est généralement faible, d'où l'intérêt de combiner ou méta-analyser les données issues de plusieurs études afin de gagner en pouvoir statistique. Or, un même métabolite mesuré dans deux études différentes (ou encore sur deux instruments différents, ou à deux instant différents) présentera des m/z et RT différents, ce qui complique l'identification des métabolites communs à plusieurs études, et donc la combinaison des données de plusieurs études. Des méthodes ont été développées pour répondre à ce problème, soit en permettant de déterminer automatiquement les caractéristiques

chimiques des métabolites[2], soit en cherchant à apparier automatiquement les métabolites communs à plusieurs études[4,5]. La plupart de ces méthodes requiert cependant des données supplémentaires qui ne sont pas toujours disponibles, telles que des mesures issues de spectrométrie de masse en tandem, ou nécessite une connaissance préalable de métabolites identifiés et communs aux études considérées pour guider l'appariement et calibrer les paramètres de la méthode.

Dans la Section 2, nous présentons une nouvelle procédure pour détecter et apparier automatiquement les métabolites communs à deux études non-ciblées en n'utilisant que leur m/z, RT et intensités de signal. Cette procédure repose sur une extension du transport optimal, la distance de Gromov-Wasserstein, déjà utilisée dans la méthode SCOT par exemple pour aligner des données multi-omiques unicellulaires[3]. Notre approche tire profit des similitudes attendues entre les structures de corrélation des intensités de signal dans deux jeux de données afin d'apparier les métabolites compatibles en termes de m/z et RT. Nous présentons en Section 3 une évaluation des performances empiriques de notre méthode sur données simulées et réelles, en la comparant à une autre approche récente qui utilise les même informations (m/z, RT et intensités de signal).

2. Appariement automatique

Notre approche repose sur l'idée que les intensités de signal de métabolites mesurés dans deux études différentes présenteront des structures de corrélation ou, de manière équivalente, des structures de distance, analogues. Cette similitude peut être utilisée afin de compléter l'information contenue dans les m/z et RT pour trouver et apparier les métabolites communs à deux études. Plus précisément, considérons $X^{(1)}$ et $X^{(2)}$ deux matrices de tailles respectives (n_1, p_1) et (n_2, p_2) où $X^{(i)}$ renferme les intensités de signal des p_i métabolites mesurés chez les n_i individus de l'étude i. Soient quatre métabolites $X_i^{(1)}, X_j^{(2)}, X_k^{(1)}$ et $X_l^{(2)}$. Si $\left(X_i^{(1)}, X_j^{(2)}\right)$ représentent les mesures d'un seul et même métabolite dans les études 1 et 2, et de même pour $\left(X_k^{(1)}, X_l^{(2)}\right)$, alors on s'attend à ce que

$$d_1\left(X_i^{(1)}, X_k^{(1)}\right) \approx d_2\left(X_j^{(2)}, X_l^{(2)}\right)$$

où, pour $i \in \{1,2\}$ et pour tout vecteurs $x, x' \in \mathbb{R}^{n_i}$, $d_i(x, x') = \frac{1}{\sqrt{n_i}} \left||x - x'|\right|^2$

Par ailleurs, les m/z et RT de ces deux variables, (m_i^1, RT_i^1) et $\left(m_j^2, RT_j^2\right)$ doivent être « compatibles » suivant des critères précisés en section 2.2.

2.1. Distance de Gromov-Wasserstein

Afin de tirer profit des similitudes attendues entre les structures de distance dans les deux jeux de données, nous utiliserons la distance de Gromov-Wasserstein (GW)[6], une extension du transport optimal au cas où les ensembles à coupler ne se situent pas dans le même espace métrique, ce qui est le cas dès que $n_1 \neq n_2$.

Plus précisément, pour p et q deux mesures de probabilité discrètes respectivement sur $X^{(1)} \subset \mathbb{R}^{n_1}$ et $X^{(2)} \subset \mathbb{R}^{n_2}$, la distance de GW entre $X^{(1)}$ et $X^{(2)}$ est donnée par :

$$GW\left(X^{(1)}, X^{(2)}\right) = \min_{\mathbb{P} \in \mathbb{U}(p,q)} \mathcal{E}_{X^{(1)}, X^{(2)}}(\mathbb{P}) \qquad (1)$$

où

$$\mathcal{E}_{X^{(1)}, X^{(2)}}(\mathbb{P}) = \sum_{i,j,k,l} \mathbb{P}_{i,j} \mathbb{P}_{k,l} \left| d_1\left(X_i^{(1)}, X_k^{(1)}\right) - d_2\left(X_j^{(2)}, X_l^{(2)}\right) \right|^2$$

et $\mathbb{U}(p, q)$ désigne l'ensemble des couplages de p et q, c'est-à-dire

$\mathbb{U}(p, q) = \left\{\mathbb{P} \in \mathbb{R}_+^{n_1 \times n_2} \middle| \mathbb{P}\mathbb{1}_{n_2} = p \text{ et } \mathbb{P}^T \mathbb{1}_{n_1} = q\right\}$

où $\mathbb{P}\mathbb{1}_{n_2} = \left(\sum_j \mathbb{P}_{i,j}\right)_i \in \mathbb{R}^{n_1}$ et $\mathbb{P}^T\mathbb{1}_{n_1} = \left(\sum_i \mathbb{P}_{i,j}\right)_j \in \mathbb{R}^{n_2}$.

Moralement, $\mathbb{P}_{i,j}$ reflète le degré de confiance avec lequel $X_i^{(1)}$peut être apparié à $X_j^{(2)}$.

Pour approximer la solution de (1), on utilisera l'algorithme de Sinkhorn pour résoudre la version avec régularisation entropique[7,9], et ainsi obtenir la matrice $\mathbb{P}$ solution de

$$\min_{\mathbb{P}\in\mathbb{U}(p,q)} \left\{\mathcal{E}_{X^{(1)},X^{(2)}}(\mathbb{P}) - \epsilon\mathbb{H}(\mathbb{P})\right\} \qquad (2)$$

où $\mathbb{H}(\mathbb{P}) = -\sum_{i,j} \mathbb{P}_{i,j}\log\left(\mathbb{P}_{i,j}\right)$ l'entropie de Shannon de $\mathbb{P}$, et pour des lois p et q discrètes uniformes sur $X^{(1)}$ et $X^{(2)}$ et ϵ à déterminer.

Par ailleurs, afin d'éliminer les métabolites non appariés, on utilisera la version déséquilibrée de la distance de GW décrite par Séjourné et al.[8] qui relaxe la contrainte sur les marginales.

2.2. Prise en compte des m/z et RT

Par ailleurs, dans notre contexte particulier, seules les métabolites $(X_i^{(1)}, X_j^{(2)})$ présentant des m/z et RT « compatibles » devraient être appariées. Le m/z d'un métabolite est une grandeur relativement stable, soumis à des erreurs de mesure de l'ordre d'1 ppm. On adjoindra donc à (2) la condition

$$\left|m_i^1 - m_j^2\right| > M_{\text{gap}} \Rightarrow \mathbb{P}_{i,j} = 0$$

avec M_{gap} choisi en fonction des appareils de spectrométrie de masse, fixé à 1ppm par défaut.

Les variations de RT d'une étude à l'autre sont plus complexes. La déviation systématique du RT peut quant à elle être non-linéaire et de grande amplitude (Figure 1), en particulier dans le cadre de protocoles de chromatographie liquide différents. Il est donc nécessaire d'estimer cette déviation afin de définir ce que seraient des RT « compatibles ». L'approche que nous avons adoptée s'inspire de méthodes de correction de la dérive des RT au sein d'un même lot d'échantillons issus d'une même expérience.

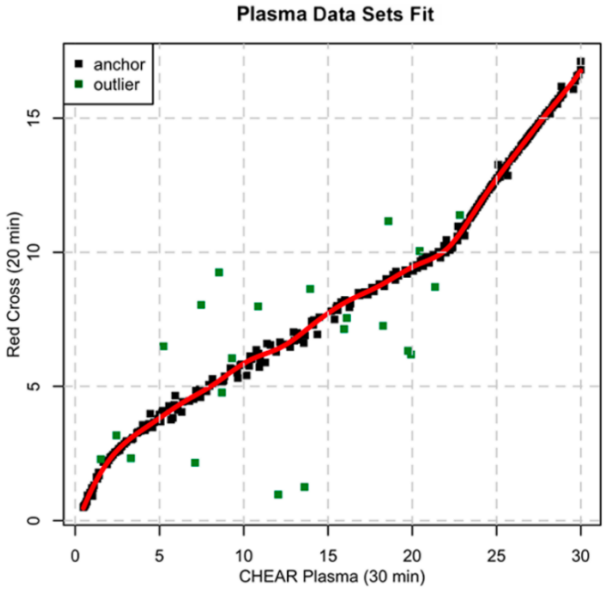


Figure 1 : Habra et al.[4] : Déviation de RT observée entre deux études utilisant des protocoles de chromatographie différents. Les axes représentent les RT mesurés (en mn, suivant deux échelles différentes) dans l'une et l'autre des études (Red Cross et CHEAR Plasma) ; les points marqués comme 'anchors' correspondent à des paires de métabolites communs aux deux études. Ceux marqués comme 'outliers' sont des exemples de paires aux m/z et intensités compatibles mais étant identifiées à tort.

On procèdera ainsi aux étapes suivantes :

1. Calcul de $\widehat{\mathbb{P}}$, couplage de GW contraint entre $X^{(1)}$ et $X^{(2)}$ réalisant
$$\underset{\mathbb{P}\in\mathbb{U}(p,q):\ \mathbb{P}_{i,j}=0 \text{ si } \left|m_i^1-m_j^2\right|>M_{\text{gap}}}{\operatorname{argmin}} \left\{\mathcal{E}_{X^{(1)},X^{(2)}}(\mathbb{P}) - \epsilon\mathbb{H}(\mathbb{P})\right\}$$

Chaque paire $\left(X_i^{(1)}, X_j^{(2)}\right)$ est associée à un coefficient $\widehat{\mathbb{P}}_{i,j}$ traduisant le degré de confiance accordée à la paire. Une paire est identifiée par le couplage de GW contraint si le coefficient associé est non nul, et on note $\mathcal{S}$ l'ensemble des paires identifiées : $\mathcal{S} = \{i, j \mid \widehat{\mathbb{P}}_{i,j} \neq 0\}$.

2. Estimation de la déviation des RT en utilisant des B-splines pondérées. Plus précisément, on cherche à identifier f B-spline pondérée minimisant la quantité.

$$\sum_{i,j \in \mathcal{S}} \widehat{\mathbb{P}}_{i,j}\left(RT_i^1 - f\left(RT_j^2\right)\right)$$

Ces paires sont pondérées par les coefficients $\widehat{\mathbb{P}}_{i,j}$ correspondant de manière à accorder plus d'importance aux paires identifiées avec confiance.

3. Détection des valeurs aberrantes et élimination a posteriori des paires identifiées à tort. Pour chaque paire identifiée $\left(X_i^{(1)}, X_j^{(2)}\right)$, on élimine cette paire de $\mathcal{S}$ si $\left(RT_i^{(1)}, RT_j^{(2)}\right)$ n'est pas compatible avec la déviation estimée.
On forme ainsi la matrice C telle que

$$C_{i,j} = \begin{cases} \widehat{\mathbb{P}}_{i,j} \text{ si } \left(RT_i^{(1)}, RT_j^{(2)}\right) \text{ est compatible avec } \hat{f} \\ 0 \quad \text{sinon} \end{cases}$$

La méthode s'arrête après cette étape et renvoie la matrice C. Afin de garantir que chaque métabolite soit apparié au maximum une fois, on ne conservera pour chaque métabolite que la paire identifiée avec le degré de confiance le plus élevé, ce qui revient à définir l'ensemble $\mathbb{S}$ des paires retenues par

$$\mathbb{S} = \left\{i, j \mid i = \underset{k}{\operatorname{argmax}}\, C_{k,j}\,, j = \underset{k}{\operatorname{argmax}}\, C_{i,k}\right\}.$$

3. Simulations

Pour évaluer les performances de notre approche, nous avons utilisé un jeu de données métabolomiques non-ciblées réelles issues de Plusquin et al. (2020)[1] comportant $n = 499$ individus et $p = 4713$ métabolites identifiés par leurs m/z et RT pour mimer le cas où l'on aurait des données provenant de deux études différentes. Les métabolites et individus du jeu de données d'origine ont été répartis dans deux sous-échantillons suivant le processus décrit en Figure 2.

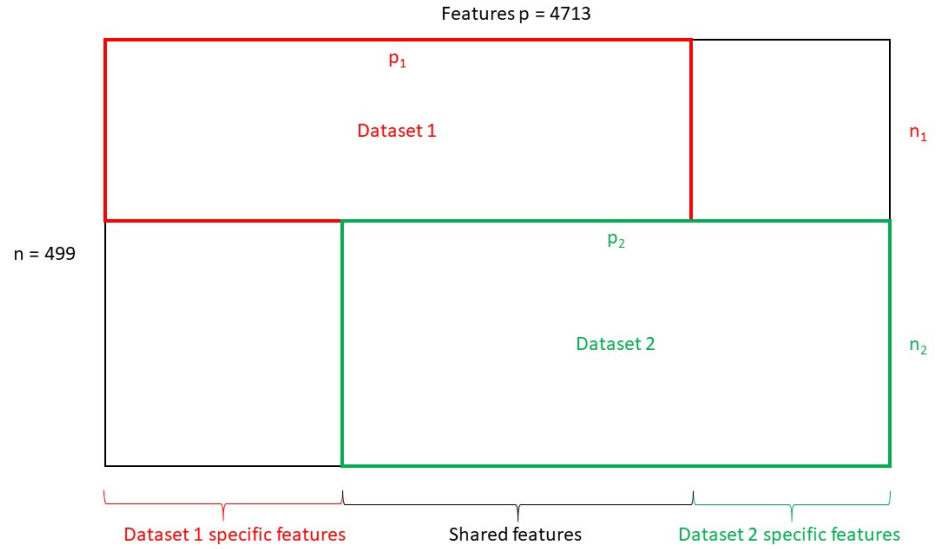


Figure 2 : Simulation des jeux de données à apparier. Le jeu de données initial est divisé en deux sous échantillons de tailles respectives $n_1 \times p_1$ et $n_2 \times p_2$, partageant des métabolites communs ('Shared features'). Le but de notre approche est de retrouver et apparier ces métabolites. Chaque sous-échantillon comporte également des métabolites qui lui sont spécifiques ('Dataset 1 specific features' et 'Dataset 2 specific features') qui ne doivent pas être appariés. Pour n_i fixé,

les individus sont assignés aléatoirement à un des deux sous-échantillon. De même, les *p* métabolites sont aléatoirement assignés au sous-échantillon n°1, au sous-échantillon n°2 ou aux deux. Une fois les deux sous-échantillons générés, les m/z et RT des métabolites qui leur ont été assignés sont bruités et déviés pour mimer des conditions d'acquisition différentes.

Nous avons comparé notre approche avec celle proposée par Habra et al. (2021)[4] qui utilise également les m/z, RT, et les intensités de signal. Pour ces deux méthodes, nous avons évalué d'une part leur *(i)* sensibilité, mesurant leur capacité à retrouver les métabolites communs aux deux jeux de données, et *(ii)* leur spécificité, mesurant leur capacité à ne pas apparier les métabolites spécifiques à l'un des jeux de données. Plusieurs configurations ont été considérées, en variant le nombre de métabolites communs, l'intensité du bruit portant sur les m/z et RT, la forme de la déviation systématique de RT, ainsi que le nombre d'individus dans chaque sous-échantillon. Pour chaque configuration, 20 paires de sous-échantillons ont été générées. Les résultats obtenus sont généralement à l'avantage de notre méthode avec une sensibilité et une spécificité respectivement comprises entre 0.95 - 0.99 et 0.91 - 0.97 dans la plupart des configurations explorées, contre 0.94 – 0.95 et 0.90 – 0.92 pour l'approche de Habra et al. (2021)[4].

Par ailleurs, les deux approches ont été comparées sur données réelles avec des résultats nettement en faveur de notre méthode.

4. Conclusion

Notre approche a montré de meilleures performances que les méthodes comparables sur des données simulées et réelles. Ces résultats doivent cependant être examinés plus en détail sur des données réelles présentant des déviations de RT potentiellement plus importantes. S'ils se confirment, notre approche pourrait avoir de nombreuses applications telles que la comparaison des protocoles expérimentaux, la collaboration entre institutions pour l'annotation de composés chimiques, ainsi que la combinaison ou la méta-analyse de données issues de différentes études. Elle contribuerait ainsi à une meilleure exploitation de l'ensemble des données métabolomiques non ciblées générées dans diverses études à ce jour, en particulier en épidémiologie.

Bibliographie

[1] Alfano R, Chadeau-Hyam M, Ghantous *et al.* (2020) A multi-omic analysis of birthweight in newborn cord blood reveals new underlying mechanisms related to cholesterol metabolism *Metabolism*, 110, 154292, ISSN 0026-0495

[2] Chen L., Lu W., Wang L. *et al.* (2021) Metabolite discovery through global annotation of untargeted metabolomics data. *Nat Methods* 18, pp.1377–1385

[3] Demetci P, Santorella R, Sandstede B, Noble WS, Singh R. (2022) SCOT: Single-Cell Multi-Omics Alignment with Optimal Transport. *J Comput Biol*. 3-18.

[4] Habra H., Kachman M., Bullock K., *et al*. (2021) metabCombiner: Paired Untargeted LC-HRMS Metabolomics Feature Matching and Concatenation of Disparately Acquired Data Sets *Analytical Chemistry* 93(12), pp.5028-5036

[5] Hsu YHH, Churchhouse C, Pers TH, *et al*. (2019) PAIRUP-MS: Pathway analysis and imputation to relate unknowns in profiles from mass spectrometry-based metabolite data. *PLOS Computational Biology* 15(1): e1006734.

[6] Mémoli, F. (2011) Gromov–Wasserstein Distances and the Metric Approach to Object Matching. *Found Comput Math* 11,17–487

[7] Peyré G. et Cuturi M. (2019) Computational Optimal Transport *Foundation and Trends in Machine Learning,* 5-6 pp.305-607

[8] Séjourné T, Vialard F-X, Peyré G (2020) The Unbalanced Gromov Wasserstein Distance: Conic Formulation and Relaxation. DOI: 10.48550/arXiv.2009.04266

[9] Solomon J, Peyré G, Kim V, Sra S. (2016) Entropic Metric Alignment for Correspondence Problems. *ACM Transactions on Graphics, Association for Computing Machinery Proc. SIGGRAPH 2016*, 35 (4), pp.72:1-72:13.

GAN Estimation of Lipschitz Optimal Transport

Lucas De Lara [1] & Alberto Gonzaléz-Sanz [2] & Louis Béthune [3]

Université Paul-Sabatier

[1] *lucas.de_lara@math.univ-toulouse.fr*
[2] *alberto.gonzalez_sanz@math.univ-toulouse.fr*
[3] *louis.bethune@univ-toulouse.fr*

Résumé. Cet article introduit le premier estimateur de la fonction de transport optimal entre deux distribitions de probabilité, basé sur des réseaux de neurones. En mettant à profit de récentes avancées théorique et pratiques dans le domaines des réseaux de neurones Lipschitz, nous définissons un réseau génératif adversarial sous contraintes Lipschitz, pénalisé par le coût de transport quadratique. Puis, nous demontrons sous des hypothèses de régularité que le générateur ainsi obtenu converge uniformément vers la fonction de transport lorsque la taille des échantillons tend vers l'infini. De plus, nous illustrons les performances de la fonction apprises sur des exemples numériques. En comparaison de travaux antérieurs se focalisant soit sur les garanties statististiques soit sur les performances pratiques, nous fournissons un estimateur implémentable et expressif qui ouvre la voie pour des applications du transport optimal où le comportement statistique doit être certifié.

Mots-clés. Transport optimal, réseaux adversariaux génératifs

Abstract. This paper introduces the first statistically consistent estimator of the optimal transport map between two probability distributions, based on neural networks. Building on theoretical and practical advances in the field of Lipschitz neural networks, we define a Lipschitz-constrained generative adversarial network penalized by the quadratic transportation cost. Then, we demonstrate that, under regularity assumptions, the obtained generator converges uniformly to the optimal transport map as the sample size increases to infinity. Furthermore, we show through a number of numerical experiments that the learnt mapping has promising performances. In contrast to previous work tackling either statistical guarantees or practicality, we provide an expressive and feasible estimator which paves way for optimal transport applications where the asymptotic behaviour must be certified.

Keywords. Optimal transport, Generative adversarial networks

1 Introduction

An *optimal transport map* transforms one distribution into another with minimal effort. Formally, given two probability distributions P and Q on $\Omega \subseteq \mathbb{R}^d$, an optimal transport

map from P to Q is a solution to,

$$\min_{T \in \mathcal{T}(P,Q)} \int_{\Omega} \|x - T(x)\|^2 \mathrm{d}P(x), \tag{1}$$

where $\mathcal{T}(P, Q)$ is the set of measurable maps $T : \Omega \to \Omega$ *pushing forward* P to Q, that is $Q(M) = P(T^{-1}(M))$ for every measurable set $M \subseteq \Omega$. This property, denoted by $T_\sharp P = Q$, means that if a random variable X follows the distribution P then its image $T(X)$ follows the distribution Q. According to Theorem 2.12 in (Villani, 2003), when P and Q admit densities with respect to the Lebesgue measure and have finite second-order moments, then there exists a unique (up to P-negligible sets) solution to Problem (1), which we denote by T_0.

In practical frameworks, one typically does not have access to the true distributions P and Q but to independent samples $x_1, \ldots, x_n \sim P$ and $y_1, \ldots, y_n \sim Q$. This raises the question of constructing a tractable approximation of the solution T_0 on the basis of these empirical observations. The simplest way to compute an empirical optimal transport map from data points is to solve Problem (1) between the empirical measures $P_n := n^{-1} \sum_{i=1}^n \delta_{x_i}$ and $Q_n := n^{-1} \sum_{i=1}^n \delta_{y_i}$ instead of P and Q. Implementing this solution suffers from three main drawbacks. The first one is the *computational cost*, since it requires at least $O(n^3)$ operations to compute the empirical optimal transport map. The second is the *memory cost*, since this map is typically stored as an $n \times n$ matrix. As a consequence of these two issues, this approach does not scale well with the size of the dataset. The third limitation of the empirical map is its *inability to generalize* to new out-of-sample observations: by construction it is only matching the set $\{x_1, \ldots, x_n\}$ to $\{y_1, \ldots, y_n\}$.

These practical drawbacks triggered a vast literature on continuous approximations of optimal transport maps. The proposed mappings all come with different practical limitations, theoretical guarantees, and experimental performances. However, state-of-the-art has mostly addressed either theoretically grounded statistical estimators of optimal transport maps, but unsuitable for large-scale implementations, or efficient heuristic approximations, at the cost of statistical guarantees. In this paper, we propose a novel GAN-based estimator G_n of T_0 which, under some assumptions, converges uniformly:

$$\|G_n - T_0\|_\infty \xrightarrow[n \to +\infty]{a.s.} 0.$$

Our construction takes root in the approximation from (Black et al., 2020), defined as the generator of a penalized Wassertein-GAN (WGAN) training problem, and improve it by assuming a setting where the optimal transport map is Lipschitz and by leveraging recent theoretical and practical advances on Lipschitz neural networks (Tanielian & Biau, 2021). Formally, G_n solves the following adversarial training:

$$\inf_{G \in \mathcal{G}_n} \left\{ \|I - G\|^2_{L^2(P_n)} + \lambda_n \sup_{D \in \mathcal{D}_n} \int D \left(\mathrm{d}(G_\sharp P_n) - \mathrm{d}Q_n \right) \right\},$$

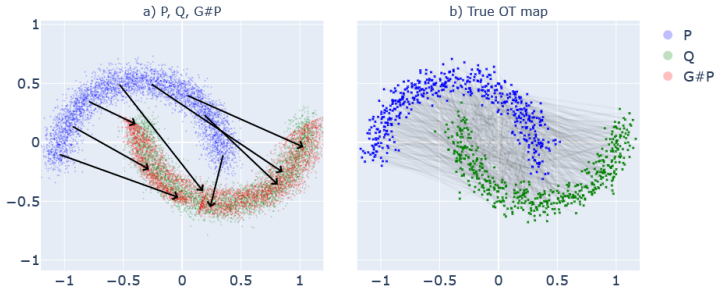


Figure 1: Estimation of the optimal transport map on the TwoMoons dataset. (a) GAN estimator G after 800 gradient steps on the generator, on the basis of $4,000$ points from each distribution. The black arrows represent the transport of specific points. (b) Empirical optimal transport map (discrete matching) between samples of size 500.

where $\mathcal{D}_n$ is a class of 1-Lipschitz discriminators providing a proxy for the Wasserstein-1 distance, and $\mathcal{G}_n$ is a class of Lipschitz generators parametrizing the space of feasible mappings. The positive parameter λ_n governs the trade-off between minimizing the quadratic transportation cost, promoting the objective of the Monge problem (1), and minimizing the distance between the generated and the target distributions, enforcing the push-forward constraint. The method we propose has the advantage of providing a theoretically sound and feasible estimation of the optimal transport map whose statistical convergence can be mathematically certified.

Notations. The absolute value of real numbers and the Euclidean norm of vectors are respectively given by $|\cdot|$ and $\|\cdot\|$. The notation B_r refers to the centered Euclidean ball of $\mathbb{R}^d$ with radius $r > 0$. If Ω is a closed convex set, then $\mathcal{P}_\Omega$ stands for the projection onto Ω. In the following $\Omega_1 \subseteq \mathbb{R}^{d_1}$ and $\Omega_2 \subseteq \mathbb{R}^{d_2}$ denote two arbitrary subsets. For a function $F : \Omega_1 \to \Omega_2$ and μ a probability measure on Ω_1, we write $\|F\|_{L^2(\mu)} := \sqrt{\int_{\Omega_1} \|F(x)\|^2 \mathrm{d}\mu(x)}$. The supremum norm of function is given by $\|\cdot\|_\infty$.For some $L > 0$, we write $\mathrm{Lip}_L(\Omega_1, \Omega_2)$ the set of L-Lipschitz functions from Ω_1 to Ω_2. For some $\alpha > 0$, we call $\mathcal{C}^\alpha(\Omega_1, \Omega_2)$ the set of α-Hölder functions from Ω_1 to Ω_2 and write $\|\cdot\|_{\alpha,\infty}$ for the α-Hölder norm of functions. For a differentiable function $F : \Omega_1 \to \Omega_2$, we call F' its derivative, where for any $x \in \Omega_1$ the quantity $F'(x)$ is a $d_1 \times d_2$ matrix. For a real symmetric matrix S and a real number γ, the relation $\gamma \preceq S$ indicates that all the eigenvalues of S are greater than γ. The relation $\succeq$ is defined similarly.

2 Lipschitz neural networks

In this paper, we focus on the *GroupSort* activation function to impose the Lipschitz constraint on the discriminator and generator. In contrast to former work, we consider general setting where the output dimension $p \geq 1$ is arbitrary.

We write σ_2 for the GroupSort activation function of grouping size 2. By definition, it splits the pre-activation input into groups of size 2, and then sorts each group by decreasing order. We call a GroupSort feed-forward neural network (with grouping size 2) any function $N_\theta : \mathbb{R}^d \to \mathbb{R}^p$ of the form

$$N_\theta = h_l \circ h_{l-1} \circ \ldots \circ h_1, \tag{2}$$

where

$$\begin{aligned}
&h_1(x) := W_1 x + b_1 \text{ with } W_1 \in \mathbb{R}^{w_1 \times d}, b_1 \in \mathbb{R}^{w_1};\\
&h_2(x) := W_2 \sigma_2(x) + b_2 \text{ with } W_2 \in \mathbb{R}^{w_2 \times w_1}, b_2 \in \mathbb{R}^{w_2};\\
&\ldots\\
&h_l(x) := W_l \sigma_2(x) + b_l \text{ with } W_l \in \mathbb{R}^{p \times w_{l-1}}, b_l \in \mathbb{R}^{p}.
\end{aligned}$$

The integer $l \geq 1$ denotes the *depth* of the network while the integers $\{w_1, \ldots, w_{l-1}\}$ refer to the *widths* of the hidden layers $\{h_1, \ldots, h_{l-1}\}$. The widths are assumed to be divisible by 2 (the grouping size). Additionally, we define $s := \sum_{i=1}^{l-1} w_i$ the *size* of the network. The parameter $\theta := (W_1, \ldots, W_l, b_1, \ldots, b_l) \in \Theta$ represents the *weights* matrices and *offset* vectors of N_θ.

For a matrix W, let $\|W\|_\infty := \sup_{\|x\|_\infty = 1} \|Wx\|_\infty$ and $\|W\|_{2,\infty} := \sup_{\|x\|=1} \|Wx\|_\infty$, where $\|x\|_\infty$ denotes the maximum norm of vectors. Consider the following assumption on the parameters:

(C) There exists a constant $C > 0$ such that for all $(W_1, \ldots, W_l, b_1, \ldots, b_l) \in \Theta$,

$$\|W_1\|_{2,\infty} \leq 1 \text{ and } \max(\|W_2\|_\infty, \ldots, \|W_l\|_\infty) \leq 1 \text{ and } \max(\|b_1\|_\infty, \ldots, \|b_l\|_\infty) \leq C.$$

In the following, we write $\mathcal{N}_C^p(l, s)$ for the class of GroupSort feed-forward neural networks with depth l, size s, output dimension p, satisfying Assumption **(C)** for the constant $C > 0$ on a compact set Ω. A trivial extension of Lemma 1 in (Tanielian & Biau, 2021) to the multivariate case ensures that GroupSort neural networks satisfying Assumption **(C)** are 1-Lipschitz. Moreover, by generalizing Theorem 2 in (Tanielian & Biau, 2021) we can show that for well-chosen bound C, depth l, and size s, the set $\mathcal{N}_C^p$ can approximate any class of bounded 1-Lipschitz functions up to a given precision. Omitted details and results can be found in the full-length paper (González-Sanz et al., 2022).

3 GAN estimator

3.1 Optimal transport setup

Set P and Q two measures on $\mathbb{R}^d$ admitting densities with respect to the Lebesgue measure and with finite second-order moments. We aim at estimating with a GroupSort neural network the unique optimal transport map T_0 between P and Q through the knowledge of the empirical distributions P_n and Q_n. As mentioned in the introduction, we consider a setting where the optimal transport map T_0 is Lipschitz.

We denote by $\Omega_P := \text{supp}(P)$ the *source* domain and $\Omega_Q := \text{supp}(Q)$ the *target* domain. Then, we let $L \geq 2$ and make the following assumptions:

(S1) The source domain $\Omega_P \subseteq B_L$ is a bounded and connected Lipschitz domain. The measure P admits a density ρ with respect to the Lebesgue measure such that $L^{-1} \leq \rho(x) \leq L$ for almost every $x \in \Omega_P$.

(S2) Let $\tilde{\Omega}_P$ denote a convex set with Lipschitz boundary such that $\Omega_P + B_{L^{-1}} \subseteq \tilde{\Omega}_P \subseteq B_L$. The optimal transport map T_0 is a differentiable function from $\tilde{\Omega}_P$ to $\mathbb{R}^d$ such that $T_0 = \nabla f_0$ where $f_0 : \tilde{\Omega}_P \to \mathbb{R}^d$ is a differentiable convex function. Additionally it satisfies:

$$T_0 \in C^2(\tilde{\Omega}_P, \mathbb{R}^d) \text{ s.t. } \|T_0\|_{2,\infty} \leq L, \text{ and } L^{-1} \preceq T_0'(x) \preceq L \text{ for all } x \in \tilde{\Omega}_P.$$

These are the same hypothesis as in Section 5 from (Hütter & Rigollet, 2021), specified with a Hölder regularity α equals to 2. This ensures the existence of a near-optimal minimax estimator of T_0, which play a key role in the proof of our estimator's consistency.

3.2 GAN setup

In the WGAN framework, the *discriminator* $D : \mathbb{R}^d \to \mathbb{R}$ is a neural network defining a proxy for the Wasserstein-1 distance, while the *generator* $G : \mathbb{R}^d \to \mathbb{R}^d$ is a neural network minimizing this proxy between $G_\sharp P_n$ and Q_n, thereby aiming at generating Q from P. Importantly, the Wasserstein-1 distance has the following dual formulation:

$$\mathcal{W}(\mu, \nu) = \sup_{f \in \text{Lip}_1(\Omega, \mathbb{R})} \int f(\mathrm{d}\mu - \mathrm{d}\nu). \tag{3}$$

The key idea of WGAN-like methods is to approximate this distance by computing the supremum over a class of neural networks included in $\text{Lip}_1(\Omega, \mathbb{R})$. Formally, we approximate the Wasserstein-1 distance through the following integral probability metric:

$$\mathcal{W}_n(\mu, \nu) := \sup_{D \in \mathcal{D}_n} \int D(\mathrm{d}\mu - \mathrm{d}\nu). \tag{4}$$

In our case, we set $\mathcal{D}_n := \mathcal{N}^1_{C_n}(l_n, s_n)$ with l_n and s_n increasing with n. The exact definition is specified in the full-length paper. We now turn to the generator. Set $\lambda_n > 0$ a regularization weight, and define the GAN estimator G_n as an optimal solution to Problem (2), that is

$$G_n \in \operatorname*{arg\,min}_{G \in \mathcal{G}_n} \mathcal{L}_n(G),$$

where

$$\mathcal{L}_n(G) := \|I - G\|^2_{L^2(P_n)} + \lambda_n \mathcal{W}_n(G_\sharp P_n, Q_n).$$

As for the discriminators, we do not enter into details for the definition of the generators. Simply note that

$$\mathcal{G}_n := \{x \in \Omega \mapsto \mathcal{P}_{B_L}(L \times N(x)), N \in \mathcal{N}^d_{C_n}(l_n, s_n)\}$$

where l_n and s_n increase with n. We are now ready to state our main theorem.

Theorem 3.1. *Let P and Q be such that the smoothness assumptions **(S1)** and **(S2)** on the optimal transport problem hold, and denote by T_0 the (almost everywhere) unique optimal transport map between P and Q. If λ_n tends sufficiently slowly to $+\infty$, then for G_n defined as a solution to Problem (2) we have*

$$\|G_n - T_0\|_\infty \xrightarrow[n \to +\infty]{a.s.} 0.$$

Critically, the speed at which λ_n must convergence depends on the dimension d of the input space. To the best of our knowledge, this is the first statistical consistency result for a neural-network-based optimal transport map. We leave the analysis of consistency rates for future work.

References

Black, E., Yeom, S., and Fredrikson, M. Fliptest: fairness testing via optimal transport. In *Conference on Fairness, Accountability, and Transparency*, pp. 111–121, 2020.

González-Sanz, A., de Lara, L., Béthune, L., and Loubes, J.-M. GAN estimation of Lipschitz optimal transport maps, 2022.

Hütter, J.-C. and Rigollet, P. Minimax estimation of smooth optimal transport maps. *The Annals of Statistics*, 49(2):1166–1194, 2021.

Tanielian, U. and Biau, G. Approximating Lipschitz continuous functions with groupsort neural networks. In *International Conference on Artificial Intelligence and Statistics*, pp. 442–450. PMLR, 2021.

Villani, C. *Topics in optimal transportation*. Number 58 in Graduate Studies in Mathematics. American Mathematical Soc., 2003.

JDCOT : an Algorithm for Transfer Learning in Incomparable Domains using Optimal Transport

Marion Jeamart [1] & Renan Bernard [1] & Nicolas Courty [1] & Chloé Friguet [1]
& Valérie Garès[2]

[1] *Univ. Bretagne-Sud, UMR 6074, IRISA, Vannes, France - prenom.nom@univ-ubs.fr*
[2] *INSA, UMR 6625, IRMAR, Rennes, France - valerie.gares@insa-rennes.fr*

Résumé. L'adaptation de domaine est un champ de l'apprentissage par transfert où les données d'entrainement (source) du modèle et celles utilisées pour le test (cible) proviennent de deux domaines dont les distributions sous-jacentes sont différentes, et il convient d'adapter le modèle appris pour qu'il puisse être utilisé sur les données cibles avec de bonnes performances. On présente ici un algorithme permettant de traiter l'adaptation de domaine dans le cas d'espaces source et cible de hétérogènes car représentés par des espaces de caractéristiques différents. La méthode développée utilise le transport optimal pour coupler les distributions des deux domaines et son implémentation est illustrée sur des données de référence.

Mots-clés. transport optimal, apprentissage par tranfert, adaptation de domaine hétérogène

Abstract. Domain adaptation is a field of transfer learning where the training data (source) and the test data (target) come from different domains. The data in these two domains have therefore different underlying distributions, and the learned model should be adapted so that it can be used on the target data with good performance. We present here a domain adaptation algorithm to adapt heterogeneous domains, *i.e.* described by different features. The developed method uses optimal transport to map the distributions of the two domains and its implementation is illustrated on benchmark data.

Keywords. optimal transport, transfer learning, heterogeneous domain adaptation

1 Introduction

Transfer learning aims at leveraging knowledge or models learnt on a specific task to different, but somehow related, learning tasks. In the specific case where the learning tasks (classification, regression) are the same, but the distribution of the learning and test data are different, domain adaptation (DA) is considered to slightly update, with no human supervision, a model trained on a source domain (S) so that it becomes more robust when being used on data of interest lying in another domain, also named target domain (T), describing the same phenomenon but with another point of view.

In general, most of the existing domain adaptation methods assume that data of both source and target domains are represented by the same features space, with identical dimensions. But in some applications, it does not hold so that development of novel methods must consider domain adaptation across heterogeneous feature spaces, which is referred to as heterogeneous domain adaptation (HDA).

Transfer learning in general is useful in a context where the availability of data is limited. Using knowledge acquired in a setting into an other setting helps to solve issues when data lack. Different cases for supervision are investigated in the literature, and main methods can be classified regarding two strategies: (1) project both data into a common subspace by jointly learning the common subspace and a classifier, then iteratively align the discriminative dimensions ([7], [1]), and (2) jointly perform implicit data reconstruction and learn a classifier ([4]).

In this paper, we focus on three kinds of HDA (Table 1): (1) unsupevised HDA when all the source labels are observed, but none in target, (2) semi-supervised HDA when all the source labels and some labels in target are observed, (3) partial HDA when source and target labels are both partially observed.

	Y^S	Y^T
Unsupervised DA	observed	unobserved
Semi-supervised DA	observed	partially observed
Partial DA	partially observed	partially observed

Table 1: Several learning cases depending on the labels availability in S and/or T. *Note: in the partial DA setting, since labels in both domains are partially observed, source and target domains can be interchanged.*

Recently, optimal transport (OT) has been investigated as an efficient tool to deal with DA issues in a unsupervised or semi-supervised context. In the following, we first present the main principles of OT for DA in section 2. Then, we propose to extend the use of OT for DA to the case of heterogeneous transfer learning in section 3. Several numerical experiments are conducted to assess the proposed algorithm. Finally, some perspectives are discussed in section 4

2 Optimal transport for domain adaptation

Notations In the following, let $(X^S, Y^S) \in \mathbb{R}^{n^S \times d^S} \times \mathcal{C}$ be the source data and $(X^T, Y^T) \in \mathbb{R}^{n^T \times d^T} \times \mathcal{C}$ the target data. Moreover, $P(x^S, y^S) \in \mathcal{P}(\Omega^S, \mathcal{C})$ and $P(x^T, y^T) \in \mathcal{P}(\Omega^T, \mathcal{C})$ denote the joint probability distribution of the data in the source and target domains, respectively.

Optimal Transport (OT) Firstly introduced by G. Monge in 1781 and recently more widely studied (see [5] for details), OT is an optimisation problem that allows to define a distance between two probability measures (discrete, semi-discrete or continuous) called the Wasserstein distance. This distance is actually used as a loss in many optimisation problems and approximation algorithms allow to scale this problem to large dimension settings. We consider two sets of weighted samples $X^S = \{(x_i^S, w_i^S), i = 1 \dots n^S, \sum_{i=1}^{n^S} w_i^S = 1\}$ and $X^T = \{(x_j^T, w_j^T), j = 1 \dots n^T, \sum_{j=1}^{n^T} w_j^T = 1\}$.

The Kantorovich formulation consists in finding a coupling matrix $P \in \mathbb{R}_+^{n^S \times n^T}$ that satisfies $\gamma = \underset{P \in U(w^S, w^T)}{\operatorname{argmin}} \sum_{i,j} C_{ij} P_{ij}$ where $U(w^S, w^T)$ is the set of matrices $P \in \mathbb{R}_+^{n^S \times n^T}$ which verifies $\sum_{i=1}^{n^S} P_{ij} = w_j^T$, $\forall j = 1 \dots n^T$ and $\sum_{j=1}^{n^T} P_{ij} = w_j^T$, $\forall i = 1 \dots n^S$, and C is a cost matrix. In the following, we will note : $\gamma = OT(w^S, w^T, C)$.

Optimal Transport for (unsupervised) Domain Adaptation (OTDA)[2]. Assuming that $P(y^T|X^T) = P(y^S|M(X^S))$ and $P(X^T) = P(M(X^S))$ with $M : \Omega^S \mapsto \Omega^T$ a nonlinear transformation of the input space, the transport map γ is obtained by solving the OT problem between X^S and X^T : $\gamma = OT(w^S, w^T, C)$ where $w^S = \mathbf{1}_{n^S}/n^S$, $w^T = \mathbf{1}_{n^T}/n^T$ and $C \in \mathbb{R}^{n^S \times n^T}$ such that $C_{ij} = d(x_i^S, x_j^T)$, with d a dissimilarity measure and $\mathbf{1}$ the indicator function.

$\gamma \in \mathbb{R}^{n^S \times n^T}$ can be considered as the empirical joint distribution of $P(X^S, X^T)$. The transport map γ is then used to calculate the barycentric coordinates of the source samples in the target domain: $M(x_i^S) = \frac{1}{w_i^S} \sum_j \gamma_{ij} x_j^T$. Finally, a classifier f is trained on the transported source data $(M(X^S), Y^S)$ and applied on the target data X^T to predict target labels : $Y_{pred}^T = f(X^T)$.

Joint Distribution Optimal Transport (JDOT) [3]. Developed to deal with changes in the marginal distributions of the features and in the conditional distributions of Y that can occur with real-world data but that are not handled by OTDA, JDOT consists in simultaneously optimising the coupling matrix γ and the classifier f by minimising:

$$\min_{\gamma, f} \sum_{i,j} \Big[\alpha \; d(x_i^S, x_j^T) + \mathcal{L}(y_i^S, f(x_j^T))\Big]\gamma_{ij}$$

where α is a hyper-parameter and $\mathcal{L}$ is a loss function. A Block Coordinate Descent (BCD) is performed to alternatively estimate γ and f in practice. The authors have shown the superiority of their approach through experiments on benchmark datasets *w.r.t.* several domain adaptation state-of-the-art methods, including previous OT-based approaches, domain adversarial neural networks or transfer components.

Nevertheless, this approach does not address the heterogeneous domain adaptation problem. In the following, JDCOT is introduced to deal with this context.

3 JDCOT: Optimal transport for heterogeneous transfer learning

The aim of the proposed method called Joint Distribution Co-Optimal Transport (JDCOT) is to deal with heterogeneous domain adaptation with different dimension data. The way OT maps two domains makes it impractical if they are on different spaces. Co-Optimal Transport (COOT) [6] has therefore been developed to deal with incomparable spaces, by simultaneously optimising two transport maps between both samples and features with a BCD. We propose to use the principle of COOT to adapt JDOT to the HDA framework.

Method The JDCOT method consists in simultaneously solving the OT problem on the samples (γ^s), the OT problem on the variables (γ^v) and the classification problem (f^T), by minimising with a BCD, with $i \in [1; n^S], j \in [1; n^T], k \in [1; d^S], \ell \in [1; d^T]$:

$$\min_{\substack{\gamma^s, \gamma^v, f^S, f^T \\ \gamma^s \in U(w^S, w^T) \\ \gamma^v \in U(v^S, v^T)}} \sum_{i,j,k,\ell} \left[\alpha \; d(X^S_{i,k}, X^T_{j,\ell}) + \mathcal{L}\Big(\tilde{f}^S(x^S_i), \tilde{f}^T(x^T_j) \Big) \right] \gamma^s_{ij} \gamma^v_{k\ell} \tag{1}$$

where v^S and v^T are the weights of the points $(X')^S$ and $(X')^T$ respectively, α is the JDOT hyper-parameter, and $\tilde{f}^S$ and $\tilde{f}^T$ are classifiers, respectively on source and target domains, such that $\tilde{f}(x) = y$ if y is observed and $\tilde{f}(x) = f(x)$ otherwise.

Experiments: Settings and datasets We have conducted extensive numerical experiments to illustrate the proposed method considering the three settings reported in Table 1 (unsupervised DA, semi-supervised DA and partial DA), using the MNIST and USPS datasets. These datasets contain gray level images of handwritten digits ($K = 10$ classes) sized 16×16 for USPS (source) and 28×28 for MNIST (target), so that the two domains do not share the same size, as expected in the heterogeneous setting. The experimental protocol consists in randomly selecting 300 images per class ($n^S = n^T = 300 \times 10$) or 30 images per class ($n^S = n^T = 30 \times 10$) in each dataset for the training set. n^S_* (resp. n^T_*) images are labelled in the source (resp. target) domain. The test set is made up of 200 target images per class and is fixed such as it's the same for all the experiments. A convolutional neural network composed of 2 convolutional and 2 dense layers (CNN) is considered as classifiers.

Unsupervised and semi-supervised HDA At first, JDCOT and COOT are compared in the unsupervised ($n^S_* = n^S$ and $n^T_* = 0$) and semi-supervised ($n^S_* = n^S$ and $\forall k \in [1; K], n^T_{k,*} \in \{1; 3; 10\}$) cases. In these settings, all source labels are known such that $\forall x^S_i \in X^S, \tilde{f}^S(x^S_i) = y^S_i$. In the semi-supervised case, f^T is initialised with the n^T_*

available target labels. Its performance on the test set is reported as a baseline score in Table 2. Table 2 also reports the correct classification rate on the test set (test accuracy) over 10 random samplings for the training sets. The test accuracy of COOT is computed with the CNNt model trained with the estimated target labels given by label propagation through γ^s. The test accuracy of JDCOT is the one given by f^T at the end of the BCD procedure. In the unsupervised case, since no target label is available to initialise f^T, a first estimate of the target labels is given by label propagation after executing the unsupervised COOT method. Both JDCOT and COOT allow to improve the baseline score and JDCOT's performance is growing along with the number of known target labels. The 3 000-sized sample is big enough to allow a good unsupervised performance for COOT so that introducing some target labels does not improve it. In the 300-sized sample case, introducing semi-supervision allows to considerably improve both COOT and JDCOT's performances. Finally, JDCOT outperforms COOT in all cases and is generally more stable.

$n^T_{k,*}$	baseline	$n^S = n^T = 3\ 000$		$n^S = n^T = 300$	
		COOT	JDCOT	COOT	JDCOT
0	-	72.96 ±8.2	77.27 ±9.1	57.27 ±16.2	58.08 ±17.2
1	39.59 ±6.0	75.81 ±4.9	78.45 ±1.1	61.74 ±14.5	69.98 ±2.8
3	56.82 ±4.4	75.35 ±6.5	79.02 ±0.9	69.71 ±7.2	73.19 ±2.4
10	80.49 ±3.1	75.75 ±6.8	88.34 ±1.7	77.25 ±1.7	85.67 ±1.7

Table 2: Mean and standard deviation of the test accuracy (%) over 10 random samplings for the training sets, considering two sample sizes. $n^T_{*,k}$ denotes the number of known labels in each class k, in target domain.

Partial HDA The partial HDA setting (semi-supervision in both domains) is then considered, with $n^S = n^T = 3\ 000$ and $n^S_{k,*} = n^T_{k,*} \in \{3; 5; 25; 100\}$ labelled samples per class in each domain. Table 3 reports the test accuracy after initialising the model with the available target labels (init) and after the whole process (final), along with the number of known labels per class. JDCOT increases the accuracy both on source and target domains which shows that the information transfer established by solving the OT problem from a domain to another works positively.

JDCOT	$n_{*,k}$	3	5	25	100
source	init	70.9 ±4.3	77.9 ±2.3	92 ±0.7	97.6 ±0.4
	final	73.5 ±5.3	84.6 ±2.5	94.6 ±0.9	98 ±0.2
target	init	62.7 ±3.2	70.5 ±3.4	90.2 ±0.9	96.1 ±0.7
	final	68.7 ±5.5	79 ± 3.2	90.3 ±0.5	97 ± 0.2

Table 3: Mean and standard deviation of the test accuracy (%) over 10 random samplings for the training set. $n^S = n^T = 3\,000$. $n_{*,k}$ denotes the number of known labels in each class k, in each domain.

4 Conclusion and discussion

Joint Distribution Co-Optimal Transport (JDCOT) is a method to deal with heterogeneous transfer learning using optimal transport: it performs domain adaptation in the case of source and target spaces of different features and different dimensions, matching both samples and features with transport maps. The method can deal with unsupervised, semi-supervised and partial domain adaptation.

An extension of JDCOT to a deep learning setting, for example to study image datasets, has also been defined to enable dealing with larger data dimensions by simultaneously optimising 2 transport plans and 2 features extractors g^S and g^T (CNNs). The vector representations of the data are compared instead of the data themselves, and optimisation is done with minibatch stochastic gradient descent.

References

[1] W.-Y. Chen, T.-M. H. Hsu, Y.-H. H. Tsai, M.-S. Chen, and Y.-C. F. Wang. Transfer neural trees: Semi-supervised heterogeneous domain adaptation and beyond. *IEEE Transactions on Image Processing*, 28(9):4620–4633, 2019. doi: 10.1109/TIP.2019.2912126.

[2] N. Courty, R. Flamary, D. Tuia, and A. Rakotomamonjy. Optimal Transport for Domain Adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(9):1853–1865, 2016.

[3] N. Courty, R. Flamary, A. Habrard, and A. Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. volume 30, 2017.

[4] H. Li, S. J. Pan, R. Wan, and A. C. Kot. Heterogeneous transfer learning via deep matrix completion with adversarial kernel embedding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33 (01):8602–8609, 2019.

[5] G. Peyré and M. Cuturi. Computational Optimal Transport. *Foundations and Trends in Machine Learning*, 11(5-6):355–206, 2018.

[6] I. Redko, T. Vayer, R. Flamary, and N. Courty. CO-Optimal Transport. In *Neural Information Processing Systems (NeurIPS)*, Online, Canada, Dec. 2020.

[7] Y. Yao, Y. Zhang, X. Li, and Y. Ye. Heterogeneous domain adaptation via soft transfer network. *CoRR*, abs/1908.10552, 2019.

La Distance de Sliced-Wasserstein pour l'Apprentissage à Grande Échelle

Kimia Nadjahi [1]

[1] *LPSM, Sorbonne Université, Paris, kimia.nadjahi@sorbonne-universite.fr*

Résumé. Les algorithmes actuels d'inférence statistique et d'apprentissage utilisent des distances définies sur l'espace des distributions de probabilité pour comparer des jeux de données. La distance de Wasserstein, issue du problème de transport optimal, est un choix intéressant mais souffre de limites computationnelle et statistique à grande échelle. Plusieurs métriques alternatives ont alors été proposées, notamment la distance de Sliced-Wasserstein (SW), de plus en plus utilisée en pratique en raison de son coût de calcul avantageux. Cependant, peu de travaux avaient analysé SW d'un point de vue théorique, ce qui limite fortement l'étude des méthodes basés sur cette distance. Nous proposons alors d'étudier les propriétés théoriques de SW afin de mieux comprendre son impact dans les problématiques modernes d'apprentissage automatique, notamment la modélisation générative. En particulier, nous prouvons un ensemble de résultats qui montrent la robustesse de SW à la dimension des données, ce qui est un avantage considérable par rapport à la distance de Wasserstein en grande dimension. Néanmoins, SW est couramment estimée en pratique avec une simple méthode de Monte Carlo, qui induit une erreur d'approximation susceptible de détériorer la performance des algorithmes basés sur SW. Pour éviter ce problème, nous introduisons une nouvelle approximation déterministe de SW en s'appuyant sur des résultats de concentration de la mesure. Nous prouvons que l'estimateur obtenu présente des garanties non asymptotiques sous une condition de dépendance faible, tout en étant plus rapide à calculer que la technique de Monte Carlo. Nos contributions sont illustrées sur des problèmes avec données synthétiques ainsi que des applications d'apprentissage profond pour la génération d'images.

Mots-clés. Apprentissage automatique, transport optimal, modélisation générative, analyse de données à grande échelle

Abstract. Modern methods for statistical inference and machine learning rely on distances defined on the space of probability distributions to compare datasets. The Wasserstein distance, which emerges from the optimal transportation problem, has been an interesting choice but suffers from computational and statistical limitations on large-scale settings. Several alternative metrics have then been proposed, including the Sliced-Wasserstein distance (SW), that has been increasingly used in practice due to its computational benefits. However, only a few works have studied SW from a theoretical point of view, which significantly limits the analysis of SW-based methods. Therefore, we provide a thorough study on the theoretical properties of SW in order to better understand its impact in modern statistical inference and machine learning problems, including generative modeling. In particular, we derive a set of results emphasizing the robustness of

SW to the data dimension, which is a strong advantage over the Wasserstein distance in large-scale settings. Nevertheless, SW is commonly estimated with a simple Monte Carlo scheme in practice, which induces an approximation error that can affect the performance of SW-based algorithms. To alleviate this issue, we introduce a novel deterministic approximation of SW by leveraging concentration of measure results. We prove that our estimate has nonasymptotical guarantees under a weak dependence condition, while being computationally more efficient than the usual Monte Carlo technique. Our contributions are illustrated on synthetic problems as well as deep learning applications for image generation.

Keywords. Machine learning, optimal transport, generative modeling, large-scale data analysis

1 Introduction

1.1 Contexte: les distances de transport optimal pour l'apprentissage automatique

Concevoir un algorithme d'apprentissage automatique efficace est un enjeu technique important en raison, entre autres, des quantités importantes de données, de la puissance de calcul limitée des ordinateurs ou encore de la difficulté à paramétrer convenablement la méthode utilisée. De nombreux travaux de recherche visent alors à développer des algorithmes d'apprentissage garantissant des résultats précis mais aussi une certaine facilité d'utilisation – par exemple, un traitement rapide de volumes massifs de données ou un paramétrage peu fastidieux.

La performance d'un algorithme d'apprentissage, ainsi évaluée en fonction de son ergonomie et de la qualité de ses résultats, dépend de plusieurs facteurs clés. En particulier, les algorithmes actuels utilisent des *distances* (ou *divergences*) définies sur l'espace des distributions de probabilité pour comparer des jeux de données. Le choix d'une distance influence fortement la performance de l'algorithme, d'où l'importance de la choisir avec précaution. Un exemple particulièrement illustratif sont les modèles génératifs, dont le but est de générer de nouvelles données similaires à des données observées. Pour cela, ces modèles cherchent généralement à minimiser une distance entre la distribution empirique des observations et une famille de lois de probabilité pouvant facilement être simulées. Plusieurs études ont alors démontré que la difficulté de ce problème de minimisation, et donc la qualité des données générées, dépend de la distance choisie [Arjovsky et al., 2017].

La *distance de Wasserstein*, issue de la théorie mathématique du *transport optimal*, se distingue des divergences existantes de par ses propriétés théoriques intéressantes. Cependant, son utilisation en science des données est *a priori* compliquée puisque son calcul

est en général très coûteux, notamment lorsque la dimension de données est grande. Au cours de ces dernières années, plusieurs méthodes numériques ont alors été développées pour contourner ces difficultés. En particulier, des alternatives pratiques à la distance de Wasserstein ont été introduites, dont la *distance Sliced-Wasserstein* (SW) [Rabin et al., 2012]. SW est définie comme la moyenne des distances de Wasserstein évaluées entre des projections unidimensionnelles aléatoires. La distance de Wasserstein entre distributions univariées étant rapide à calculer, SW bénéficie de coûts de calcul faibles même en grande dimension, expliquant ainsi l'émergence récente de nouvelles méthodologies basées sur SW pour résoudre efficacement plusieurs problèmes, comme des *auto-encodeurs* [Kolouri et al., 2019b], des *réseaux adverses génératifs* [Deshpande et al., 2018, Wu et al., 2019] ou des *flots de gradient* [Liutkus et al., 2019], entre autres.

1.2 Objectifs

Comme expliqué dans la section précédente, les avantages pratiques ainsi que le lien avec la distance de Wasserstein font de SW un outil très prometteur pour de nombreuses applications en grande dimension. Cependant, la grande majorité des contributions sur SW étaient des analyses empiriques mettant essentiellement en valeur la qualité des résultats obtenus et sa vitesse de calcul : peu de travaux avaient étudié SW d'un point de vue théorique, ce qui limitait fortement l'analyse des méthodes d'apprentissage basées sur cette distance.

Partant de ce constat, l'objectif de notre contribution est d'approfondir l'étude de SW, notamment lorsqu'elle est utilisée dans les problématiques modernes d'inférence statistique et d'apprentissage automatique, par exemple la génération d'images avec des réseaux de neurones profonds. Plus précisément, nous souhaitons dans un premier temps apporter de nouvelles propriétés théoriques, principalement statistiques et topologiques, qui permettent de mieux comprendre la performance des algorithmes existants basés sur cette distance et de motiver de nouvelles applications (Section 2). Nous tirons ensuite partie de cette étude théorique et de précédentes études empiriques pour identifier les limites de SW et proposer des solutions pratiques (Section 3). Nos résultats théoriques sont illustrés par des analyses empiriques utilisant des données synthétiques et réelles.

2 Étude théorique de SW et des algorithmes associés

Dans un premier temps, nous nous intéressons aux estimateurs obtenus par minimisation de SW, définis par

$$\hat{\theta}_n = \operatorname{argmin}_{\theta \in \Theta} \mathrm{SW}(\mu_\theta, \hat{\mu}_n) \ ,$$

où μ_θ est une distribution de probabilité paramétrée par $\theta \in \Theta$ et $\hat{\mu}_n$ est la mesure empirique d'un échantillon d'observations $(Y_1, \ldots, Y_n)$ indépendantes et identiquement distribuées selon une distribution inconnue $\mu_\star$. Ces estimateurs ont des retombées pratiques

importantes : par exemple, en inférence statistique traditionnelle, ils sont utilisés pour inférer les paramétres d'une distribution dont la densité n'admet pas de forme analytique; de plus, plusieurs modèles génératifs récemment introduits reposent sur ces estimateurs [Deshpande et al., 2018, Kolouri et al., 2019b, Wu et al., 2019]. L'analyse de ces méthodes reste cependant très limitée, puisque les propriétés théoriques de ces estimateurs n'ont pas été suffisamment étudiées auparavant.

Nous prouvons que ces estimateurs sont *bien définis* puisque la solution au problème de minimisation existe, sous certaines hypothèses sur le modèle statistique. Nous démontrons ensuite un résultat asymptotique qui montre la *cohérence* de ces estimateurs, au sens où plus le nombre de données observées est grand, plus l'estimateur est proche de la solution optimale. Ces résultats justifient par des arguments théoriques l'utilisation des algorithmes basés sur ces estimateurs, comme les modèles génératifs dont l'analyse était majoritairement empirique. Notre étude [Nadjahi et al., 2019] s'inspire de l'analyse de Bernton et al. [2019] pour la distance de Wasserstein, et l'adaptation de leurs techniques de preuve a conduit à l'identification de nouvelles propriétés topologiques liées à SW. En particulier, nous prouvons que la convergence selon SW implique la *convergence faible* des mesures de probabilité en général.

D'autre part, nous prouvons plusieurs résultats qui confirment *l'efficacité de SW en grande dimension* pressentie dans les études empiriques précédentes. Nous établissons d'abord un *théorème central limite*, qui montre qu'en augmentant le nombre n de données observées, les estimateurs convergent vers une variable aléatoire spécifique à une vitesse en $\sqrt{n}$, donc *indépendante de la dimension des données* [Nadjahi et al., 2019]. Nous étudions aussi la *complexité d'échantillonnage*, une propriété statistique importante pour les distances de probabilité : comme l'utilisateur a accès à un nombre limité de données en pratique, une distance entre deux distributions de probabilité est typiquement estimée comme la distance entre des approximations empiriques de ces distributions ; la complexité d'échantillonnage correspond alors à la vitesse de convergence de cette estimation vers la distance entre les distributions d'origine quand le nombre de données croît. Nous démontrons que la complexité d'échantillonnage de SW ne dépend pas de la dimension des données, du fait de l'utilisation de l'opération de "slicing" [Nadjahi et al., 2020].

SW présente donc des propriétés statistiques *robustes à la dimension*, ce qui constitue un avantage considérable par rapport à la distance de Wasserstein. En effet, la complexité d'échantillonnage de la distance de Wasserstein augmente de façon exponentielle en la dimension, ce qui peut entraîner des coûts de calculs très élevés en grande dimension [Bernton et al., 2019, Weed and Bach, 2019]. Notons que certains de nos résultats s'appliquent plus généralement à la famille de distances obtenues par "slicing", introduite par Nadjahi et al. [2020].

3 Identification des limites de SW et solutions

L'étude théorique permet également d'identifier les limites de SW, déjà observées en pratique. En effet, la technique couramment utilisée pour calculer SW est une simple méthode de Monte Carlo qui, d'après des travaux antérieurs, introduit une erreur d'approximation susceptible de déteriorer la performance des algorithmes basés sur SW [Deshpande et al., 2019]. Nous confirmons cette observation empirique par un nouveau résultat théorique : l'évaluation de SW entraîne effectivement une erreur d'approximation causée par le nombre limité de données et par la méthode de calcul traditionnellement employée [Nadjahi et al., 2020]. Plus précisément, cette méthode introduit une variance supplémentaire dans l'erreur, qui peut être élevée en grande dimension. De plus, elle dépend d'un paramètre dont le choix de la valeur reste peu clair.

Plusieurs approches ont alors été développées pour contourner les limites de l'estimation par Monte Carlo et consistent généralement à choisir des échantillons plus informatifs [Deshpande et al., 2019, Kolouri et al., 2019a, Nguyen et al., 2020]. Cependant, ces méthodes ne sont pas soutenues par des garanties théoriques et s'appuient sur des problèmes d'optimisation difficile ou très coûteux à résoudre.

Nous adoptons alors une autre perspective pour proposer une alternative efficace à la méthode de Monte Carlo, inspirée par la littérature sur la *concentration des projections aléatoires* [Nadjahi et al., 2021]. Nous montrons que les estimations de SW ainsi obtenues ont des garanties théoriques satisfaisantes au sens où plus la dimension augmente, plus l'estimation est précise (sous une hypothèse de dépendance faible sur les données). Cette nouvelle technique d'estimation est *non paramétrique* et s'appuie sur des opérations déterministes, moins coûteuses que la méthode traditionnelle. Notre analyse empirique confirme que notre méthode peut facilement être incorporée dans des modèles génératifs existants, originellement basés sur l'estimateur de SW par Monte Carlo, afin d'accélérer leur exécution et d'améliorer la précision des résultats en grande dimension.

Bibliographie

M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein Generative Adversarial Networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

E. Bernton, P. E. Jacob, M. Gerber, and C. P. Robert. On parameter estimation with the Wasserstein distance. *Information and Inference: A Journal of the IMA*, 8(4):657–676, 2019.

I. Deshpande, Z. Zhang, and A. Schwing. Generative modeling using the sliced Wasser-

stein distance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3483–3491, 2018.

I. Deshpande, Y.-T. Hu, R. Sun, A. Pyrros, N. Siddiqui, S. Koyejo, Z. Zhao, D. Forsyth, and A. G. Schwing. Max-Sliced Wasserstein distance and its use for GANs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 10648–10656, 2019.

S. Kolouri, K. Nadjahi, U. Şimşekli, R. Badeau, and G. Rohde. Generalized Sliced Wasserstein Distances. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019a.

S. Kolouri, P. E. Pope, C. E. Martin, and G. K. Rohde. Sliced Wasserstein Auto-Encoders. In *International Conference on Learning Representations*, 2019b.

A. Liutkus, U. Simsekli, S. Majewski, A. Durmus, and F.-R. Stöter. Sliced-Wasserstein flows: Nonparametric generative modeling via optimal transport and diffusions. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4104–4113, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

K. Nadjahi, A. Durmus, U. Şimşekli, and R. Badeau. Asymptotic Guarantees for Learning Generative Models with the Sliced-Wasserstein Distance. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

K. Nadjahi, A. Durmus, L. Chizat, S. Kolouri, S. Shahrampour, and U. Şimşekli. Statistical and Topological Properties of Sliced Probability Divergences. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

K. Nadjahi, A. Durmus, P. E. Jacob, R. Badeau, and U. Şimşekli. Fast Approximation of the Sliced-Wasserstein Distance Using Concentration of Random Projections. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

K. Nguyen, N. Ho, T. Pham, and H. Bui. Distributional Sliced-Wasserstein and applications to generative modeling, 2020.

J. Rabin, G. Peyré, J. Delon, and M. Bernot. Wasserstein barycenter and its application to texture mixing. In *Scale Space and Variational Methods in Computer Vision*, pages 435–446, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. ISBN 978-3-642-24785-9.

J. Weed and F. Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *Bernoulli*, 25(4A):2620–2648, 11 2019.

J. Wu, Z. Huang, D. Acharya, W. Li, J. Thoma, D. P. Paudel, and L. V. Gool. Sliced Wasserstein Generative Models. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3713–3722, 2019.

15 juin 2022

Programme

SFdS - Società Italiana di Statistica **667**

Analysing spatio-temporal patterns in the association of confirmed deaths by COVID-19 and intensive care admissions, Andrea Bucci [et al.] 668

Sieve bootstrap based on extreme learning machines in time series analysis, Cira Perna [et al.] 674

Advances in clustering asymmetric proximity data, Donatella Vicari 676

SFdS - ENBIS **678**

Scalable additive stacking models for electricity demand forecasting, Euan Enticott679

Metamodeling and sensitivity analysis for models with spatial output. Application to coastal flooding models., Tran Vi-Vi Élodie Perrin [et al.] 680

Textual Data for Electricity Demand Forecasting, David Obst [et al.] 686

On forecasting bounded variables with latent stochastic bounds, Amandine Pierrot [et al.] 687

SFdS - Sport **693**

L'origine de la déficience visuelle est-elle un facteur de performance ? Analyse des para-nageurs et para-athlètes de niveau international, Julien Schipman [et al.] . . 694

Relative age effect among French swimmers, Audrey Difernand 701

Parametric and non-parametric modeling by a Makovian multistate model of the evolution of performance in French alpine skiing, Quentin De Larochelambert [et al.] 707

Reparlons de notes en patinage, Léo Gerville-Réache 713

Modélisation de la probabilité de réussite d'une passe au football, Sébastien Déjean [et al.] 718

Estimation of race scenarios in standard triathlon according to level density, Thibaut Ledanois 724

Statistique mathématique **731**

Estimateurs non paramétriques à noyau général birnbaum-saunders des densités multivariées à support non négatifs, Smail Adjabi 732

Integration of bounded monotone functions: Revisiting the nonsequential case, with a focus on unbiased Monte Carlo (randomized) methods, Subhasish Basak [et al.] 738

Subsampling under constraint, Florian Combes [et al.] 744

Recursive conditional distribution estimation using kernel method and Bernstein polynomials, Salima Helali [et al.] 750

Explainability as statistical inference, Hugo Senetaire [et al.] 751
La Géométrie pour l'Unicité, la Parcimonie et le Regroupement des Estimateurs Pénalisés, Patrick Tardivel [et al.] 759
PLENIERE : Li-Chun Zhang **765**
PLENIERE : Emilie Kaufmann **766**

PLENIERE : Markus Reiss

Rank detection for time-varying covariance matrices and how Le Cam Theory may help

Markus Reiß

Humboldt-Universität zu Berlin

Rank detection for covariance matrices is one of the fundamental inference problems in statistics. Here we focus on the case of a time-varying instaneous (or spot) covariance matrix $S(t)$ of a continuous-time process $X(t)$. The data are given by high-frequency observations of X on $[0,T]$, possibly corrupted by noise. We ask for testing the hypothesis $\mathbb{H}_1 : \int_0^T \lambda_{r+1}(S(t))\mathrm{d}t \geq v_n$ i.e. that the mean $(r+1)$st eigenvalue is larger than some signal detection rate v_n, tending to zero with sample size n. This problem can be embedded in the classical nonparametric signal detection framework, but it has many unexpected features. For instance, the optimal detection rate depends on a regularity assumption on $S(t)$ under the null, not the alternative and a possible spectral gap leads to significantly better detection rates. We show how an asymptotically equivalent nonparametric Gaussian white noise model may help to understand the structure of the statistical problem and to come up with efficient and implementable methods. The rank detection is illustrated with applications to intraday data from government bonds.

Based on joint work with Lars Winkelmann, Markus Bibinger, Nikolaus Hautsch, Peter Malec

SFdS – Società Italiana di Statistica

Analysing spatio-temporal patterns in the association of confirmed deaths by COVID-19 and intensive care admissions

Andrea Bucci [1,a] & Luigi Ippoliti [1,b] & Pasquale Valentini [1,c]

[1] *'G. d'Annunzio' University of Chieti-Pescara, Italy*
[a] *andrea.bucci@unich.it*
[b] *luigi.ippoliti@unich.it*
[c] *pasquale.valentini@unich.it*

Résumé. En considérant les données régionales NUTS-2, ce document explore si, et dans quelle mesure, les données sur les unités de soins intensifs (USI) peuvent expliquer la dynamique des décès confirmés par COVID-19. Sur la base d'une approche non paramétrique Bayésienne, les problèmes de regroupement spatial et de prédiction temporelle sont pris en compte.

Mots-clés. COVID-19, Analyse Bayésienne non paramétrique, Mélange de gaussiennes, Modéle de paramétres variant dans le temps

Abstract. Understanding whether and to what extent the intensive care capacity strain for COVID-19 affects the death rate is critical to preparing authorities for further potential pandemic waves. As responses from governments to the COVID-19 pandemic have been varied across countries and through time, interest may also focus on finding groups of Regions with similar behaviour in terms of that association. To this end, we propose a flexible Bayesian nonparametric approach, based on a mixture of Gaussian processes coupled with the Dirichlet process, capable of clustering time series in terms of the time-varying parameter relative to this association, and allowing spatial correlation, measurement errors and predictions. We evaluate the proposed methodology on the weekly counts of deaths and new admissions to intensive care recorded at the NUTS-2 regional level for several European countries.

Keywords. COVID-19, Bayes nonparametric, Model-based clustering, Time-varying parameter model

1 Introduction

The ongoing coronavirus disease 2019 (COVID-19) has severely affected the worldwide economy, society and human health. Since its spread, the responses from governments to the pandemic have been varied across countries and through time. This has lead to different dynamics in terms of deaths and hospitalizations in individual countries, and

within the administrative regions. Consequently, considering both the temporal and the spatial components of the pandemic has become crucial to understand the patterns of the disease and assess the effectiveness of preventive interventions on health outcomes.

Recently, several studies have analysed the spatio-temporal evolution of the COVID-19 pandemic from different perspectives. For instance, several Bayesian models have been developed by Chen et al. (2020), Sahu & Böhning (2021), Lee et al. (2021) and Bartolucci & Farcomeni (2021), to assess and predict the distribution of the pandemic at different phases. From a different perspective, a strand of the recent literature has focused on finding groups of administrative units at several levels in terms of pandemic-related variables, see Zhang et al. (2020), Shariati et al. (2020), D'Urso et al. (2021) and Bucci et al. (2021).

What remains still unclear is whether and to what extent hospital capacity strain, especially in intensive care units (ICU), impacts patients outcomes. In other words, policy makers may be interested in understanding if and how the increased patient volume in ICU affects the number of confirmed deaths, and if this association varies through time and space. Prior to COVID-19, the evidence on the association between mortality and ICU strain have been mixed. Some studies suggest that higher ICU occupancy has been associated with increased mortality risk (lapichino et al. 2004, Chrusch et al. 2009), while others find no association (Taccone et al. 2021). A clearer understanding of how this association operates, like finding groups of regions with similar effects of ICU occupancy on deaths, can provide policy makers with more accurate information about the potential consequences of allowing health services to be overwhelmed.

Few papers have analysed the impact of ICU strain on deaths by COVID-19 (Baumer et al. 2020, Burrell et al. 2021, Bravata et al. 2021, Wilde et al. 2021). In this work, we contribute to this literature by proposing to cluster European Regions through a Bayesian nonparametric approach with multiple goals: i) to understand if an association between ICU admissions and confirmed deaths exists; ii) to analyse if this relationship has changed through the different phases of the pandemic; iii) to find groups of regions with similar patterns of the association; iv) to produce short-term forecasts of confirmed number of deaths. In fact, modelling the weekly time series patterns as a mixture of Gaussian processes (GP) with a Dirichlet process (DP) prior (Ferguson 1973) over mixture components allows us to automatically detect a number of clusters and to account for time-varying parameters, spatial correlation, and measurement noise. This specification permits also to produce out-of-sample forecasts, which can be helpful in preparing authorities to the effects of further pandemic waves.

2 Motivating example: ICU admissions and confirmed COVID-19 deaths in Spain and France

The data used to study the association between ICU admissions and confirmed COVID-19 deaths are sampled at the NUTS-2 level (*i.e.* Regions) for Spain and France at weekly frequency. These data provide the basis for challenging modelling issues. The deaths and ICU admissions per milion inhabitants data, stratified by country, are depicted in Figure 1. The temporal evolution of the pandemic is characterised by three or four peaks in the periods March - May 2020, October 2020 - March 2021, July-September 2021 and at the end of December 2021 and by low death counts for all regions during the summer. As expected, the deaths and ICU admissions time series within the same country present similar patterns; however, a high variability can be observed in all the country-level plot, especially for the second and third waves. This suggests that the impact of the pandemic has been highly asymmetric within countries and that there might exist a supranational pattern across regions.

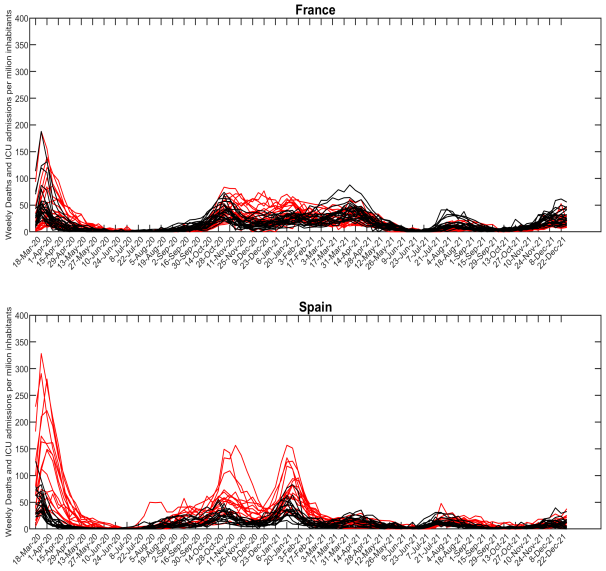


Figure 1: *Deaths (red) and ICU admissions (black) per milion inhabitants due to COVID-19 from March 2020 to December 2021 in France and Spain.*

3 The statistical model

Following Section 2, we work with population level summaries of confirmed COVID-19 deaths (Y) and ICU admissions (X) collected over time over a fixed study area, $\mathcal{D}$. Henceforth, the complete set of information for the deaths data is denoted as $\{Y_t(\mathbf{s}_i), \mathbf{s}_i \in \mathcal{D}\}$, $i = 1, \ldots, n;\ t = 1, \ldots, T$, where n is the number of areal units, $\mathbf{s}_i$, within $\mathcal{D}$ and T is the length of the time series. We also assume that at the first two levels of the hierarchy, the model for the data is given by

$$Y_t(\mathbf{s}_i) \overset{ind}{\sim} Poisson(\exp(Z_t(\mathbf{s}_i))) \quad (1)$$

$$Z_t(\mathbf{s}_i) = \eta_t(\mathbf{s}_i) + \varepsilon_t(\mathbf{s}_i) \quad (2)$$

where $\eta_t(\mathbf{s}_i)$ is the true (latent) process of interest and $\varepsilon_t(\mathbf{s}_i)$ is a spatially and temporally uncorrelated Gaussian error term, assumed to have zero mean and time varying variance, *i.e.* $\varepsilon_t(\mathbf{s}_i) \sim N\left(0, \sigma^2_{\varepsilon,t}\right)$.
At the third level of the hierarchy, the true latent process is given by

$$\eta_t(\mathbf{s}_i) = \mu_t(\mathbf{s}_i) + X_t(\mathbf{s}_i)\beta_t(\mathbf{s}_i) + \nu_t(\mathbf{s}_i), \quad (3)$$

where $\mu_t(\mathbf{s}_i)$ is a general time-varying mean, $X_t(\mathbf{s}_i)$ is the number of new admissions to ICU by COVID-19 in the previous week, $\beta_t(\mathbf{s}_i)$ is a site specific regression coefficient which varies over time and $\nu_t(\mathbf{s}_i)$ is an error term uncorrelated in time but correlated in space (*i.e.*, specified as a Conditional Autoregressive process).
Since our goal is the clustering of the regression coefficients, we place DP priors on the temporal regression coefficients, $\beta_t(\mathbf{s}_1), \ldots, \beta_t(\mathbf{s}_n)$, thereby inducing ties among their values.
The general model given in equations (1) to (3) is developed specifying the prior distributions of all hyperparameters within a state-space framework. The full probabilistic inference for the parameters is facilitated by a Markov chain Monte Carlo (MCMC) scheme for multivariate dynamic systems.

Preliminary results from the general and other nested models suggest that the lagged weekly number of new ICU admissions is strongly associated with weekly confirmed deaths and that this association is time-varying and clustered at regional level. Inferential issues and the statistical analysis will be discussed in an extended version of this abstract.

References

Bartolucci, F. & Farcomeni, A. (2021), 'A spatio-temporal model based on discrete latent variables for the analysis of COVID-19 incidence', *Spatial Statistics* .

Baumer, T., Phillips, E., Dhadda, A. & Szakmany, T. (2020), 'Epidemiology of the First Wave of COVID-19 ICU Admissions in South Wales—The Interplay Between Ethnicity

and Deprivation', *Frontiers in Medicine* **7**.
URL: *https://www.frontiersin.org/article/10.3389/fmed.2020.569714*

Bravata, D. M., Perkins, A. J., Myers, L. J., Arling, G., Zhang, Y., Zillich, A. J., Reese, L., Dysangco, A., Agarwal, R., Myers, J., Austin, C., Sexson, A., Leonard, S. J., Dev, S. & Keyhani, S. (2021), 'Association of Intensive Care Unit Patient Load and Demand With Mortality Rates in US Department of Veterans Affairs Hospitals During the COVID-19 Pandemic', *JAMA Network Open* **4**(1), e2034266–e2034266.
URL: *https://doi.org/10.1001/jamanetworkopen.2020.34266*

Bucci, A., Ippoliti, L., Valentini, P. & Fontanella, S. (2021), 'Clustering spatio-temporal series of confirmed COVID-19 deaths in Europe', *Spatial Statistics* p. 100543.
URL: *https://www.sciencedirect.com/science/article/pii/S2211675321000531*

Burrell, A. J., Pellegrini, B., Salimi, F., Begum, H., Broadley, T., Campbell, L. T., Cheng, A. C., Cheung, W., Cooper, D. J., Earnest, A., Erickson, S. J., French, C. J., Kaldor, J. M., Litton, E., Murthy, S., McAllister, R. E., Nichol, A. D., Palermo, A., Plummer, M. P., Ramanan, M., Reddi, B. A., Reynolds, C., Trapani, T., Webb, S. A. & Udy, A. A. (2021), 'Outcomes for patients with covid-19 admitted to australian intensive care units during the first four months of the pandemic', *Medical Journal of Australia* **214**(1), 23–30.
URL: *https://doi.org/10.5694/mja2.50883*

Chen, Z.-L., Zhang, Q., Lu, Y., Guo, Z.-M., Zhang, X., Zhang, W.-J., Guo, C., Liao, C.-H., Li, Q.-L., Han, X.-H. & Lu, J.-H. (2020), 'Determining the spatial effects of COVID-19 using the spatial panel data model', *Chinese Medical Journal* **133**, 10441050.

Chrusch, C. A., Olafson, K. P., McMillan, P. M., Roberts, D. E. & Gray, P. R. (2009), 'High occupancy increases the risk of early death or readmission after transfer from intensive care', *Critical Care Medicine* **37**(10).

D'Urso, P., Mucciardi, M., Otranto, E. & Vitale, V. (2021), 'Community mobility in the European regions during COVID-19 pandemic: a partitioning around medoids with noise cluster based on space-time autoregressive models', *Spatial Statistics* .

Ferguson, T. S. (1973), 'A Bayesian Analysis of Some Nonparametric Problems', *Annals of Statistics* **1**(2), 209–230.

lapichino, G., Gattinoni, L., Radrizzani, D., Simini, B., Bertolini, G., Ferla, L., Mistraletti, G., Porta, F. & Miranda, D. R. (2004), 'Volume of activity and occupancy rate in intensive care units. association with mortality', *Intensive Care Medicine* **30**(2), 290–297.
URL: *https://doi.org/10.1007/s00134-003-2113-4*

Lee, D., Robertson, C. & Marques, D. (2021), 'Quantifying the small-area spatio-temporal dynamics of the COVID-19 pandemic in Scotland during a period with limited testing capacity', *Spatial Statistics* .

Sahu, S. & Böhning, D. (2021), 'Bayesian spatio-temporal joint disease mapping of COVID-19 cases and deaths in local authorities of England', *Spatial Statistics* .

Shariati, M., Mesgari, T., Kasraee, M. & Jahangiri-rad, M. (2020), 'Spatiotemporal analysis and hotspots detection of COVID-19 using geographic information system (March and April, 2020)', *Journal of Environmental Health Science and Engineering* **18**(2), 1499–1507.

Taccone, F. S., Van Goethem, N., De Pauw, R., Wittebole, X., Blot, K., Van Oyen, H., Lernout, T., Montourcy, M., Meyfroidt, G. & Van Beckhoven, D. (2021), 'The role of organizational characteristics on the outcome of COVID-19 patients admitted to the ICU in Belgium', *The Lancet Regional Health – Europe* **2**.
URL: *https://doi.org/10.1016/j.lanepe.2020.100019*

Wilde, H., Mellan, T., Hawryluk, I., Dennis, J. M., Denaxas, S., Pagel, C., Duncan, A., Bhatt, S., Flaxman, S., Mateen, B. A. & Vollmer, S. J. (2021), 'The association between mechanical ventilator compatible bed occupancy and mortality risk in intensive care patients with covid-19: a national retrospective cohort study', *BMC Medicine* **19**(1), 213.
URL: *https://doi.org/10.1186/s12916-021-02096-0*

Zhang, X., Rao, H., Wu, Y., Huang, Y. & Dai, H. (2020), 'Comparison of spatiotemporal characteristics of the COVID-19 and SARS outbreaks in mainland China', *BMC Infectious Diseases* **20**(1), 805.

Sieve bootstrap based on extreme learning machines in time series analysis

Michele La Rocca [1] & Cira Perna [1]

[1] *Department of Economics and Statistics*
University of Salerno (IT)
Via Giovanni Paolo II, 132 - 84084 Fisciano (SA)
larocca@unisa.it; perna@unisa.it

Abstract. The aim of the talk is to propose and discuss a sieve bootstrap scheme based on Extreme Learning Machines for non linear time series. The use of Extreme Learning Machines in the resampling scheme can dramatically reduce the computational burden of the bootstrap procedure, with performances comparable to the NN-Sieve bootstrap and computing time similar to the AR-Sieve bootstrap. Moreover, the proposed procedure is fully nonparametric in its spirit and retains the conceptual simplicity of the residual bootstrap. A Monte Carlo simulation experiment has been implemented, in order to evaluate the performance of the proposed procedure and to compare it with the NN-Sieve bootstrap.

Keywords. Sieve bootstrap; nonlinear time series; extreme learning machines; neural networks.

1 Soft computing tools for statistical inference

In the last decades, in the context of time series, several bootstrap procedures have been proposed to infer properties of a statistic of interest, including both parametric and non-parametric resampling schemes. When it is reasonable to impose parametric type assumptions on the process, so that the dependence structure of the series is explicitly modelled, the residual bootstrap is usually a sensible choice. However, it is very sensitive to model misspecification, in which case it leads to bootstrap estimators, which might be not consistent. In this latter case, more complex and fully non-parametric bootstrap methods are required and, among them, sieve bootstrap schemes have become useful and powerful tools. For general nonlinear processes, an approach, which employs neural network models to approximate the original nonlinear process has been proposed in Giordano et.al (2007). The method is flexible and easy to implement as a standard residual bootstrap scheme while retaining the advantage of being a nonparametric technique (see Giordano et al, 2011). It is model-free within a general class of nonlinear processes and avoids the specification of a finite dimensional parametric model for the data generating process.

However, despite their proven theoretical capabilities of non-parametric data-driven universal approximation of general nonlinear functions, NNs face challenging issues concerning the computational burden, which can be very heavy especially for complex nonlinear generating processes. Moreover, in many applications of the bootstrap, to incorporate uncertainty due to model estimation, the statistical model needs to be estimated on each bootstrap resampled data. This makes the use of the bootstrap unfeasible when applied to complex estimation procedures. Finally, many statistical problems could be solved in practice only by using iterative bootstrap, which of course requires very efficient estimation procedures. To overcome these problems, our proposal (La Rocca and Perna, 2019) is to estimate the neural network model in the NN-Sieve procedure by using learning without iterative tuning. This approach, known as Extreme Learning Machines (ELM) in the computational intelligence and machine learning literature, has become a very popular learning tool for Single-hidden Layer Feedforward Networks. In this algorithm, the input weights and the hidden layer bias are randomly chosen with remarkable computational advantages with respect to standard neural network models. The ELM formulation leads to solving a system of linear equations in terms of the unknown weights connecting the hidden layer to the output layer. The solution of this general system of linear equations is usually obtained using Moore-Penrose generalised pseudoinverse. The aim of the talk is to propose and discuss a sieve bootstrap scheme based on ELM for nonlinear time series. The procedure is fully nonparametric in its spirit and retains the conceptual simplicity of the AR-Sieve bootstrap. Using ELMs in the resampling scheme can dramatically reduce the computational burden of the bootstrap procedure, with performances comparable to the NN-Sieve bootstrap and computing time similar to the AR-Sieve bootstrap. Moreover, NAR-Sieve bootstrap could be easily extended to multivariate nonlinear time series making this bootstrap scheme suitable for the construction of inference tools for complex vector time series.

References

Giordano, F., La Rocca, M. and Perna, C. (2007). Forecasting nonlinear time series with neural network sieve bootstrap, *Comput. Stat. Data Anal.*, 51, pp 3871–3884.
Giordano, F., La Rocca, M. and Perna, C. (2011). Properties of the neural network sieve bootstrap, *J. Nonparametr. Stat.*, 23, pp. 803–817.
La Rocca, M. and Perna, C. (2019). Nonlinear autoregressive sieve bootstrap based on extreme learning machines, *Math. Biosci. Eng.*, 17, pp .636–653

ADVANCES IN CLUSTERING ASYMMETRIC PROXIMITY DATA

Donatella Vicari

Dipartimento di Scienze Statistiche, Sapienza Università di Roma, Italy
e-mail: donatella.vicari@uniroma1.it

Résumé. L'ensemble à classer est caractérisé par une matrice de proximité carrée asymétrique telle qu'une matrice d'échange. Quelques modèles de classification non hiérarchiques sont présentés, basés sur la décomposition de cette matrice dans sa partie symétrique et antisymétrique.

Mots-clés. Données asymétriques, classification.

Abstract. Some non-hierarchical clustering models for asymmetric proximity data are presented which rely on the decomposition into symmetric and skew-symmetric components.

Keywords. Asymmetric data, clustering.

1. Introduction and background

Asymmetric relations arise in many applicative domains where the focus is the analysis of systems of relationships between pairs of entities, such as confusion data, sociomatrices, brand switching, import–export data, flows, migration data, just to give some non-exhaustive examples. The asymmetric nature of such data has been ignored for long time and standard methods have been applied after symmetrizing under the assumption that asymmetry was negligible and/or due to noise. But when data are intrinsically asymmetric by nature such an assumption is wrong and appropriate methods need to be used to take into account the directions of the exchanges as well as their intensities.

Many methodologies have been proposed to explain the asymmetric effects especially in the context of multidimensional scaling for representing objects in low-dimensional spaces, while less attention has been paid to clustering methodologies. Actually, the first main clustering proposals followed a hierarchical approach by extending the standard hierarchical agglomerative algorithms which have been adapted as well as the graphical representations of the resulting hierarchy to display generalized dendrograms (see Saito and Yadohisa, 2005, and Bove et al., 2021, for comprehensive reviews).

Some clustering models are presented here which are framed in a non-hierarchical framework and rely on the decomposition of the asymmetric proximity data in its symmetric and skew-symmetric components.

2. Asymmetric clustering models

Let $\mathbf{\Omega}=[\omega_{ij}]$ be an $(N\times N)$ two-way one-mode matrix of dissimilarities measured on N objects, where ω_{ij} is generally different from ω_{ji} $(i,j=1,\ldots,N)$. Let us recall that any square matrix $\mathbf{\Omega}$ can be uniquely decomposed into a sum of a symmetric matrix **M** and a skew-symmetric matrix **N**

$$\mathbf{\Omega} = \mathbf{M} + \mathbf{N} = \tfrac{1}{2}\left(\mathbf{\Omega} + \mathbf{\Omega}'\right) + \tfrac{1}{2}\left(\mathbf{\Omega} - \mathbf{\Omega}'\right), \tag{1}$$

where **M** and **N** are orthogonal to each other, i.e. $trace(\mathbf{MN}) = 0$. The entry m_{ij} of **M** represents the average amount of the exchange between objects *i* and *j*, while the entry n_{ij} of **N** is the imbalance between *i* and *j*, i.e. the amount by which n_{ij} differs from the mean m_{ij}, ($i,j=1,\ldots,N$).

The models presented here jointly handles the symmetric and skew-symmetric components which both subsume the same partition of the *N* objects into *J* clusters identified by the membership matrix **U**. Generally speaking, two clustering structures **S** and **K** for the two components both depending by the common partition are defined which are in turn further decomposed into their *within* and *between* parts

$$\begin{aligned} \mathbf{M} &= \mathbf{S}(\mathbf{U}) + \mathbf{E}_{\mathbf{S}} = \mathbf{S}_W(\mathbf{U}) + \mathbf{S}_B(\mathbf{U}) + \mathbf{E}_{\mathbf{S}} \\ \mathbf{N} &= \mathbf{K}(\mathbf{U}) + \mathbf{E}_{\mathbf{K}} = \mathbf{K}_W(\mathbf{U}) + \mathbf{K}_B(\mathbf{U}) + \mathbf{E}_{\mathbf{K}} \end{aligned}, \tag{2}$$

where $\mathbf{E}_{\mathbf{S}}$ and $\mathbf{E}_{\mathbf{K}}$ denote the parts of **M** and **N** not accounted for by the models.

A) In the first nonhierarchical clustering model (Vicari, 2014) the two clustering structures **S** and **K** in (2) are specified which rely on the same partition **U** and on different sets of weights to account for the within and between cluster effects. In the resulting partition, objects in the same cluster share the same behaviors in terms of exchanges directed towards the other clusters, i.e., some clusters are mainly origins towards (destinations from) some other clusters.

B) An extension of the model is proposed in Vicari (2018) which incorporates possible external variables **X** into the skew-symmetric clustering structure **K**(**U**, **X**) to be used as covariates to explain the imbalances between objects.

C) Differently, in Vicari (2020) the symmetric and skew-symmetric components of **Ω** are jointly fitted by two different partitions nested one into the other so that in the partition obtained from the skew-symmetries some objects are allowed to remain unassigned to any cluster: this indicates that such objects present a low degree of asymmetry.

All models are fitted in a least-squares framework by efficient ALS algorithms and results are graphically displayed.

References

Bove, G., Okada, A. and Vicari, D. (2021), *Methods for the Analysis of Asymmetric Relationships*, Series: Behaviormetrics: Quantitative Approaches to Human Behavior, Springer Nature, Singapore.

Saito, T. and Yadohisa, H. (2005), *Data analysis of asymmetric structures*. *Advanced approaches in computational statistics*. Marcel Dekker.

Vicari, D. (2014), Classification of asymmetric proximity data, *Journal of Classification*, 31, 386–420.

Vicari, D. (2018), CLUSKEXT: CLUstering model for SKew-symmetric data including EXTernal information, *Advances in Data Analysis and Classification*, 12, 43–64.

Vicari, D. (2020), Modelling asymmetric exchanges between clusters. In: Imaizumi T., Nakayama A. and Yokoyama S. (Eds.), Advanced studies in behaviormetrics and data science. *Behaviormetrics: Quantitative approaches to human behavior*, 297–313, Springer Nature.

SFdS - ENBIS

Scalable additive stacking models for electricity demand forecasting

Euan Enticott [1]

[1]University of Bristol – Royaume-Uni

Résumé

This submission is for a special session.
Short-term electricity demand forecasts are regularly used to inform decisions in grid management. However, the increasing reliance on renewable production will create a new challenge in demand forecasting. Renewable energy sources are less centralised and often dependent on external factors such as weather. To limit the need for large-scale and expensive storage infrastructure, smart grid management systems can be employed. These systems will require reliable demand forecasts at lower levels of aggregation, possibly down to the individual household level. At this level, demand is characterised by a low signal-to-noise ratio, with frequent abrupt changepoints in demand dynamics. The challenges posed by forecasting at a low level of aggregation motivate the use of an ensemble approach that can incorporate information from several models and across households. The idea of additive stacking for probabilistic forecasting was proposed by Capezza (2020). However, the number of models that could be used was limited as the number of unknown parameters in the stacking model scales linearly with the number of experts.
In this talk, we will discuss more scalable solutions for additive stacking. Specifically, we explore the idea of adding structure to the weights to reduce the number of parameters required. This provides modelling advantages by reducing both computational complexity and the risk of overfitting. The area of application is not limited to demand forecasting, but extend to any setting where ensemble modelling is used.

Metamodelling and sensitivity analysis for models with spatial output. Application to coastal flooding models.

Tran Vi-Vi Élodie Perrin [1] & Olivier Roustant [2] & Jérémy Rohmer [3] & Jean-Philippe Naulin [4]

[1] *ONERA/DTIS, Université de Toulouse, F-31055 Toulouse, France*
tran-vivi-elodie.perrin@outlook.fr
[2] *Institut de Mathématiques de Toulouse, Université de Toulouse, INSA, 135, Avenue de Rangueil, 31077 Toulouse Cedex 4, France*
roustant@insa-toulouse.fr
[3] *BRGM, 3 av. Claude Guillemin, BP 36009, 45060 Orléans Cedex 2, France*
J.Rohmer@brgm.fr
[4] *CCR, 157 Boulevard Haussmann, 75008 Paris, France*
jnaulin@ccr.fr

Résumé. Motivé par l'évaluation des risques de submersions marines, on considère les modèles hydrodynamiques numériques développés par le BRGM et la CCR. La sortie de ces simulateurs est une carte d'inondation. L'objectif est de réaliser une analyse de sensibilité (AS) afin de mesurer et de hiérarchiser l'influence des paramètres d'entrée sur la sortie. Afin de réduire le temps de calcul des modèles et la dimension de la sortie spatiale, on propose d'utiliser l'ACP fonctionnelle (ACPF). La sortie est décomposée dans une base de fonctions, adaptée pour traiter les variations locales, telle que les ondelettes ou les B-splines. Une ACP avec une métrique ad-hoc est appliquée aux coefficients les plus importants, selon un critère d'énergie après orthonormalisation de la base, ou directement sur la base originale avec une approche de régression pénalisée. Des méta-modèles (comme le krigeage) sont construits sur les premières composantes principales, sur lesquels peut être réalisée l'AS. Comme résultat complémentaire, une formule analytique est obtenue pour les indices de sensibilité basés sur la variance, généralisant celle connue pour des bases orthonormées. L'ensemble des travaux a été appliqué à un cas d'inondation côtière. Les processus d'inondation ont été simulés avec les modèles numériques du BRGM et de la CCR.

Mots-clés. Analyse de sensibilité globale, données spatiales, analyse en composantes principales fonctionnelle, ondelettes, B-splines.

Abstract. Motivated by the risk assessment of coastal flooding, the numerical hydrodynamic models of the BRGM and the CCR are considered. Their outputs are flood maps. The aim is to perform a sensitivity analysis (SA) to quantify and hierarchize the influence of the input parameters on the output. The application of functional PCA (FPCA) is proposed to reduce both computation time and spatial output dimension. The output is decomposed on a basis of functions designed to handle local variations, such as wavelets

or B-splines. PCA with an ad-hoc metric is applied on the most important coefficients, according to an energy criterion after basis orthonormalization, or on the initial basis with a penalized regression approach. Fast-to-evaluate metamodels (such as Kriging) are built on the first principal components, on which SA can be done. As a by-product, we obtain analytical formulas for variance-based sensitivity indices, generalizing a known formula assuming the orthonormality of basis functions. The whole methodology is applied on a case of coastal flooding. Coastal flooding processes are simulated with the BRGM and CCR numerical models.

Keywords. Global sensitivity analysis, spatial data, functional principal component analysis, wavelet, B-splines

This communication presents the work in [Perrin et al., 2021] and [Perrin, 2021] on metamodelling and global sensitivity analysis for models with spatial output, with an application to coastal flooding model.

Coastal flooding may lead to major natural disasters in coastal regions, as exemplified by several recent events like cyclone Irma in 2017 or Hurricane Sandy in 2012. In France, the last major event is Xynthia storm that induced 53 deaths, 79 injured people and 2.5 billions euros of damages, whose 700 million euros for coastal flooding (see e.g., [Naulin et al., 2015],[FFSA and Gema, 2011]). The technical pillar of any flooding risk assessment is the capability for accurate and robust predictions of the inland consequences (i.e. water levels at the coast, flood spatial extent, etc.) given any offshore meteo-oceanic conditions (like surge peak, tide peak, storm duration, wave characteristics, etc.). This can be done using high-resolution hydrodynamic numerical models (i.e. simulators), such as those developed by the BRGM (the French geological survey) and the CCR (the French state reinsurance company), which are called hazard models.

In the current study, we consider the spatial distribution of the maximum value of the water depth (calculated over the time duration of a given storm event) as a typical indicator of flooding : this is the output of the considered numerical simulator. However, the numerical models have significant uncertainties, which come from the input data such as offshore meteo-oceanic conditions (waves and sea or ocean water level) or the altitude of the ground numerical model (landforms). Other uncertainties come from the model parameters such as ground roughness. Therefore, we aim at evaluating the influence of these uncertainties on the spatial distribution of the maximum water depth.

To do so, we are interested in non-intrusive methods : the simulator is considered as a black box which can only be evaluated. In this context, several sensitivity analysis techniques have been proposed (see, e.g., [Iooss and Lemaître, 2015, Saltelli et al., 2008]). However, there are two main issues. First, Monte Carlo methods commonly used to estimate sensitivity indices of each input parameter, require a large number of simulator runs

(more than 10^4). Hence, they are hardly applicable directly on the simulator, which typically presents large computation time cost for a single run (of several minutes, even hours). Second, the output is functional : the maximum water depth is a function of the location. In practice, the locations are discretized on a meshgrid, and the output is represented by a high dimensional vector of length equal to the number of pixels. Depending on the processes involved in the flooding (overflow, wave-induced overtopping, coastal defences' breaching), the required level of discretization can be very fine (down to a few meters). This might add difficulty for sensitivity analysis by imposing to manipulate vector of high dimension. For instance, the case study considers maps of size $K = 256 \times 256 = 65\ 536$.

In this context, a standard methodology fixes these two problems in the following way (see e.g. [Marrel et al., 2015, Ma et al., 2019, Li et al., 2020]). First, the output dimension is reduced, most often by principal component analysis (PCA) or by using functional basis decomposition (for instance, Fourier or wavelets basis). This provides a lower dimensional output vector, formed by the largest components (PCA components or basis coordinates). Second, a fast-to-evaluate proxy, called metamodel or surrogate model, is built on that vector. This is usually done by considering independently each coordinate as a scalar output as in [Chen et al., 2011]. Among all metamodels (e.g. linear regression, neural networks, etc.), we select the Gaussian process (GP) regression model [Williams and Rasmussen, 2006], because they provide both an interpolation of the data and an uncertainty at unknown area. Moreover, the method is parameterized by a covariance function (or kernel), which makes it flexible, and allows to exploit expert knowledge.

Unfortunately, PCA treats each coordinate independently and misses the spatial dependence. Furthermore, the output of the simulator exhibits strong local variations : the water depth is not a smooth function of the location. As a result, even with suitable functional bases such as wavelets, a large number of coefficients, typically several hundreds, must be kept to get an accurate approximation. This problem has clearly been highlighted in previous studies, e.g. [Marrel et al., 2015]. This weakens the benefits of reducing dimension.

To tackle this issue, we propose to use functional PCA (FPCA), a common technique in functional data analysis [Ramsay, 2006]. This is equivalent to performing PCA on the coefficients of a functional basis decomposition, with the metric given by the Gram matrix of basis functions. It can be used for popular bases, including Fourier, wavelets, and B-splines. Notice that for non-orthonormal bases such as B-splines, the PCA step uses a different metric than the usual PCA. In addition, we add a preliminary selection step, by choosing the basis terms which are most influential, based on the energy decomposition after the orthonormalization of the functional basis, or according to a penalized regression approach by using directly the original basis. With these two ideas, the method is applicable for large dimensional vectors. For instance, we can deal with maps made of several

tens of thousands of pixels. Furthermore, we cumulate the advantages of PCA and basis decomposition by accounting for spatial dependence of the output, which is ignored by standard PCA, since functions are decomposed in a suitable functional space. Besides, by doing PCA in a second time, dimension reduction is ensured, even when a large number of basis coefficients must be kept : the final number of principal components is small. Finally, as remarked when doing PCA, building a metamodel independently for each coordinate has some sense, since the principal components are uncorrelated (though not necessarily independent).

The use of FPCA for sensitivity analysis has been proposed for instance by [Lamboni et al., 2011]. There, orthonormal basis functions are obtained as eigenfunctions of a Hilbert-Schmidt operator associated to a covariance kernel, by Karhunen-Loève decomposition. In our approach, we define the functional basis first. In theory, the two approaches are equivalent, since a covariance kernel can be built from a predefined basis corresponding to its Karhunen-Loève decomposition. However, in practice, here, there is a clear advantage in defining the functional basis first, which is to deal with non stationarity without expert knowledge. Indeed, contrarily to usual kernels in RKHS which are guided by global regularity assumptions, several functional basis such as wavelet basis have been designed to fit functions with strong local variations.

As other contribution, we give a closed-form expression for Sobol indices by using FPCA metamodels. As explained in the previous paragraph, the formula can be made equivalent to the expression found in [Lamboni et al., 2011] (these indices are named "generalized sensitivity indices") in the case of orthonormal basis functions, when using a kernel constructed from the basis functions. The formula that we obtain is also valid for non-orthonormal popular basis functions such as B-splines.

The whole methodology has been applied on a case of coastal flooding. The study site is "Les Bouchôleurs" (french Atlantic coast, near "La Rochelle" city), which was hit during the Xynthia storm in 2010. Coastal flooding processes are simulated with the BRGM and CCR numerical codes. We consider five scalar inputs which correspond to parameters of tide and storm surge. We are interested in their sensitivity on the model output, which is the spatial distribution of the maximum water depth (calculated over the time duration of a given storm event). The output corresponds to a map with regular discretizations of 256×256 pixels (each pixel being of 25×25 m^2).

Three metamodelling methods have been compared : dimension reduction with standard PCA (GP^{PCA}), and the proposed method based on FPCA by using wavelets ($GP^{FPCA}_{wavelet}$) and B-splines ($GP^{FPCA}_{B-splines}$) basis. Metamodels have been trained using a learning dataset of size 400. 100 test maps have been predicted for both models based on the learning sample. The spatial root mean square error (RMSE) of the three metamodels predictions

is computed. We show the RMSE maps for the BRGM model in Figure 1. We notice that the spatial RMSE of GP^{PCA} is higher than $GP^{FPCA}_{wavelet}$ and $GP^{FPCA}_{B-splines}$ in the urban area (red circle), where there are the highest irregularities.

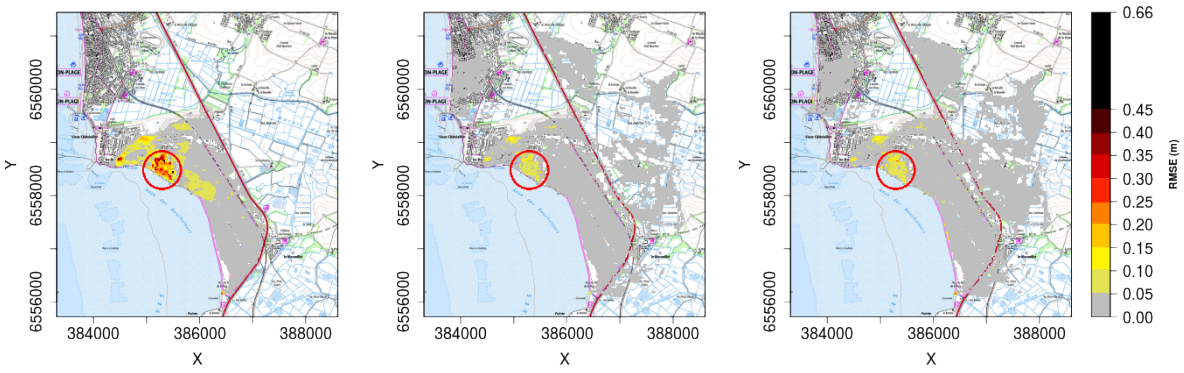


FIGURE 1 – **(BRGM model)** The RMSE maps (using 100 test samples) obtained with GP^{PCA}, $\text{GP}^{FPCA}_{wavelet}$, and $\text{GP}^{FPCA}_{B-spline}$ for the coastal flooding case. In the maps, there are locations without any given values. They correspond to locations where RMSE is strictly less than 1cm, which is negligible. The red circle corresponds to the location of the main urban area. The background layer (SCAN 25® from the National Institute of Geographic and Forest Information IGN) indicates the locations of the urban areas and key topographic elements (roads, railways, marshlands, etc.).

A global sensitivity analysis has been performed for the BRGM and CCR coastal flooding models by replacing the simulators by the metamodel ($GP^{FPCA}_{wavelet}$) trained with 500 simulations. We are interested in the influence of five scalar inputs, which characterize the tide and storm surge. The generalized sensitivity indices have been estimated. The tide level appears to have the highest influence for both models, especially for the CCR simulator. Nevertheless, for the CCR simulator, there is also a small influence of the time duration of the increase and the decrease of the storm surge signal. For the BRGM simulator, there is also an influence of the storm surge peak amplitude and the phase difference between surge peak and high tide.

Acknowledgement The project is funded the BRGM and the CCR. The authors thank Olivier Alata, Professor at Jean Monnet University, in Saint-Etienne, for his expertise on wavelets. The acknowledgments are also addressed to Deborah Idier and Rodrigo Pedreros, from BRGM, for establishing the BRGM model and their opinions on the results, and, David Moncoulon and Pierre Tinard, from CCR, for their expertise in reinsurance.

Références

[Chen et al., 2011] Chen, T., Hadinoto, K., Yan, W., and Ma, Y. (2011). Efficient meta-modelling of complex process simulations with time–space-dependent outputs. *Computers & chemical engineering*, 35(3) :502–509.

[FFSA and Gema, 2011] FFSA and Gema (2011). La tempête Xynthia du 28 février 2010. Bilan chiffré au 31 décembre 2010. Technical report, Association française de l'assurance.

[Iooss and Lemaître, 2015] Iooss, B. and Lemaître, P. (2015). A review on global sensitivity analysis methods. In *Uncertainty management in simulation-optimization of complex systems*, pages 101–122. Springer.

[Lamboni et al., 2011] Lamboni, M., Monod, H., and Makowski, D. (2011). Multivariate sensitivity analysis to measure global contribution of input factors in dynamic models. *Reliability Engineering & System Safety*, 96(4) :450–459.

[Li et al., 2020] Li, M., Wang, R.-Q., and Jia, G. (2020). Efficient dimension reduction and surrogate-based sensitivity analysis for expensive models with high-dimensional outputs. *Reliability Engineering & System Safety*, 195 :106725.

[Ma et al., 2019] Ma, P., Mondal, A., Konomi, B., Hobbs, J., Song, J., and Kang, E. (2019). Computer model emulation with high-dimensional functional output in large-scale observing system uncertainty experiments. *arXiv preprint arXiv :1911.09274*.

[Marrel et al., 2015] Marrel, A., Perot, N., and Mottet, C. (2015). Development of a surrogate model and sensitivity analysis for spatio-temporal numerical simulators. *Stochastic environmental research and risk assessment*, 29(3) :959–974.

[Naulin et al., 2015] Naulin, J., Moncoulon, D., Le Roy, S., Pedreros, R., Idier, D., and Oliveros, C. (2015). Estimation of insurance related losses resulting from coastal flooding in France. *Natural Hazards and Earth System Sciences Discussions*, 3 :2811–2846.

[Perrin et al., 2021] Perrin, T., Roustant, O., Rohmer, J., Alata, O., Naulin, J., Idier, D., Pedreros, R., Moncoulon, D., and Tinard, P. (2021). Functional principal component analysis for global sensitivity analysis of model with spatial output. *Reliability Engineering & System Safety*, 211 :107522.

[Perrin, 2021] Perrin, T. V.-V. E. (2021). *Méta-modélisation et analyse de sensibilité pour les modèles avec sortie spatiale. Application aux modèles de submersion marine.* PhD thesis, Université de Lyon.

[Ramsay, 2006] Ramsay, J. O. (2006). *Functional data analysis.* Wiley Online Library.

[Saltelli et al., 2008] Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., and Tarantola, S. (2008). *Global sensitivity analysis : the primer.* John Wiley & Sons.

[Williams and Rasmussen, 2006] Williams, C. K. I. and Rasmussen, C. E. (2006). *Gaussian processes for machine learning.* MIT Press Cambridge, MA.

Textual Data for Electricity Demand Forecasting

David Obst[1,2], Badih Ghattas[2], Sandra Claudel[1], Jairo Cugliari [3,4], Yannig Goude[5], and Georges Oppenheim[6]

[1]EDF Labs – EDF Recherche et Développement – France
[2]Institut de Mathématiques de Marseille – Aix Marseille Université : UMR7373 – France
[3]Entrepôts, Représentation et Ingénierie des Connaissances (ERIC) – Université Lumière - Lyon 2 : EA3083 – Université Lumière Lyon 2 5 avenue Pierre Mendès-France 69676 Bron Cedex, France
[4]Entrepôts, Représentation et Ingénierie des Connaissances (ERIC) – Université Lumière - Lyon 2 : EA3083, Université Claude Bernard Lyon 1 – Université Lumière Lyon 2 5 avenue Pierre Mendès-France 69676 Bron Cedex, France
[5]EDF Labs – Electricité de France - EDF – France
[6]Paris Saclay – Université de Paris-Sud Orsay – France

Résumé

Les modèles traditionnels de prevision de l'electricite à moyen terme s'appuient sur des informations calendaires et meteorologiques telles que la temperature et la vitesse du vent pour obtenir des performances elevees. Cependant, le fait de dependre de ces variables presente des inconvenients, car elles peuvent ne pas etre suffisamment informatives en cas de conditions meteorologiques extremes, ce qui entraine des erreurs de prevision plus importantes, et les donnees meteorologiques historiques sont souvent disponibles avec retard ou à un prix eleve. Bien qu'omnipresentes, les sources d'information textuelles sont rarement incluses dans les algorithmes de prediction des series temporelles, malgre les informations pertinentes qu'elles peuvent contenir. Dans ce travail, nous proposons d'exploiter les rapports meteorologiques librement accessibles pour les problèmes de prediction de la demande d'electricite et des series temporelles meteorologiques. Nos experiences sur des donnees de charge françaises et britanniques montrent que les sources textuelles considerees permettent d'ameliorer la precision globale du modèle de reference, en particulier par temps froid. De plus, nous appliquons notre approche au problème de l'imputation des valeurs manquantes dans les series temporelles meteorologiques, et nous montrons que notre approche textuelle bat les methodes standard. De plus, l'influence des mots sur les predictions des series temporelles peut etre interpretee pour les schémas d'encodage consideres du texte, ce qui conduit à une plus grande confiance dans nos resultats.

On forecasting bounded variables with latent stochastic bounds

Amandine Pierrot [1] & Pierre Pinson [2]

[1] *Elektrovej, Building 325, 2800 Kongens Lyngby, Denmark, amapi@dtu.dk*
[2] *Akademivej, Building 358, 2800 Kongens Lyngby, Denmark, ppin@dtu.dk*

Résumé. En lien avec l'application à la prévision éolienne, nous nous intéressons à la prévision de variables continues et bornées quand les bornes de l'intervalle varient au cours du temps ou en fonction de variables exogènes sans pouvoir être observées.

Mots-clés. Variables bornées, prévision, prévision éolienne, séries temporelles, variables manquantes, variables latentes, algorithmes MCEM.

Abstract. Out of the application to wind power forecasting, we are interested in forecasting bounded continuous variables when the bounds may vary over time or depending on exogenous variables while not being observed.

Keywords. Bounded variables, forecasting, wind power forecasting, times series, missing variables, latent variables, MCEM algorithms.

1 Introduction

Many forecasting applications involve response variables which are both continuous and bounded. This is especially the case when one wants to forecast rates, percentages or proportions, for example the spread of an epidemic as in Guolo and Varin (2014), the unemployment rates of a given country as in Wallis (1987), or the proportion of time spent by animals in a certain activity, for example in Cotgreave and Clayton (1994). Indeed proportional data are widely made use of in ecology, see Warton and Hui (2011).

In all those applications, the bounds of the interval the response variable lies in are assumed to be fixed to the same values over the sample or throughout the time series. If this assumption makes sense in some cases, it can be misleading and negatively impacts the quality of the forecasts when the bounds actually vary over time or depending on exogenous variables whilst not being observed. In particular it is highly relevant for energy applications, such as wind power forecasting. Indeed wind power generation is a stochastic process that is theoretically double-bounded by zero and the nominal power of the turbine. However in practice the upper bound may also change over time, while being unknown, for example in case of curtailment actions which for the information is not available or not reliable.

2 Preliminaries and motivation

2.1 General framework

We first set the general statistical framework of having a continuous response variable bounded by varying bounds. Let X be a continuous random variable lower bounded by a and upper bounded by b. We want to make a good prediction of X, which we denote by $\hat{X}$. If X takes values in (a, b) then so should $\hat{X}$. We argue that in many forecasting applications, (a, b) might differ for a realisation x_i of X, depending on exogenous variables or over time for stochastic processes, without being observed. Thus we propose to look at the bounds of the continuous response variable X as two missing random variables A and B. The probability distribution of X conditional on A and B is any continuous probability distribution $P(a, b)$ such as

$$\mathbb{P}(X|A = a, B = b) = 0 \text{ if } X \leq a \text{ or } X \geq b.$$

We circumscribe our framework to a parametric framework, that is to probability distributions $P_\theta(a, b)$ of X conditional on A and B which are governed by unknown parameters θ. Let $f(x|a, b; \theta)$ be the density of X, conditional on the bounds (a, b) and on the parameters θ w.r.t. Lebesgue measure as X is a continuous variable. The conditional density $f(x|a, b; \theta)$ can be any probability density function restricted to $(0, 1)$ and rescaled to (a, b) through the function g of the variable $Y \in (0, 1)$ such as $X = g(Y) = (b - a)Y + a$. This includes for instance the beta distribution, truncated distributions and distributions of transformed normal variables introduced in Johnson (1949). In this setup, to predict the value (or distribution) of X comes down to predicting the value (or distribution) of X conditional on (A, B). Let $f(a, b|\beta)$ be the marginal density of the random bounds (A, B) conditional on the parameter β w.r.t Lebesgue measure if the bounds are continuous variables, or counting measure if they are discrete variables. We need to estimate all the unknown parameters $\psi = (\theta, \beta)$, that is to estimate the parameters ψ out of the sample $\mathbf{x} = (x_1, \ldots, x_n)$ without actually observing $(\mathbf{a}, \mathbf{b}) = \big((a_1, b_1), \ldots, (a_n, b_n)\big)$.

In order to predict the value or the distribution of X conditional on (A, B) we thus need distributional assumptions both on $X|(A, B)$ and (A, B). The assumptions we make on the distributions of both the bounded variable and the bounds depend on the forecasting application and the level of complexity in the model that can or should be handled. They do not change the general framework. In this work we will focus on distributions we think are well suited for the application to wind power forecasting.

2.2 Motivating synthetic examples

We show the impact of a fixed bound assumption through a synthetic example built upon the application to wind power forecasting. In this motivating simulation, we focus on

both point and probabilistic forecasting and evaluate the accuracy of the forecasts in the worst and best cases: when wrongly assuming the bound to be fixed to 1 compared to having perfect knowledge of the random bound. We also show the impact of moving from the perfect knowledge of the bound, that is for both the inference and the prediction tasks, to having to forecast the bound before forecasting the target bounded variable. Let X_t be the wind power generation scaled by the nominal power at time t. X_t is a continuous bounded random variable which takes values in $(0, 1)$. Following Pinson (2012) and Pierrot and Pinson (2021), we assume that the distribution of X_t conditional on the previous information set $\mathcal{F}_{t-1}^X$ (the σ-algebra generated by $X_1, \ldots, X_{t-1}$) is a generalized logit-normal distribution, that is the density is

$$f(x_t|\mathcal{F}_{t-1}^X;\theta) = \begin{cases} \frac{1}{\sqrt{2\pi\sigma^2}} \frac{\nu}{x_t(1-x_t^\nu)} \exp\left[-\frac{1}{2}\left(\frac{\gamma(x_t;\nu)-\mu_t}{\sigma}\right)^2\right] & \text{if } 0 < x_t < 1 \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where $\theta = (\Lambda, \sigma^2, \nu)$, with $\Lambda = (\lambda_1, ..., \lambda_p)$ the p parameters entering the auto-regressive process which characterizes the conditional expectation $\mu_t = \sum_{k=1}^p \lambda_k \gamma(x_{t-k}; \nu)$, $\gamma(x; \nu) = \ln\left(\frac{x^\nu}{1-x^\nu}\right)$, $\nu > 0$. In (1) it is assumed that X_t is upper bounded by 1. Moving to the hypothesis that X_t is now upper bounded by a bound B_t which is a random variable and might take values different from 1, the density of X_t conditional on $\mathcal{F}_{t-1}^X$ and on $\mathcal{F}_t^B$ (the σ-algebra generated by $B_1, \ldots, B_t$) becomes

$$f(x_t|\mathcal{F}_{t-1}^X, \mathcal{F}_t^B;\theta) = \begin{cases} \frac{1}{\sqrt{2\pi\sigma^2}} \frac{\nu}{x_t(1-(x_t/b_t)^\nu)} \exp\left[-\frac{1}{2}\left(\frac{\gamma(x_t/b_t;\nu)-\mu_t}{\sigma}\right)^2\right] & \text{if } 0 < x_t < b_t \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where $\mu_t = \sum_{k=1}^p \lambda_k \gamma\left(\frac{x_{t-k}}{b_{t-k}}; \nu\right)$, $x_{t-k} \in (0, b_{t-k})$. Let note that the densities in equations (1) and (2) are typical examples of taking a density whose support is $(0, 1)$ and rescaling it to the interval (a, b), with $a = 0$ here. We then consider different simulation cases corresponding to different assumptions for the probability distribution of the upper bound B_t and show the improvement we can get over the fixed-to-1-bound assumption.

3 Methodology

3.1 Statistical inference

For ease of notation and in adequacy with the wind power forecasting application, without loss of generality, from now we will consider only the upper bound B as being a random bound, that is $a = 0$. Let $\mathbf{x} = (x_i)$ be an i.i.d. sample from $f(x|b;\theta)$ and $\mathbf{b} = (b_i)$ an i.i.d. sample from $f(b|\beta)$. The density $f(\mathbf{x}, \mathbf{b}|\psi)$ is the joint density of the samples $\mathbf{x}$ and $\mathbf{b}$,

$f(\mathbf{x}|\mathbf{b};\theta)$ the density of $\mathbf{x}$ conditional on $\mathbf{b}$ and $f(\mathbf{b}|\beta)$ the marginal density of $\mathbf{b}$. Because we do not have access to the sample $\mathbf{b}$, the likelihood function of the observations $\mathbf{x}$ is

$$L(\psi|\mathbf{x}) = \int f(\mathbf{x},\mathbf{b}|\psi)\mathrm{d}\mathbf{b} = \int f(\mathbf{x}|\mathbf{b};\theta)f(\mathbf{b}|\beta)\mathrm{d}\mathbf{b}. \tag{3}$$

A direct maximization of (3) involves complicated high-dimensional integration and can be computationally infeasible. In order to compute the maximum likelihood estimates of the parameters ψ we thus propose to use an Expectation-Maximization (EM) algorithm, see Dempster, Laird and Rubin (1977), Wu (1983) and Boyles (1983). In the EM setup the data likelihood in equation (3) is the incomplete data likelihood. The complete data likelihood is $L(\psi|\mathbf{x},\mathbf{b}) = f(\mathbf{x},\mathbf{b}|\psi)$ and the complete data log-likelihood is then

$$l(\psi|\mathbf{x},\mathbf{b}) = \ln f(\mathbf{x},\mathbf{b}|\psi) = \ln f(\mathbf{x}|\mathbf{b};\theta) + \ln f(\mathbf{b}|\beta).$$

At each iteration $(i+1)$ the E-step of the EM algorithm computes the conditional expectation

$$Q(\psi|\psi^{(i)}) = \mathbb{E}\left[\ln f(\mathbf{x}|\mathbf{b};\theta)|\mathbf{x};\psi^{(i)}\right] + \mathbb{E}\left[\ln f(\mathbf{b}|\beta)|\mathbf{x};\psi^{(i)}\right],$$

and the M-step maximizes $\psi \longmapsto Q(\psi|\psi^{(i)})$. The EM algorithm is an ascent algorithm for the incomplete data log-likelihood: it can be showed that the algorithm (weakly) increases the log-likelihood in every step. When the bounded variable X and the random bound B are assumed to be following continuous probability distributions, it is generally not possible to write the E-step of the EM algorithm in closed form. We then need to use Monte Carlo integration, that is a Monte Carlo approximation Q_N of Q at the E-step as proposed by Wei and Tanner (1990):

$$Q_N(\psi|\psi^{(i)}) = \frac{1}{N}\sum_{n=1}^{N}\left[\ln f(\mathbf{x}|\mathbf{b}^{(n)};\theta)\right] + \frac{1}{N}\sum_{n=1}^{N}\left[\ln f(\mathbf{b}^{(n)}|\beta)\right],$$

where the Monte Carlo sample $(\mathbf{b}^{(n)})$ is drawn according to the posterior distribution $f(\mathbf{b}|\mathbf{x};\psi^{(i)})$. Because of the introduction of the Monte Carlo error at the E-step, the incomplete data log-likelihood is not guaranteed to increase at every iteration anymore. Thus implementing a Monte Carlo EM (MCEM) algorithm involves additional challenges compared to the usual EM algorithm, not only in terms of achieving convergence and deciding on a stopping rule, but also while choosing the Monte Carlo method of simulation and the size N of the Monte Carlo sample $(\mathbf{b}^{(n)})$.

3.2 Prediction task

Once we have recovered the parameter estimates $\hat{\theta}$ of the distribution of the bounded variable X and the parameter estimates $\hat{\beta}$ of the distribution of the bound B, in order to get the new expected $\hat{x}$ or its full predictive distribution we need to achieve the following steps:

1. We draw a sample (b_i) according to $f(b|\hat{\beta})$ or take the expected value $\hat{b}$.

2. We get $\hat{x}$ or its predictive distribution out of the conditional density $f(x|b_i;\hat{\theta})$ or $f(x|\hat{b};\hat{\theta})$.

When (X_t) is a Markov stochastic process, and we wish to predict the next realisation x_{t+1} which depends on the p former observations of (X_t), a step 0 should come before the step 1, as an estimation of the p former (non-observed) values of (B_t) are needed for computing $f(x_{t+1}|\mathcal{F}_t^X, \mathcal{F}_{t+1}^B; \theta)$ as in (2).

4 Preliminary results and conclusions

This work is driven by our application to wind power forecasting and we show some inference and forecasting results when (X_t) is a Markov stochastic process distributed according to (2) and (B_t) is a stochastic process. It is worth noting that in this particular setup the upper bound can also be considered as an additional parameter of the scaled distribution of the bounded response variable. This parameter would have to be estimated in an online fashion though, as the bound varies over time. This would involve the use of multiple forgetting factors, as the other parameters of the scaled distribution shall not vary likewise.

References

Boyles, R.A (1983). On the convergence of the EM algorithm, *Journal of the Royal Statistical Society: Series B*, 45(1), pp. 47-50.

Booth, J.G. and Hobert, J.P. (1999). Maximizing general linear mixed model likelihoods with an automated Monte Carlo EM algorithm, *Journal of the Royal Statistical Society: Series B*, 61(1), pp. 265-285.

Caffo, B.S, Jank, W. and Jones, G.L. (2005). Ascent-based Monte Carlo Expectation-Maximization, *Journal of the Royal Statistical Society: Series B*, 67(2), pp. 235-251.

Cotgreave, P. and Clayton, D.H. (1994). Comparative analysis of time spent grooming by birds in relation to parasite load, *Behaviour*, 131(3), pp. 171-187.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society: Series B*, 39(1), pp. 1-38.

Dempster, A.P., Rubin, D.B. and Tsutakawa, R.K. (1981). Estimation in covariance components models, *Journal of the American Statistical Association*, 76(374), pp. 341-353.

Doob, J.L. (1935). The limiting distributions of certain statistics, *Annals of Mathematical Statistics*, 6(3), pp. 160-169.

Fisher, R.A. (1925). Theory of statistical estimation, *Mathematical Proceedings of the Cambridge Philosophical Society*, 22(5), pp. 700-725.
Guolo, A. and Varin, C. (2014). Beta regression for time series analysis of bounded data, with application to Canada Google® flu trends, *Annals of Applied Statistics*, 8(1), pp. 74-88.
Johnson, N.L. (1949). Systems of Frequency Curves Generated by Methods of Translation, *Biometrika*, 36(1), pp. 149-176.
Levine, R.A. and Casella, G. (2001). Implementations of the Monte Carlo EM algorithm, *Journal of Computational and Graphical Statistics*, 10(3), pp. 422-439.
Louis, T.A. (1982). Finding the observed information matrix when using the EM algorithm, *Journal of the Royal Statistical Society: Series B*, 44(2), pp. 226-233.
McCulloch, C.E. (1994). Maximum likelihood variance components estimation for binary data, *Journal of the American Statistical Associations*, 89(425), pp. 330-335.
McCulloch, C.E. (1997). Maximum likelihood algorithms for generalized linear mixed models, *Journal of the American Statistical Association*, 92(437), pp. 162-170.
Meilijson, I. (1989). A fast improvement to the EM algorithm on its own terms, *Journal of the Royal Statistical Society: Series B*, 51(1), pp. 127-138.
Pierrot, A. and Pinson, P. (2021). Adaptive generalized logit-normal distributions for wind power short-term forecasting, *Proceedings of the IEEE Powertech 2021*, Madrid, Spain.
Pinson, P. (2012). Very-short-term probabilistic forecasting of wind power with generalized logit-normal distributions, *Journal of the Royal Statistical Society: Series C*, 61(4), pp. 555–576.
Redner, R.A and Walker, H.F. (1984). Mixture densities, maximum likelihood and the EM algorithm, *SIAM Review*, 26(2), pp. 195–239.
Robert, C.P. and Casella, G. (2004). *Monte Carlo statistical models*, Springer, New York.
Wallis, K.F. (1987). Time series analysis of bounded economic variables, *Journal of Time Series Analysis*, 8(1), pp. 115–123.
Warton, D.I. and Hui, F.K.C (2011). The arcsine is asinine: the analysis of proportions in ecology, *Ecology*, 92(1), pp. 3–10.
Wei, G.C.G. and Tanner, M.A. (1990). A Monte Carlo implementation of the EM algorithm and the Poor Man's data augmentation algorithms, *Journal of the American Statistical Association*, 85(411), pp. 699-704.
Wu, C.F.J. (1983). On the convergence properties of the EM algorithm, *Annals of Statistics*, 11(1), pp. 95-103.

SFdS - Sport

L'ORIGINE DE LA DÉFICIENCE VISUELLE EST-ELLE UN FACTEUR DE PERFORMANCE ? ANALYSE DES PARA-NAGEURS ET PARA-ATHLÈTES DE NIVEAU INTERNATIONAL

Julien Schipman [1], Bryan Le Toquin [2], Quentin De Larochelambert [3], Guillaume Saulière [4], Stephanie Duncombe [5] & Jean-François Toussaint [6]

[1,2,3,4,5,6] Irmes/ Insep. *11 avenue du Tremblay 75012 Paris.*
julien.schipman@insep.fr; bryan.letoquin@insep.fr; quentin.delarochelambert@insep.fr; guillaume.sauliere@insep.fr; s.duncombe@uq.net.au; jean-francois.toussaint@aphp.fr;

Résumé. L'objectif de cette étude était d'analyser l'effet des déficiences visuelles congénitales et acquises sur les performances internationales en natation et athlétisme Paralympiques issues de l'ensemble des compétitions entre 2009 et 2019, soit 20 689 performances récoltées. L'origine de la déficience a été recueillie sur le site du Comité International Paralympique (CIP) et catégorisée en deux groupes (déficience congénitale et acquise). Dans les classes sportives des déficients visuels (11-12-13), le niveau de performance et la relation âge/performance ont été étudiés en fonction de l'origine de la déficience. Dans les classes 11 et 12, les performances maximales ont été atteintes plus tôt par les nageurs et nageuses atteints d'une déficience congénitale que par ceux et celles atteints d'une déficience acquise ($p < 0.05$). Aucune différence n'était présente dans la classe 13 ni dans aucune classe en para-athlétisme ($p > 0.05$). Un niveau de performance similaire a été observé entre les deux disciplines sportives pour chaque classe ($p > 0.05$). Cette étude a démontré que l'origine de la déficience peut influencer le parcours de performance chez les nageurs déficients visuels.

Mots-clés. Natation ; Athlétisme ; Paralympique ; Déficience ; Handicap ; Détection ; Para sport ;

Abstract. The aim of this study was to investigate the effect of congenital and acquired visual impairments on the international performance of Para swimmers and Para track and field athletes. We collected results from visually impaired Para athletes competing in Para swimming or Para athletic events at all IPC-labelled competitions between 2009 and 2019. The dataset contained 20,689 events results. Impairment origin was collected from the International Paralympic Committee (IPC) website. We separated impairment origin into two groups to distinguish those with a congenital impairment from those with an acquired impairment. In visual impairment sport classes (11–12-13), the performance level and the age performance relationship were investigated according to the impairment origin. In classes 11 and 12, peak performance was achieved earlier by male and female swimmers with a congenital impairment compared with those who had an acquired impairment ($p < 0.05$). No differences were present in class 13 or in any class in Para athletics ($p > 0.05$). A similar performance level was observed among the two sport disciplines for each class ($p > 0.05$). This study demonstrated that impairment origin can influence the performance pathway among visually impaired swimmers.

Keywords. Swimming ; Athletics ; Paralympiques ; Impairment origin ; Talent Identification ; Para sport ;

1. Introduction

One of the challenges that Para sport faces today is the achievement of fair competition. This can be done by minimizing the impact of an individual's impairment on the outcome so that sporting ability, skill level and training alone determine success and the final result. Historically, the classification system was based on a medical diagnosis for all Paralympic disciplines and sport classes. In the 1980s, a sport-specific and evidence-based functional classification system was implemented. However, because of the lack of guidance and scientific evidence of visual impairment (VI) functional impact, the VI classification system is still based on a medical diagnosis. In order to assess the needs for a sport-specific and evidence-based VI classification system, two studies, based on a systematic interactive survey method address several key points for Para swimming and Para athletics classification systems. In both studies, expert panels unanimously agreed that the age at which a VI is acquired will impact the swimming and running performance and the technical skills acquisitions (Allen et al., 2020; H. J. C. Ravensbergen et al., 2018). To our knowledge, no studies have been conducted comparing quantitatively the performance level between Paralympic athletes with a congenital impairment (CI) and an acquired impairment (AI). At an individual scale, the age at which the impairment is acquired could also have a major influence on this pathway. Therefore, it seems necessary to include the impairment origin in the age-related performance analysis in order to provide information on how it can influence a career pathway.

The objective of this study was to investigate the international performances of Para Swimmer (PS) and Para Athletes (PA) with an AI or a CI. To meet this objective, we first compared the performance level in each impairment class between the CI and AI groups. Then we investigated the relationship between age and performance in each impairment class according to the VI origin for the same events.

2. Methods

Publicly available data was extracted from the IPC website: www.paralympic.org. We extracted each athletes' top annual performance for each event between 2009 and 2019. Before 2009, the data were unavailable on the IPC website. In Para swimming, athletes' times were collected for the following long course events: 50 m freestyle, 100 m freestyle, 400 m freestyle, 100 m butterfly, 100 m backstroke, 100 m breaststroke, 200 m individual medley. For Para Athletics, times were collected for the following events: 100 m, 200 m (only females), 400 m, 1500 m, 5000 m (only males), marathon, and long jump. These events were selected because they were in the Tokyo 2021 Paralympic Games program. The following variables were included in the dataset: event type, event time converted to speed ($m.s^{-1}$), year of the event, last and first name, age (in years), sex, and impairment classification. A categorical variable "VI origin" with three factors (congenital impairment, acquired impairment, missing information) was added to the dataset. We defined performance level as the mean performance for the CI group and AI group in each Para swimming and Para athletics event for each class. We defined performance as the best event time converted into speed (m. s−1) for each Para swimmer with a congenital impairment (PSCI) and Para swimmer with acquired impairment (PSAI) in each event for each class. Similarly, the best event time for each Para track and field athlete with congenital impairment (PACI) and Para athlete with acquired impairment (PAAI) in each event by classes was defined as the athlete's performance.

2.1 Relative performance

We analyzed the best individual performance of each PS and PA according to VI origin in class 11, 12 and 13 for each event. To do so, we calculated the Relative Performance (RP) for each athlete *(i)* in percentage of the all-time best performance for each event *(j)* and for each class *(k)* (Equation (1)). This method allowed us to compare intra-class performance by grouping all events.

$$Eqn\,1 : RP_i(\%) = \frac{World\ record_{jk}}{Best\ individual\ performance_{jk}} * 100$$

2.2 Performance level in each event

In order to go further in the investigation, we analyzed the best individual performance of each PS and PA in each class for the 21 events in each discipline. The performance was calculated in $m.s^{-1}$ and meters (m) for long jump event.

2.3 Age related performance according to the VI origin

In order to investigate the relationship between age and performance, we used the best performance of each PS and PA by age and according to VI origin in each event and for each class.

We modelled the relationship between age and performance according to an Integrative Model of Age-Performance (IMAP1). This model accurately describes the age-related changes of performance in multiple time-series phenotypic traits. IMAP1 (Equation (2)) is based on an enhanced biological foundation (Berthelot, Bar-Hen et al., 2019). The formula of IMAP1 is:

$$Eqn\,2 : P(t) = a \cdot e^{\frac{b}{c}\left(1-e^{-\alpha}\right)} \cdot \left(1-e^{d(t-f)}\right)$$

P represents the performance according to *t* (the time), *a* and *b* are two growth parameters, *c, d* and *f* are decay parameters. In order to find optimal parameters, we used the method of least squares to employ a method of parameter optimization under constraint. Parameters *a, b* and *c* were forced to be higher than 0 and parameter *f* was forced to be between 90 and 120. For each event, we calculated the estimated age of peak performance (EAPP) that corresponds to the maximal estimated performance by age. The quality of each fit was estimated with the coefficient of determination, R^2.

2.4 Statistical analysis

We conducted statistical tests to compare the performance indicators within each sample and therefore, test the effect of the observed factor: the VI origin.

To test for performance difference according to each impairment origin group, a Kruskal Wallis test with a pairwise Wilcoxon test with Bonferroni adjustment was performed in order to control the multiple tests effects and reduce the probability of type I errors. Data were reported as mean ± standard deviation of RP. To test the age performance relationship difference according to impairment origin group we computed and compared mean EAPP for all events with a Wilcoxon test. Data were reported as mean ± standard deviation of EAPP.

For all statistical analysis, the significance level was set at $\alpha = 0.05$. Data processing, graphical visualization and statistical analyses were conducted in R (version 4.0.3; The R Foundation for Statistical Computing, Vienna, Austria).

3. Results

3.1 Relative performance

The Para swimming dataset contained 10,619 event times: 6380 events from 408 males; 4239

events from 227 females. The Para athletics dataset contained 10,942 event times: 7658 events from 1031 males; 3284 events from 447 females

The relative performance analysis showed a statistically significant difference between CI, AI and MI groups ($p < 0.05$) in each sport class in para swimming and para athletics. The RP were lower in each sport class for the MI group for men and women in Para swimming and Para athletics (Figure 1). Post-hoc tests showed statistically significant differences for MI group with the CI and AI groups ($p < 0.05$). Between CI and AI groups, the post-hoc tests showed no statistical differences for men and women in Para swimming and Para athletics ($p > 0.05$).

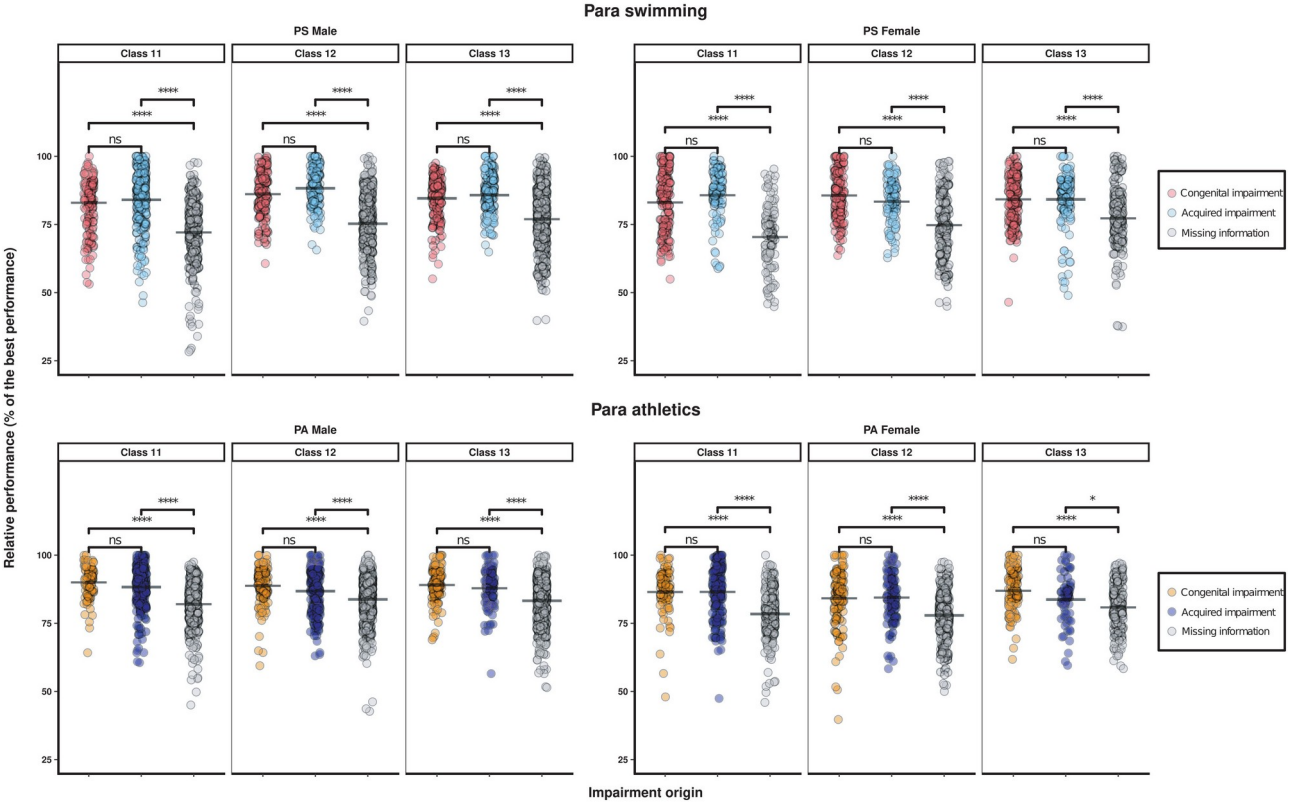


Figure 1: Relative performance (in percentage (%) of the best performance for each event) in congenital impairment, acquired impairment and missing information groups for each sport class in para swimming and para athletics events. *: Indicate a significant mean relative performance difference ($p < 0.05$) with the two others groups****: Indicate a significant mean relative performance difference ($p < 0.001$) with the two others groups

3.2 Performance level

Para swimming

The performance level analysis for men showed no statistical differences between PSCI and PSAI in mean performance among all sport classes in each event (Figure 2). The performance level analysis for women showed no statistical differences between PSCI and PSAI in mean performance in 20 of 21 events (Figure 2). A statistically significant difference was found for class 13 in the 400 m freestyle. In this event, the performance level was significantly higher for PSAI ($p < 0.04$). Mean speed was 1.29 ± 0.11 $m.s^{-1}$ for PSCI and 1.36 ± 0.07 $m.s^{-1}$ for PSAI.

Para athletics

The performance-level analysis for men showed no statistical differences between PACI and PAAI in mean performance in 19 of 21 events (Figure 2). A statistically significant difference was found for class 12 in the 5000 m. In this event, the performance level was significantly higher for PACI ($p < 0.03$). Mean speed was 5.40 ± 0.30 $m.s^{-1}$ for PACI and 5.03 ± 0.51 $m.s^{-1}$ for PSAI. A statistically

significant difference was also found for class 12 in the marathon. In this event, the performance level was significantly higher for PACI ($p < 0.04$). Mean speed was 4.65 ± 0.27 m.s^{-1} for PACI and 4.32 ± 0.037 m.s^{-1} for PAAI.

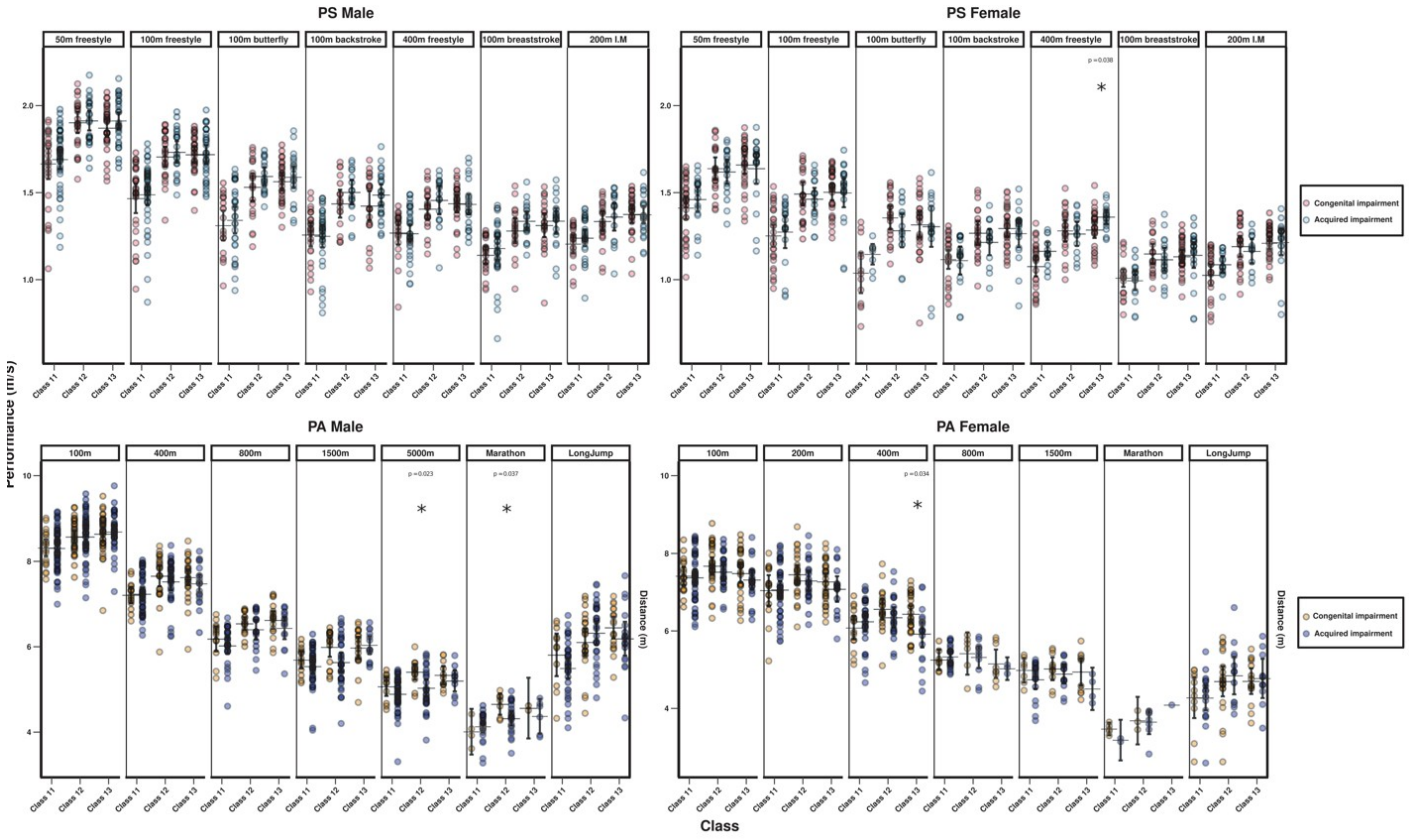


Figure 2: Performance level (mean ± sd) in each event for each sport class in para swimming and para athletics. Performance is expressed as speed (m.s^{-1}) or distance (m) for long jump. *: Indicate a significant mean performance difference between congenital impairment and acquired impairment ($p < 0.05$).

The performance level analysis for women showed no statistical differences between PACI and PAAI in mean performance in 20 of 21 events (Figure 2). A statistically significant difference was found for class 13 in the 400 m. In this event, the performance level was significantly higher for PACI ($p < 0.04$). Mean speed was 6.43 ± 0.54 m.s^{-1} for PACI and 5.92 ± 0.72 m.s^{-1} for PAAI.

3.3 Age-related performance

In each event, the age-performance relationship displayed a similar pattern between the CI and AI groups in Para swimming and Para athletics. Performance increased progressively up to a peak and then decreased gradually. Based on the event, the EAPP was different between impairment classes and VI origin.

Para swimming

The age-performance relationship analysis for females showed a statistically significant difference in mean EAPP for class 11 and 12 between PSCI and PSAI (Figure 3). For class 11, the mean EAPP were 18.08 ± 0.50 years for PSCI and 23.05 ± 1.36 years for PSAI ($p < 0.01$). For class 12, the mean EAPP was 18.19 ± 0.52 years for PSCI and 21.20 ± 3.11 years for PSAI ($p < 0.03$). The age-performance relationship analysis for males showed a statistically significant difference in mean EAPP for class 11 and 12 between PSCI and PSAI (Figure 3). For class 11, the mean EAPP was 20.51 ± 2.45 years for PSCI and 23.77 ± 1.46 years for PSAI ($p < 0.03$). For class 12, the mean

EAPP was 22.62 ± 2.80 years for PSCI and 26.04 ± 1.57 years for PSAI ($p < 0.05$). In class 13 for both genders, the age-performance relationship analysis showed no statistical differences between in mean EAPP for PSCI and PSAI ($p > 0.05$).

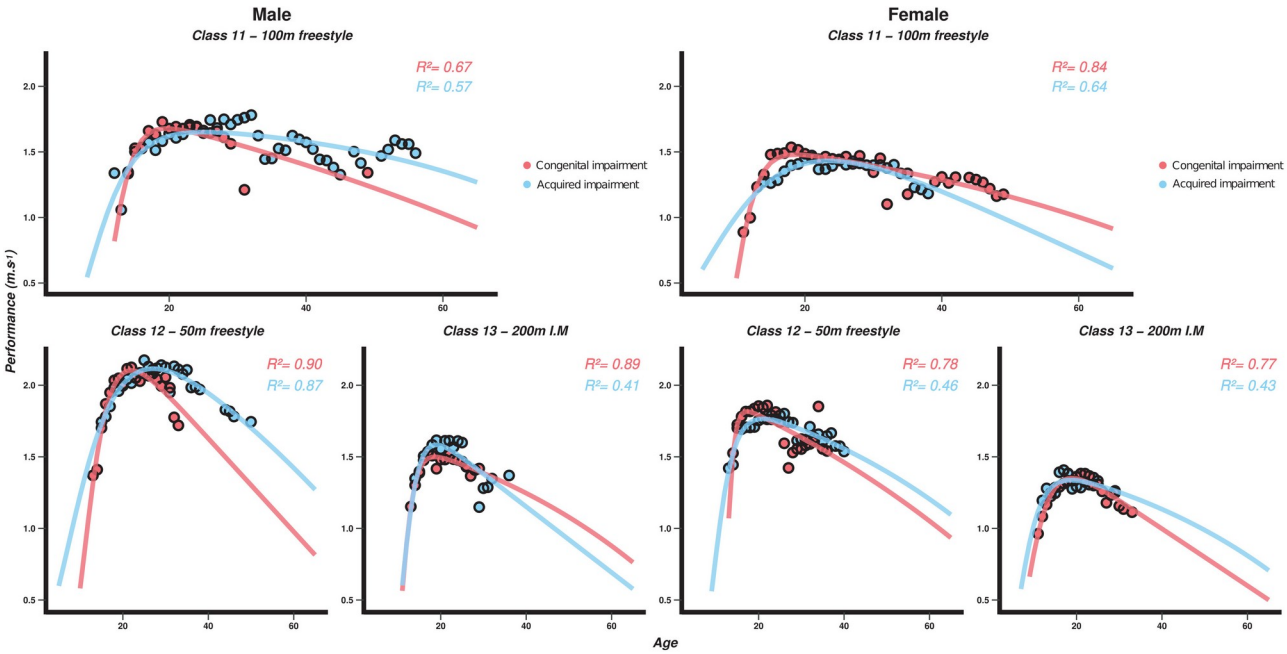


Figure 3: Age related performance modelling in para swimming by visual impairment origin (Congenital impairment and acquired impairment) in 100 m freestyle – class 11, 50 m freestyle – class 12 and 200 m individual medley (I.M) – class 13.

In para-athletics, the age-performance relationship analysis for males and females showed no statistical differences ($p > 0.05$) in mean EAPP for PACI and PAAI in each sport class

4. Conclusion

This study is the first to provide an initial investigation into the relationship between the origin of the impairment and performance in Paralympic disciplines. Based on international performances data since 2009, it aimed to determine if the origin of VI was a performance factor in Para swimming and Para athletics. The main findings showed that: 1) Performance level is similar among all impairment classes in both disciplines (Figures 1–2). 2) Age at peak performance is reached at any earlier age for male and female PSCI in class 11 and 12 in almost all events (Figure 3). The findings of this study through the quantitative analysis of performance in visual impaired PS and PA provide an initial response to the expert panel's queries in both disciplines (Allen et al., 2020; H. J. C. Ravensbergen et al., 2018). These results could allow a better individualization of athlete support and provide additional information about each swimmer's potential. Further studies should be conducted to corroborate these observations and questions the necessity to add the impairment origin in the classification system.

5. References

-Allen et al. (2020), Evidence-based Classification in Track Athletics for Athletes with a Vision Impairment: A Delphi Study, Optometry and Vision Science: Official Publication of the American Academy of Optometry, 97(11):984-94.
-H. J. C. Ravensbergen et al. (2018), Expert consensus statement to guide the evidence-based classification of Paralympic athletes with vision impairment: a Delphi study, British Journal of

Sports Medicine. 50(7):386-91.
-Berthelot et al. (2019), An integrative modeling approach to the age-performance relationship in mammals at the cellular scale. Science Report,9(1):418.

RELATIVE AGE EFFECT AMONG FRENCH SWIMMERS

Audrey Difernand[1], Quentin De Larochelambert[1], Robin Pla[1], Juliana Antero[1], Jean-François Toussaint[1] & Adrien Sedeaud[1]

[1] *Institut de Recherche bio-Médicale et d'Epidémiologie du Sport (IRMES), Institut National du Sport, de l'Expertise et de la Performance (INSEP), 11 avenue du Tremblay, 75012 Paris*
audrey.difernand@insep.fr; qdelarochelambert@ffs.fr; robin.pla@ffnatation.fr; juliana.antero@insep.fr; jean-francois.toussaint@aphp.fr; adrien.sedeaud@insep.fr

Résumé. L'objectif de cette étude était de révéler la présence de l'effet de l'âge relatif parmi les nageurs Français et de proposer une méthode de rééquilibrage afin de mieux apprécier le potentiel de l'athlète en fonction de sa catégorie et de sa discipline. 62 610 nageurs entre 10 et 16 ans sur la discipline du 100m nage libre en bassin de 50m sont considérés pour cette étude. Parmi eux, moins d'un nageur sur cinq entre 13 et 16 ans est né dans le dernier trimestre de l'année. Pour éviter l'abandon ou la perte de vue des nageurs, nous avons mis en place une méthode de rééquilibrage basée sur la performance du nageur, son âge exact au moment de la compétition et le coefficient de régression entre la performance et l'âge au sein de la catégorie considérée. Après application de la méthode, nous avons remarqué aucune différence significative entre les performances rééquilibrées et les performances réalisées a posteriori excepté pour la catégorie des 13 ans.

Mots-clés. Age relatif, natation, détection, méthodes de rééquilibrage

Abstract. The aim of this study was to highlight the presence of the relative age effect among French swimmers and to propose a method of rebalancing in order to better appreciate the potential of each athlete according to his category and discipline. 62 610 males' swimmers between the ages of 10 and 16 in the 100m freestyle in a 50m pool are considered for this study. Less than one in five swimmers aged 13-16 was born in the last quarter of the year. We implemented a rebalancing method based on the swimmer's performance, his exact age at the time of the competition and the regression coefficient between performance and age within the category considered. After applying the method, we found no significant differences between the rebalanced performances and the performances achieved except for the 13 years old category.

Keywords. Relative age, swimming, talent identification, rebalancing method

1. Introduction

The issue of talent identification in sport is very complex. Indeed, the objective is to identify, detect and choose the future Olympic champions 10 years in advance. Age is one of the key factors in this Olympic quest. The younger the athlete is identified, the more time he or she will have to build up in high performance structures. Relative age is defined by the exact age as opposed to the integer age which does not consider the day and month of birth (Helsen, Starkes, et Van Winckel 1998). Thus, within the same age category, there can be up to 365 days difference between two athletes. This age difference can also be translated into differences in height, weight, strength, experience and agility. For this reason, older athletes within the same age group may be favored in selections compared to younger ones: this is called relative age effect (RAE) (Musch et Grondin 2001). A study highlights the over-representation of British rugby players with 73% born in the first half of the year compared to 27% in the second half (McCarthy et Collins 2014). The RAE is also present in individual sports such as swimming (Abbott et al. 2020). After having shown the presence of RAE, they aim to find a method to remove this effect (Cobley et al. 2019). In our study, the objective was not only to show that a relative age effect exists among French swimmers using the example of male 100m freestyle, but also to test a method that allows us to reveal the potential of a swimmer with respect to his age, performance and age category.

2. Methods

All performances from local to international level were collected by the French Swimming Federation between 2000 and 2020. In this study, we were interested in males' 100m freestyle event in the Olympic pool. The database was composed of the date of birth, the date of performance and the performance in seconds of each swimmer.
In the first part, we kept only the season best performance for each swimmer. To better visualize the differences between swimmers, we defined the relative age gap. It was the time between the last birthday and the competition date. Swimmers were divided into birth quarters (Q1: January, February, March, Q2: April, May, June, Q3: July, August, September, Q4: October, November, December). To test for a significant difference between the birth quarters, chi-square test, odds ratios and 95% confidence intervals were calculated.
Within each age category, we related the relative age gap to performance in order to reveal the nature of the relationship between these two variables. Then, using the regression coefficient, we calculated the rebalanced performance:

$$T_r = T_0 - c * m$$

where:
T_r: rebalanced performance time
T_0: initial performance time
c: slope coefficient
m: time until the next birthday

In the second part, in order to validate our rebalancing method, we used a database composed of swimmers having achieved 2 performances in the season with a minimum of 6 months difference. The rebalancing method is applied to the first performance considering the difference with the second performance.

$$T_r = T_1 - c * (T_2 - T_1)$$

where:
T_r: rebalanced performance time
T_1: initial performance time
T_2: second performance time
c: slope coefficient

Thus, it has been possible to compare the rebalanced performance with the (second) performance achieved.

3. Results

The dataset consisted of 62 610 swimmers between the ages of 10 and 16 who have competed at least once in the last 20 years. The distribution in birth quarters is presented in table 1. At 15, for example, we were twice as likely to find a swimmer born in the first quarter as in the last quarter (OR = 2.13, $p<0.05$). Indeed, 32.2% of them were born in the first quarter compared to only 18.2% in the last quarter. Also, we noted that among the top 10% swimmers, the proportion born in the first quarter increases with age in contrast to the proportion born in the last quarter. Among the 14 years old, 19.1% of swimmers were born in the last quarter for all swimmers combined, compared to 16.1% in the top 10%.

Age	Top %	Total	Q1 (%)	Q2 (%)	Q3 (%)	Q4 (%)	χ^2	p-value	OR	p	95% CI
10	100	3767	897 (23.81)	871 (23.12)	1087 (28.86)	912 (24.21)	30.78	***	0.98	**	[0.97 ; 0.99]
11	100	6982	1706 (24.43)	1738 (24.89)	1941 (27.8)	1597 (22.87)	35.46	***	1.09	**	[1.06 ; 1.12]
12	100	9807	2661 (27.13)	2498 (25.47)	2554 (26.04)	2094 (21.35)	75.20	***	1.37	**	[1.28 ; 1.47]
13	100	12787	3655 (28.58)	3426 (26.79)	3152 (24.65)	2554 (19.97)	211.99	***	1.60	**	[1.5 ; 1.7]
14	100	15140	4559 (30.11)	4147 (27.39)	3540 (23.38)	2894 (19.11)	418.50	***	1.82	**	[1.72 ; 1.93]
15	100	14127	4547 (32.19)	3976 (28.14)	3034 (21.48)	2570 (18.19)	679.78	***	2.13	**	[2.01 ; 2.25]
10	10	376	48 (12.77)	92 (24.47)	134 (35.64)	102 (27.13)	40.26	***	0.39	**	[0.29 ; 0.52]
11	10	698	91 (13.04)	179 (25.64)	267 (38.25)	161 (23.07)	90.15	***	0.50	**	[0.43 ; 0.58]
12	10	980	156 (15.92)	262 (26.73)	365 (37.24)	197 (20.1)	101.69	***	0.75	**	[0.71 ; 0.79]
13	10	1278	281 (21.99)	386 (30.2)	381 (29.81)	230 (18.0)	55.39	***	1.28	**	[1.2 ; 1.37]
14	10	1514	428 (28.27)	450 (29.72)	392 (25.89)	244 (16.12)	68.26	***	2.05	**	[1.73 ; 2.43]
15	10	1412	426 (30.17)	416 (29.46)	342 (24.22)	228 (16.15)	70.95	***	2.24	**	[1.86 ; 2.7]

Table 1: *Distribution by birth quarters with respect to age groups and level of competitiveness among male's 100m Freestyle in 50m pool (top %: top percentage, Total: number of observations, Q1-4: birth quarters 1 to 4, χ^2: chi-square adjustment value, ***: $p<0.01$, **: $p<0.05$, OR: odds-ratios between Q1 and Q4, 95% CI: 95% Confidence Interval.)*

Linear regressions were run in each age category (*Figure 1*). The relationship between the relative age gap and the average performance was negative and the absolute value of the coefficient decreased with age. At 12, on average, one month more of relative age gap gained 0.826 seconds, which over a year gave us more than 9 seconds.

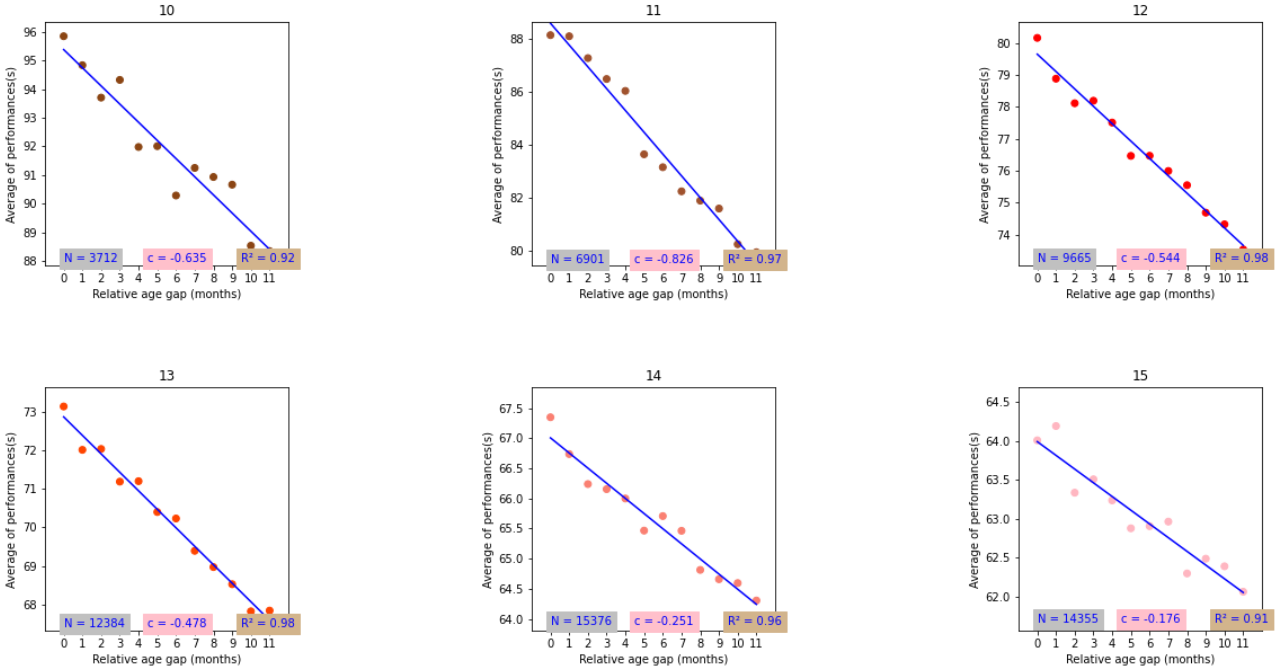


***Figure 1**: Average of performances (in seconds) by relative age gap (in months) among male's 100m Freestyle in 50m pool (N: number of observations, c: slope regression coefficient, R²: coefficient of determination)*

The rebalancing method is illustrated in *figure 2*. We observed that the best performances before rebalancing were mostly achieved by relatively old swimmers within the age category considered. After rebalancing, the top 10 best performances was renewed and new potentials emerged. Indeed, 132, 118, 187, 65, 63 and 42 swimmers had rebalanced performances that were better than the threshold of the pre-rebalancing top 10 best performances. The rate of stability on the other hand tended to be higher with age.

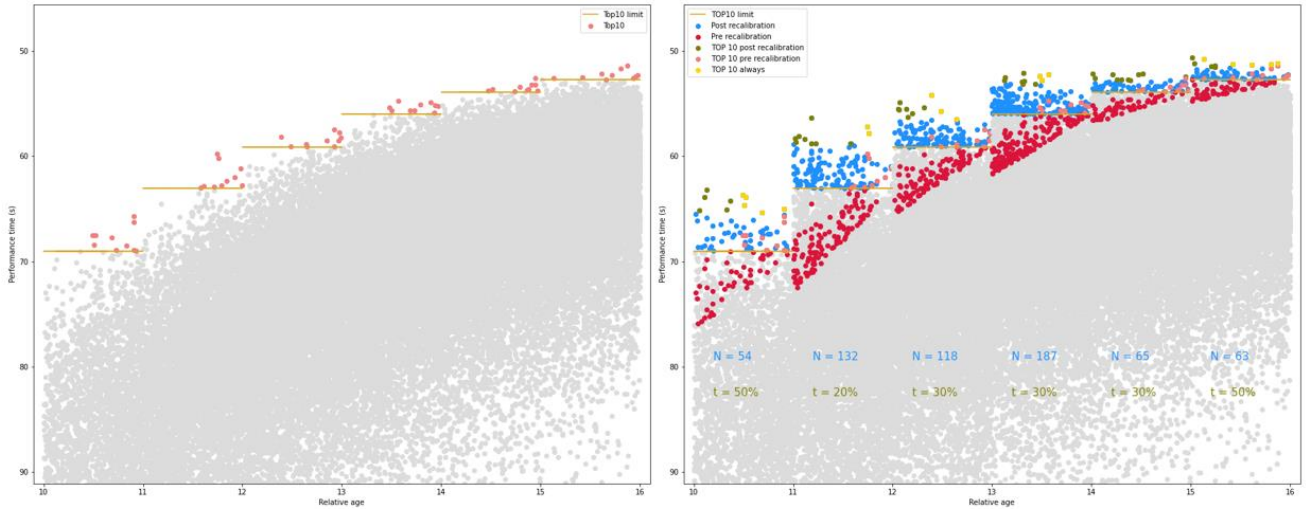


***Figure 2**: On the left, pre-rebalanced performances between 10 and 16 years old among male's 100m Freestyle in 50m pool. Horizontal bars (gold) mark the top10 best performances (pink) threshold. On the right, post-rebalanced performances are illustrated. Performances from the old and new top 10 best performances are shown in gold, those from the new top 10 best performances after rebalancing are in green and those from the old top 10 best performances before rebalancing are still in pink. Performances above the old top10 threshold are shown in blue and the red dots represent their original performances before rebalancing.*

Since the sample does not follow the normal distribution hypothesis, a non-parametric Friedman test is performed to detect a significant difference between the three types of performance. We assume the null hypothesis that the three types of performance follow the same distribution. We have for each age category a p-value smaller than 0.01, which allows us to reject the null hypothesis. Therefore, a Wilcoxon test is performed for each pair of possible performances to reveal which performance type is significantly different from the other two. *Figure 3* shows the comparisons between the original, second and rebalanced performances. Indeed, we found no significant differences on average between the actual second performances and the rebalanced performances except for the 12- and 15-years old category ($p<0.05$). Indeed, for 10, 11, 13 and 14 years-old categories, the average of rebalanced performances was 82.46, 77.25, 65.01 and 61.79 seconds and the average of realized second performances was 84.09, 77.51, 64.01 and 61.86 seconds respectively.

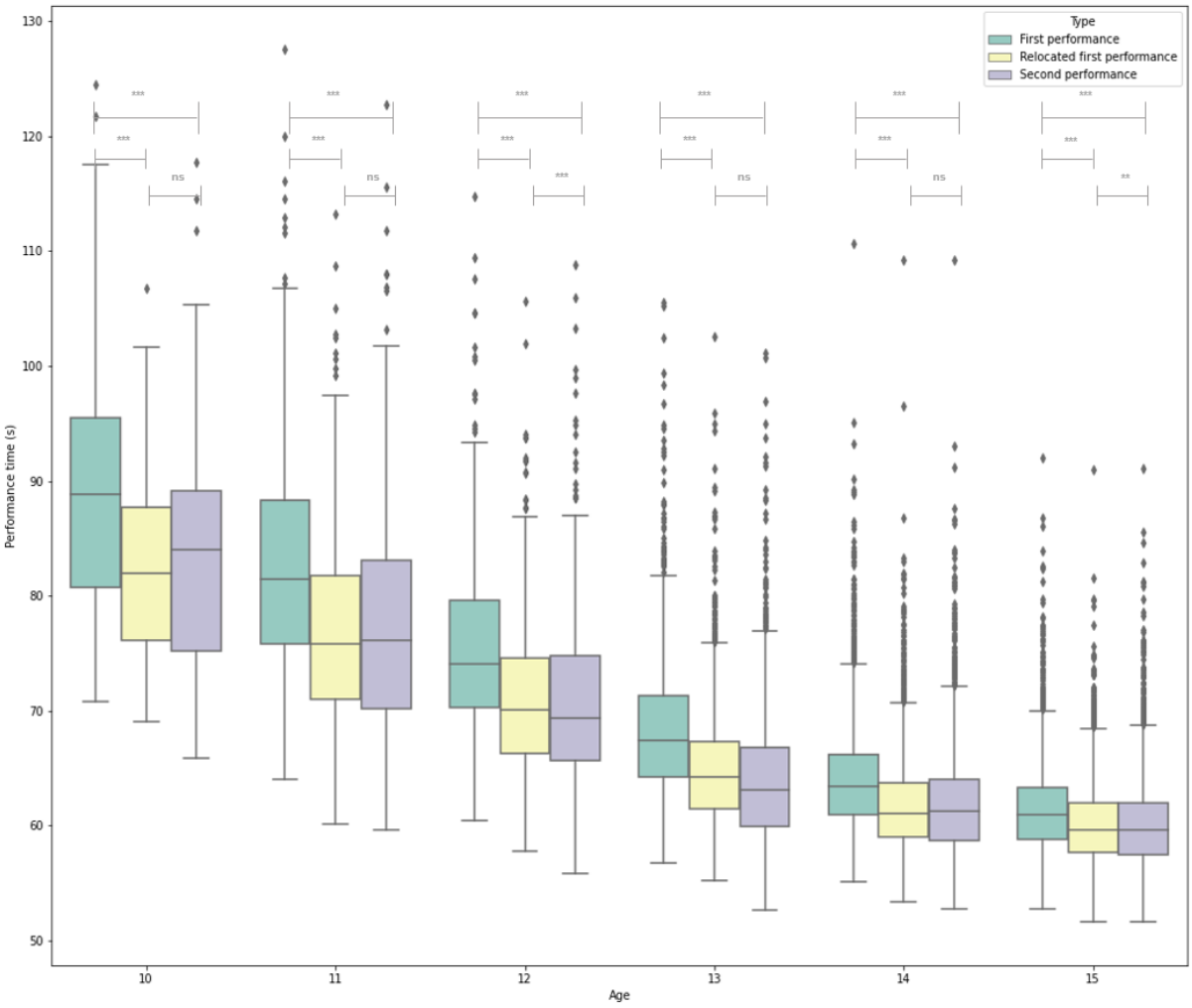


***Figure 3**: Boxplots of first, rebalanced and second performances by age between 10 and 15 years old among male's 100m Freestyle in 50m pool.*

4. Conclusion

This study is the first to demonstrate the relative age effect among French swimmers, particularly among males' 100m Freestyle in 50m pools. It is worth noticing that this effect is more important as the level of competitiveness increases. Also, thanks to the rebalancing method, it is possible to objectify swimmers' performances in relation to their exact age and thus better appreciate their potential. As performance is multifactorial and complex to predict, further studies should be considered, including other variables as maturity for example.

Bibliographie

Abbott, Shaun et al. 2020. « Testing the Application of Corrective Adjustment Procedures for Removal of Relative Age Effects in Female Youth Swimming ». *Journal of Sports Sciences* 38(10): 1077-84.

Cobley, Stephen et al. 2019. « Removing Relative Age Effects from Youth Swimming: The Development and Testing of Corrective Adjustment Procedures ». *Journal of Science and Medicine in Sport* 22(6): 735-40.

Helsen, Werner F., Janet L. Starkes, et Jan Van Winckel. 1998. « The Influence of Relative Age on Success and Dropout in Male Soccer Players ». *American Journal of Human Biology: The Official Journal of the Human Biology Council* 10(6): 791-98.

McCarthy, Neil, et Dave Collins. 2014. « Initial Identification & Selection Bias versus the Eventual Confirmation of Talent: Evidence for the Benefits of a Rocky Road? » *Journal of Sports Sciences* 32(17): 1604-10.

Musch, Jochen, et Simon Grondin. 2001. « Unequal Competition as an Impediment to Personal Development: A Review of the Relative Age Effect in Sport ». *Developmental Review* 21(2): 147-67.

Parametric and non-parametric modeling by a Makovian multi-state model of the evolution of performance in French alpine skiing

De Larochelambert Quentin[1,2,3,4,*] Difernand Audrey[2,4] Antero Juliana[2,4]
Sedeaud Adrien[2,4] Toussaint Jean-François[2,5,4] Coulmy Nicolas[3]
Louis Pierre-Yves[6,1]

2022-05-04

Abstract

We estimate the level of future performance in alpine skiing by modeling the probabilities of transition between performance levels categorized into performance lanes, through a time-inhomogeneous Markov chain as well as nonparametric and parametric multi-state models. We show that the chances of reaching the best level of performance in adulthood are very slightly higher for an individual in the first level of performance at age 12 than an individual in the last level of performance (5% difference). This difference increases with age (16% at 15 years).

Nous estimons le niveau de performance future en ski alpin en modélisant les probabilités de transition entre les niveaux de performance catégorisés en couloirs de performance, à travers une chaîne de Markov inhomogène dans le temps ainsi que des modèles multi-états non paramétriques et paramétriques. Nous montrons que les chances d'atteindre le meilleur niveau de performance à l'âge adulte est très faiblement supperieur pour un individu dans le premier niveau de performance à 12 ans qu'un individu dans le dernier niveau de performance (5% de différence). Cette différence augmente avec l'âge (16% à 15 ans).

[1] Institut de Mathématiques de Bourgogne, UMR 5584 CNRS, Université Bourgogne Franche-Comté, F-21000, Dijon, France
[2] Institute for Research in bioMedicine and Epidemiology of Sport (IRMES), Paris, France
[3] Scientific Department, French Ski Federation, Annecy, France
[4] Institut National du Sport, de l'Expertise et de la Performance (INSEP), Paris, France
[5] Hôtel-Dieu, Assistance Publique—Hôpitaux de Paris, Paris, France
[6] PAM UMR 02.102, Université Bourgogne Franche-Comté, AgroSup Dijon, 1 esplanade Erasme, F-21000, Dijon, France

[*] Correspondence: De Larochelambert Quentin <qdelarochelambert@ffs.fr>

1 Introduction

Performance : An age-related variable complex to predict in the long term

Estimating the future potential of young athletes has always been an essential issue in the field of sport to detect future champions. Large databases make it possible to take an innovative look at the performance and the understanding of its determinants, in particular with regard to already well-established physiological principles. DH Moore (Moore 1975) was the first to theorize the relationship between age and performance by a sum of two exponential laws. More recently, G Berthelot (Berthelot et al. 2019) took up DH Moore's equation by including a biological component. A family of models increasingly used in biostatistics can answer this question: Multi-state models.

Multi-state models

Multi-state models are a specific family of survival models, making it possible to study a categorical variable over time, generally a disease, and which can take more than two categories (Commenges 1999). This type of model has the particularity of taking into account incomplete, censored or truncated data.They are based on the transitions between the different states. Hougaar et al. (Hougaard 1999) lists the most common multi-state models. In this study, we develop a performance evolution model based on the markov property, by estimating the transitions intensities between the performance levels in a parametric and non-parametric way.

2 Methods

Dataset

The population studied included all French alpine skiers having participated in at least one competition in Giant Slalom (GS), Slalom (SL) Super Giant Slalom (SG) and Downhill (DH) between the age of 10 up to 30 from the 2004/2005 season up to the 2020/2021 season.

Longitudinal categorization of performance

To characterize the level at each age of each individual, we create performance lanes (PL). We define k be the number of PL and q be a vector of size $k+1$ containing the quantiles representing the level of each cluster at each age. Let X_j be the set of performances achieved in alpine skiing at age j. We calculate the upper limit of each PL i for each age j :

$$\forall t \in [1 \dots k], PL_{tj}^{+} = \mathrm{X}_{j}^{q_t} \tag{1}$$

To define the levels of interest in the objective of our method, we have chosen to have inhomogeneous PL with $q = \{0, 0.05, 0.1, 0.25, 0.5, 1\}$.

Modeling of transitions between performance lanes

Let the states corresponding to the PL in the Multi-state model with $C = \{1, 2, 3, 4, 5\}$, and let $X(t)$ be a random process taking values in C.

Empirical estimation by a time nonhomogeneous Markov chain

This first model makes it possible to obtain a trend of future lane probabilities. We define $M^t = (m^t_{i,j})$ transition matrices of size 5x5 with $m^t_{i,j}$ the probability of going from lane i to lane j from age t to age $t+1$, obtained empirically with the real transitions made by the skiers.

We define $P_{t\leftarrow s}(j|i)$ the probability of moving from lane i to lane j from age s to age t defined by (2)

$$P_{t\leftarrow s}(j|i) = \sum_{r\in R} \prod_{l}^{N-1} m^{a_l}_{r_l,r_{l+1}} \tag{2}$$

with $R = \{i\}\{1,...5\}^{N-2}\{j\}$ all possible paths from i to j, $a = \{r,...,t\}$ age vector from r to t, $N = card(a)$.

Estimation by a continuous-time Markovian multi-state model

Model property The multi-state models are based on the intensity of the transition or the risk rate $\alpha_{ij}(t)$, which expresses the instant risk of passing from state i to state j at time t. The multi-state model is a clock-forward model (Iacobelli and Carstensen 2013) based on the Markov property, i.e. all the information useful for predicting the future is included in the present state. All transitions are possible between the different PL.

We will seek to estimate in this model $P_{ij}(s,t) = P(X(t) = j \mid X(s) = i)$ the transition matrix, corresponding to the probability of going, directly or indirectly, from lane i performance to lane j performance from time s to time t.

Cumulative transition-specific hazards estimate

Non parametric In this section, the chosen estimator does not rely on any a priori hypothesis to describe the transition intensities; We use the Nelson Aalen estimator (Aalen 1978; Nelson 1972) to estimate the cumulative transition intensities allowing to take into account the right censorship and the left truncation.

Parametric In the parametric approach, we will try to $\mathrm{A}^{ij}_t = \int_0^t \alpha_{ij}(u)du$ by different laws commonly used (Kalbfleisch, n.d.; Lawless, n.d.), with the following probability density function : Exponential, Weibull, Gamma, Gompertz and Log-normal. The parameters were estimated using the likelihood maximization method.

Model quality To assess the quality of the different models, we calculated for each distribution the mean square error (MSE) between the multi-state estimations (MS) and the empiric estimations (EM) at each age t.

In order to have the most robust model possible, we chose a hybrid model by selecting, for each transition, the one with the weakest AIC.

Transition matrix estimation We then calculate the probabilities,using the parametric and non-parametric model, by means of a product integral (Andersen et al. 1993).

The estimates were calculated forward and backward. In forward estimations, we fix the PL start i and the start age s in $P_{ij}(s,t)$. We thus seek to calculate the probability of evolving in each of the PL

according to the starting PL i at age s. In the backward estimations, we fix j and t. Thus, we seek to know how the probabilities of reaching the PL j at age t evolve for each start PL.

3 Results

Longitudinal categorization of performance

Figure 1 (A) represents the performance density at each age from 10 to 30 years. Each distribution is split with the quantiles of order 0.05,0.1,0.25 and 0.5. By connecting the upper bounds of the quantiles of each density, we obtain the PL for the age of 10 to 30 years Figure 1 (B).

Modeling of transitions between performance lanes

The best model with the AIC criterion is the Weilbull model (AIC $= 1.2969162 \times 10^5$). Note that for all the transitions, the cumulative intensities have the form of a straight line (constant risk), or of a slight curve (risk decreasing monotonously). Therefore, we exclude the estimate by the log normal distribution in our hybrid model (where we select the best model for each transition), which considers an arc-shaped risk which increases then which decreases.

Figure 2 shows the forward probabilities of reaching PL 1 according to the starting lane at 10 years. The individuals present at the PL 1 to 10 years are those having the most chance of being in the PL 1 to 24 years but the differences in probabilities are low (2 % additional chance compared to PL 5, 1 % compared to PL 4, 1 % compared to PL 3, 1 % compared to PL 2). Again, by setting the entry PL at 15 years, individuals closer to PL 1 are more likely to be present in PL 1 to 24 years but the probabilities are more heterogeneous (10 % additional chance compared to PL 5, 9 % compared to PL 4, 8 % compared to PL 3, 4 % compared to PL 2).

4 Conclusion

We modeled the probabilities of evolution of the level of performance according to the starting level over the ages. We show that the difference in accession to the highest national level according to the starting level is very small among young skiers, then increases with age. In perspective, the model can be refined by adding other individual co-variables, such as anthropometric data (heights, weights), or by adding memory to the Markov process, and therefore taking into account the kinetics of progression in the process estimation of potential.

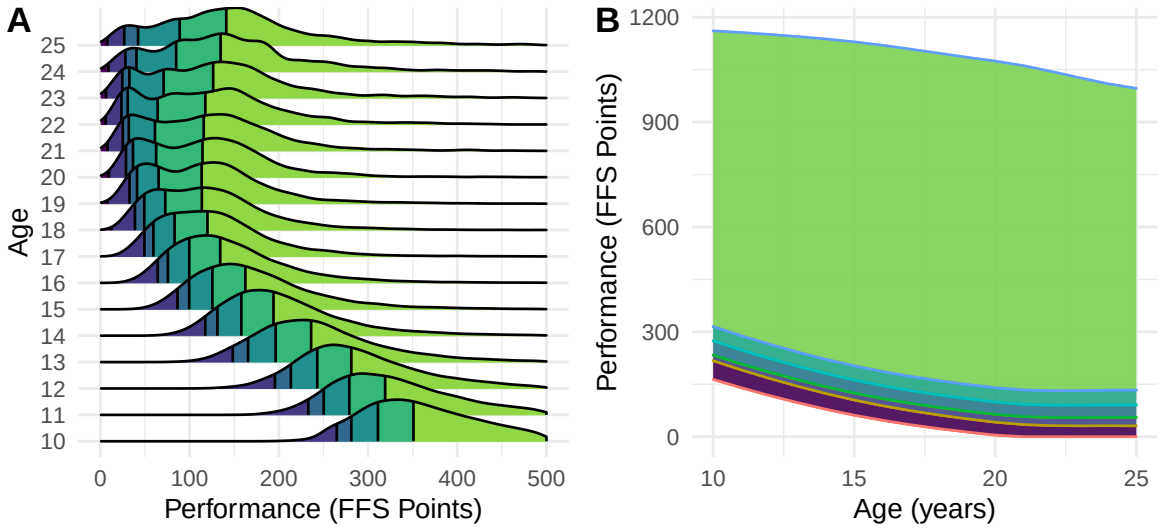


Figure 1: The five performance lanes from the age of 10 to 30 years old

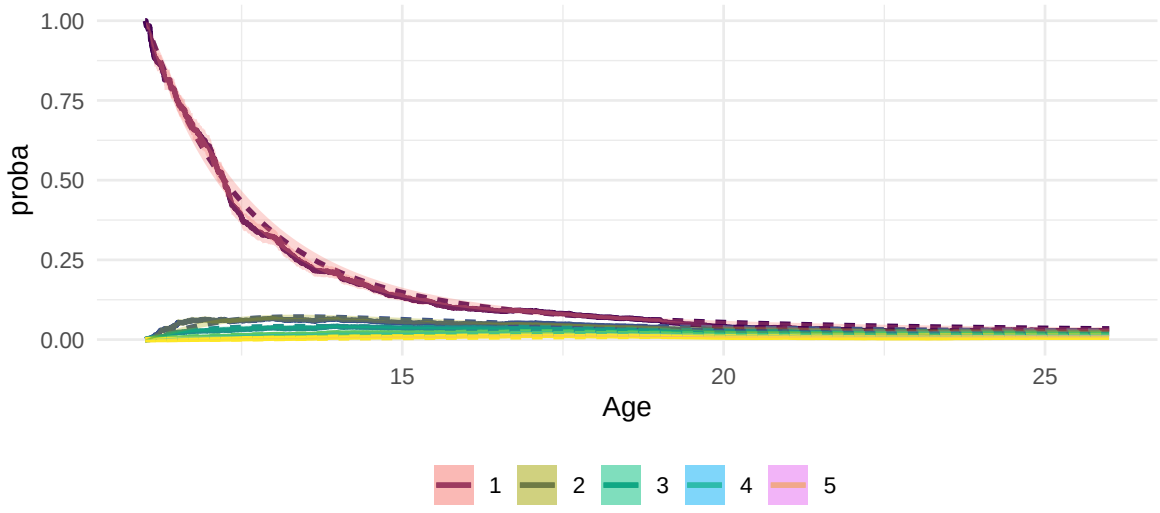


Figure 2: Probabilities transition

Bibliography

Aalen, Odd. 1978. "Nonparametric Inference for a Family of Counting Processes." *The Annals of Statistics* 6 (4): 701–26. https://doi.org/10.1214/aos/1176344247.

Andersen, PER KRAGH, Ornulf Borgan, Richard D. Gill, and Niels Keiding. 1993. *Statistical Models Based on Counting Processes*. Springer Series in Statistics. New York: Springer-Verlag. https://doi.org/10.1007/978-1-4612-4348-9.

Berthelot, Geoffroy, Avner Bar-Hen, Adrien Marck, Vincent Foulonneau, Stéphane Douady, Philippe Noirez, Pauline B. Zablocki-Thomas, et al. 2019. "An Integrative Modeling Approach to the Age-Performance Relationship in Mammals at the Cellular Scale." *Scientific Reports* 9 (1): 418. https://doi.org/10.1038/s41598-018-36707-3.

Commenges, D. 1999. "Multi-State Models in Epidemiology." *Lifetime Data Analysis* 5 (4): 315–27. https://doi.org/10.1023/A:1009636125294.

Hougaard, Philip. 1999. "Multi-State Models: A Review." *Lifetime Data Analysis* 5 (3): 239–64. https://doi.org/10.1023/A:1009672031531.

Iacobelli, Simona, and Bendix Carstensen. 2013. "Multiple Time Scales in Multi-State Models." *Statistics in Medicine* 32 (30): 5315–27. https://doi.org/10.1002/sim.5976.

Kalbfleisch, John. n.d. "The Statistical Analysis of Failure Time Data, 2nd Edition | Wiley." *Wiley.com*. https://www.wiley.com/en-us/The+Statistical+Analysis+of+Failure+Time+Data%2C+2nd+Edition-p-9781118031230.

Lawless, Jerald. n.d. "Statistical Models and Methods for Lifetime Data | Wiley Series in Probability and Statistics." https://onlinelibrary.wiley.com/doi/book/10.1002/9781118033005.

Moore, D. H. 1975. "A Study of Age Group Track and Field Records to Relate Age and Running Speed." *Nature* 253 (5489): 264–65. https://doi.org/10.1038/253264a0.

Nelson, Wayne. 1972. "Theory and Applications of Hazard Plotting for Censored Failure Data." *Technometrics* 14 (4): 945–66. https://doi.org/10.1080/00401706.1972.10488991.

REPARLONS DE NOTES EN PATINAGE

Léo Gerville-Réache[1]

[1]*Université de Bordeaux, CNRS, UMR 5218 IMS, France, leo.gerville-reache@u-bordeaux.fr*

Résumé. Les Jeux olympiques d'hiver de 2022 étant terminés, cette communication vous propose de replonger dans l'une des épreuves qui a permis à la France de capitaliser une médaille d'or de plus. Il s'agit de la danse sur glace. Comme d'habitude, c'est un jury composé de 9 juges qui, via un ensemble de notes, a classé les prestations des 23 couples en compétition. En retirant toujours la meilleure et la moins bonne note de chaque évaluation, il nous a semblé opportun de rappeler et réutiliser les possibilités statistiques qu'offre le modèle de Gauss-Markov.

Mots-clés. Jury, Patinage, Gauss-Makov, JO 2022

Abstract. With the end of the 2022 Winter Olympic Games, this talk invites you to dive back into one of the events that allowed France to capitalize on one more gold medal. This is ice dancing. As usual, it is a jury composed of 9 judges who, through a set of notes, ranked the performances of the 23 couples in competition. By always removing the best and worse score from each evaluation, it seemed appropriate to recall and reuse the statistical possibilities offered by the Gauss-Markov model.

Keywords. Jury, Ice skating, Gauss-Makov, Olympic Games 2022

1. Introduction

Nombre de sports fondent (pour tout ou partie) les résultats d'une compétition sur des notes attribuées par des juges. Dans les sports d'hiver, c'est le patinage artistique qui vient rapidement à l'esprit, mais le plongeon ou encore la gymnastique sont évidemment du même ressort…

Si vous regardez attentivement comment cela fonctionne, vous verrez que chaque juge (un expert reconnu de la discipline) attribue pour chaque prestation de la compétition, une ou plusieurs notes selon un barème de cotation. Pour autant, il est rare que les juges mettent exactement la même note à une même prestation. Il persiste une part de subjectivité (de variabilité) qui est bien normale.

La plupart du temps, sur l'ensemble des notes distribuées par le jury à une prestation, pour calculer la « note finale », on retire la note la plus basse et la note la plus haute. L'idée est de limiter l'impact d'une note, intentionnellement ou non, trop éloignée de la moyenne des notes attribuées. Mais est-ce une bonne d'idée ? Utilise-t-on bien toutes les possibilités de l'analyse statistique ?

Sans faire appel à l'Intelligence Artificielle pour une notation plus juste, via de la reconnaissance automatique liée à des caméras 3D [2], cette communication propose de revenir sur les résultats qui ont conduit le couple Gabriella PAPADAKIS / Guillaume CIZERON à remporter une médaille d'or aux JO de 2022. Via le modèle de Gauss-Markov déjà présenté aux JDS de 1999 [3], cette communication propose de rappeler cette approche statistique sur des données très actuelles.

2. Les données

En cherchant sur le site officiel https://olympics.com, on peut trouver le rapport officiel [1] de la compétition de patinage artistique où le couple français Gabriella PAPADAKIS / Guillaume CIZERON a été sacré champion olympique de danse rythmique sur glace. En moyennant les notes « Program Componen » de chaque jury pour chaque prestation, on obtient le tableau suivant (ici dans l'ordre du classement final qui prend en compte également les notes « Executed Element »). « CL1 » est le couple classé globalement premier et ayant donc obtenu la médaille d'or.

Notes	J1	J2	J3	J4	J5	J6	J7	J8	J9
CL1	9,85	9,70	9,90	9,85	9,85	9,60	9,80	9,70	9,85
CL2	9,75	9,65	9,45	9,65	9,65	9,70	9,65	9,75	9,50
CL3	9,85	9,55	9,75	9,70	9,65	9,45	9,45	9,50	9,35
CL4	9,65	9,15	9,60	9,50	9,50	9,45	9,55	9,35	9,10
CL5	9,60	9,15	9,05	9,25	9,25	9,40	9,25	9,40	9,10
CL6	9,40	9,40	9,15	9,45	9,50	9,30	9,25	9,45	9,05
CL7	9,15	9,20	9,25	9,40	9,15	9,35	9,15	9,25	8,85
CL8	8,65	8,65	8,70	8,95	8,80	8,50	8,65	8,75	8,45
CL9	8,85	8,55	8,70	8,85	8,70	8,65	8,65	8,85	8,70
CL10	8,60	8,50	8,85	8,65	8,85	8,70	8,55	8,65	8,90
CL11	8,50	8,45	8,60	8,80	8,50	7,90	8,30	8,15	8,60
CL12	8,10	8,40	8,40	8,50	8,20	8,15	8,40	8,45	8,15
CL13	8,35	8,15	8,25	8,45	8,30	8,05	7,90	8,10	8,40
CL14	8,05	7,90	8,05	8,30	8,15	8,00	7,90	8,35	8,15
CL15	7,30	7,40	7,70	7,55	7,85	7,80	7,65	8,00	8,00
CL16	8,10	7,55	7,75	7,75	7,70	7,70	7,90	7,95	7,60
CL17	7,40	7,75	7,70	7,60	7,35	7,45	7,35	7,40	7,85
CL18	7,35	7,40	7,45	7,55	7,45	7,25	7,55	7,55	7,95
CL19	7,15	7,50	7,95	8,10	7,55	7,45	7,35	7,40	7,85
CL20	7,65	7,40	7,65	7,40	7,50	7,25	7,65	7,75	7,40
CL21	7,60	7,35	7,50	7,35	7,90	7,25	7,45	7,65	7,45
CL22	7,75	7,40	7,70	7,50	7,55	7,25	7,15	7,25	7,65
CL23	5,70	6,20	7,30	6,10	5,75	6,85	6,50	6,55	7,25

3. Le modèle de Gauss-Markov

On suppose que I compétiteurs sont notés par J juges. Chaque juge note chaque compétiteur indépendamment les uns des autres avec la même échelle. Plus la performance est de qualité, plus la note est haute. Soit X_{ij} la note du juge j sur la prestation i. On suppose que la note X_{ij} est la réalisation de la variable aléatoire suivante :

$$X_{ij} = a_i + b_j + Y_{ij}, i = 1, \dots, I \text{ et } j = 1, \dots, J.$$

Où a_i est la « vraie » valeur de la prestation i, b_j est « l'erreur systématique » du juge j et Y_{ij} est une erreur aléatoire. Dans le modèle de Gauss-Makov, les erreurs Y_{ij} sont des variables aléatoires indépendantes et identiquement distribuées $N(0; \sigma^2)$.

Les valeurs a_i, b_j et σ^2 sont inconnues mais ce n'est pas difficile de les estimer. En posant que $b_1 + b_2 + \cdots + b_J = 0$, on obtient alors des estimations des MCO particulièrement simples des a_i et b_j :

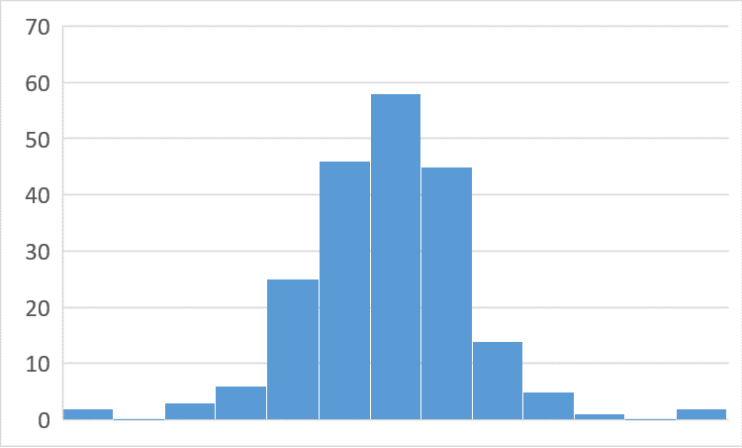
$$\hat{a}_i = \frac{1}{J}\sum_{j=1}^{J} x_{ij} \text{ et } \hat{b}_j = \frac{1}{I}\sum_{i=1}^{I} x_{ij} - \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J} x_{ij}$$

Pour identifier une note aberrante, il est possible de mettre en place un test statistique spécifique à ce modèle de Gauss-Markov. Une note aberrante est nécessairement une note qui s'écarte fortement de

la note « attendue » par un juge donné à une prestation donnée. C'est en regardant les y_{ij} (qui sont par définition des écarts (ou erreurs) entre le modèle et les observations que cela est possible. Voici le tableau des $y_{ij} = x_{ij} - \hat{a}_i - \hat{b}_i$.

Y_{ij}	J1	J2	J3	J4	J5	J6	J7	J8	J9	a_i
CL1	0,06	0,00	0,02	-0,02	0,05	-0,10	0,07	-0,12	0,03	**9,79**
CL2	0,11	0,10	-0,28	-0,07	0,00	0,15	0,07	0,08	-0,17	**9,64**
CL3	0,27	0,06	0,08	0,04	0,06	-0,05	-0,07	-0,12	-0,27	**9,58**
CL4	0,23	-0,19	0,09	-0,01	0,06	0,11	0,18	-0,11	-0,36	**9,43**
CL5	0,33	-0,03	-0,31	-0,10	-0,03	0,21	0,04	0,09	-0,20	**9,27**
CL6	0,08	0,16	-0,26	0,04	0,16	0,06	-0,02	0,09	-0,31	**9,33**
CL7	-0,04	0,10	-0,03	0,13	-0,05	0,24	0,02	0,02	-0,38	**9,19**
CL8	-0,02	0,06	-0,06	0,19	0,11	-0,09	0,03	0,04	-0,26	**8,68**
CL9	0,13	-0,08	-0,11	0,05	-0,03	0,01	-0,01	0,09	-0,05	**8,72**
CL10	-0,09	-0,10	0,07	-0,12	0,15	0,09	-0,08	-0,08	0,17	**8,69**
CL11	0,08	0,12	0,09	0,30	0,07	-0,44	-0,06	-0,31	0,15	**8,42**
CL12	-0,20	0,18	0,01	0,12	-0,12	-0,07	0,16	0,11	-0,19	**8,31**
CL13	0,14	0,02	-0,05	0,16	0,07	-0,08	-0,26	-0,15	0,15	**8,22**
CL14	-0,04	-0,10	-0,13	0,13	0,05	-0,01	-0,13	0,22	0,02	**8,09**
CL15	-0,39	-0,20	-0,08	-0,22	0,15	0,19	0,02	0,27	0,27	**7,69**
CL16	0,33	-0,14	-0,11	-0,11	-0,09	0,01	0,18	0,14	-0,21	**7,78**
CL17	-0,14	0,30	0,07	-0,02	-0,20	0,00	-0,13	-0,17	0,28	**7,54**
CL18	-0,15	-0,01	-0,14	-0,03	-0,06	-0,16	0,11	0,02	0,42	**7,50**
CL19	-0,44	0,00	0,27	0,43	-0,05	-0,05	-0,18	-0,22	0,23	**7,59**
CL20	0,14	-0,03	0,05	-0,19	-0,03	-0,18	0,19	0,20	-0,15	**7,52**
CL21	0,10	-0,06	-0,09	-0,23	0,39	-0,16	0,01	0,12	-0,08	**7,50**
CL22	0,29	0,02	0,15	-0,04	0,07	-0,13	-0,26	-0,25	0,15	**7,47**
CL23	-0,76	-0,18	0,75	-0,44	-0,73	0,47	0,09	0,05	0,75	**6,47**
b_j	**0,00**	**-0,09**	**0,09**	**0,08**	**0,01**	**-0,09**	**-0,06**	**0,03**	**0,03**	

Le plus gros écart en valeur absolue est pour le juge 1 sur la prestation du couple 23. L'écart est de précisément -0,76 et la note donnée de 5,7. Ce juge a mis une note très élevée à cette prestation.



NB : La normalité des Y_{ij} est vérifiée via l'histogramme des y_{ij} ci-contre.

Mais cette note est-elle significativement trop élevée ?

Si l'on souhaite tester l'hypothèse H_0 selon laquelle il n'existe aucune note aberrante, on peut procéder ainsi [3,4]. On définit Λ :

$$\Lambda = \frac{IJ}{(I-1)(J-1)} \frac{\max_{ij} y_{ij}}{S^2} \quad \text{avec} \quad S^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} y_{ij}^2.$$

Ici, $\Lambda = 0{,}087$. Sous l'hypothèse H_0 la p-value est déterminée de la manière suivante :

$$p = IJ\left[1 - I_z\left(\Lambda; \frac{1}{2}; \frac{1}{2}(IJ - I - J)\right)\right]$$

Où I_z est la fonction Béta incomplète.

La p-value vaut environ 0,015. Cela signifie que la note attribuée par le juge 1 à la prestation du couple 23 est significativement trop basse. La raison n'est pas connue mais la p-value est nette ! Il est alors possible de proposer une correction de cette note grâce à la formule suivante :

$$\tilde{x}_{23;1} = x_{23;1} + \frac{IJ}{(I-1)(J-1)} y_{23;1}$$

La note corrigée $\tilde{x}_{23;1}$ serait alors de 6,60 (au lieu de 5,70).

Maintenant, testons l'hypothèse qu'un juge soit moins précis que les autres juges. Pour cela on commence par mesurer la précision de chaque juge avec la formule suivante [3,4] :

$$\Lambda_j = \frac{J}{(J-1)S^2} \sum_{i=1}^{I} y_{ij}^2$$

Sous l'hypothèse H_0 la p-value est déterminée de la manière suivante :

$$p = J\left[1 - I_z\left(\mathrm{Max}_j \Lambda_j; \frac{I-1}{2}; \frac{(I-1)(J-1)}{2}\right)\right]$$

On trouve ici que $\mathrm{Max}_j \Lambda_j = \Lambda_9 = 0{,}245$. La p-value pour ce juge 9 est de 0,003. Le juge 9 est significativement moins précis que les autres ! C'est à vous de prendre votre décision quant à la qualité de ce juge et la pertinence de ses notes…

Pour finir, qu'en est-il des résultats de la compétition ? En tous les cas, de la partie « Program Componen » de cette compétition. Qu'est-ce-que ce jury est en capacité de juger ?

Cette fois-ci, l'objectif est de tester si un ensemble de prestations donné est statistiquement indiscernable ou non. Pour reformuler : Le jury est-il suffisament précis pour affirmer une différence significative, par exemple, entre les couples CL1 et CL2 par exemple

On cherche à tester l'hypothèse selon laquelle la « vraie » valeur de k prestations d'indice $(i_1, \ldots, i_k) = \Phi$ est la même. C'est-à-dire que :

$$H_0 : a_{i_1} = a_{i_2} \ldots = a_{i_k}.$$

Pour cela, on calcule la statistique [3,4] :

$$Z(\Phi) = \frac{J(J-1)(I-1)}{(k-1)S^2} \sum_{i \in \Phi} \left(a_i - \frac{1}{k} \sum_{i \in \Phi} a_i\right)^2$$

Sous l'hypothèse H_0, $Z(\Phi)$ suit une loi de Fisher à $k-1$ et $(I-1)(J-1)$ degrés de liberté. La p-value du test est alors la suivante:

$$p = F^{-1}_{k-1,(J-1)(I-1)}\big(Z(\Phi)\big)$$

En testant, $a_1 = a_2$, c'est-à-dire que les prestations des couples CL1 et CL2 sont égales (ou « indiscernables ») on obtient $Z(\Phi) = 2{,}25$ et on trouve $p = 0{,}14$. Aussi la différence n'est pas statistiquement significative. Le jury n'est pas en capacité de produire une différence significative entre le couple CL1 et le couple CL2.

Pour cette question de la différence significative entre CL1 et CL2, on aurait pu penser réaliser un test de Student sur deux échantillons appariés. C'est après tout, une approche tout à fait classique. La p-value vaut dans ce cas $p = 0{,}034$. C'est-à-dire qu'avec une approche via le test de Student, la différence entre CL1 et CL2 est significative. Il n'est pas rare que deux approches statistiques produisent des décisions différentes. Pour autant, dans notre cas, il semble que prendre en compte toute l'information contenue dans le tableau de données soit plus pertinente…

4. Discussion

Avec une approche via le modèle de Gauss-Markov, il a été possible de tester la qualité des juges et finalement mesurer la qualité du jury. Ce jury est capable de distinguer certaines prestations mais pas toutes. Par exemple, bien que les moyennes des couples CL1 et CL2 soient différentes (9,79 contre 9,64) celles-ci ne sont pas significativement différentes.

Mais cette approche demande d'attendre la fin de la compétition pour que le classement soit réalisé. Un peu comme une délibération (statistique) du jury... On perd cependant cette « magie » de l'attente du couple en tête de la compétition au fur et à mesure des prestations.

Maintenant, qu'en est-il de la procédure actuelle qui consiste à systématiquement éliminer (du calcul de la moyenne de chaque prestation) la note la plus basse et la note la plus haute ; procédure ayant clairement pour but d'éviter les notes anormales (dans un sens ou dans l'autre).

Le plus important est de comprendre qu'au lieu d'avoir dans notre cas, 9 notes par prestation, il n'en reste plus que 7. Pourtant, si les juges sont « honnêtes », c'est une perte d'information et de précision importante. En retirant systématiquement 2 notes de manière préventive, on se prive potentiellement d'une plus grande précision du jury.

Pour autant, chaque juge sachant que s'il note trop bas ou trop haut par rapport aux autres juges, sa note ne sera pas prise en compte a un avantage. Ce principe « oblige » les juges à maitriser leur notation qui, dans le cas contraire, pourraient voir leur note non prise en compte…

L'un des écueils de la procédure qui élimine systématiquement la meilleure et la moins bonne note, c'est qu'une note anormale peut parfaitement se glisser dans les notes conservées [2]. Clairement, un juge qui note bas comme le juge 2 et qui une fois note haut (mais pas trop) va voir sa note anormale conservée ! De plus, un juge qui aurait une bonne évaluation des prestations mais qui surnoterait systématiquement verrait ses notes retirées alors qu'il classe correctement les prestations.

La recherche d'une procédure « optimale » est délicate. Cette optimalité n'existe sans doute pas. Comme toute procédure d'évaluation, elle doit être honnête et claire. Il est sans doute difficile d'admettre l'idée que les notes soient retenues et « délibérées » en fin de compétition via le modèle de Gauss-Markov, même si statistiquement, cette approche est pertinente. Il est également difficile d'admettre qu'un jury, dont la précision n'est pas infaillible, classe des prestations alors qu'elles sont clairement indiscernables pour ce jury.

On dit souvent que le jury est souverain. C'est sans doute le principe qui permet d'accepter, le plus souvent, le résultat prononcé…

Bibliographie

[1] https://olympics.com/beijing-2022/olympic-games/static/owg2022/pdf/OWG2022/FSK/OWG2022_FSK_C77B_FSKXICEDANCE----------QUAL000100--.pdf

[2] *JO 2022.* En patinage artistique, l'Intelligence Artificielle pour une notation plus juste ?, *Journal Ouest France,* 13 Février 2022.

[3] Gerville-Réache L., Nikulin M.S. (1999). Analyse statistique de l'évaluation des sportifs par un jury, *Journées de statistique,* Grenoble, France, 51-54.

[4] Greenwood P.E., Nikulin M.S. (1996) *A Guide to Chi-Squared Testing* John Wiley & Sons. Chap. 24 : Statistical Evaluation of Ratings in Sports Compétitions Such as Skating or Gymnatics.

Modélisation de la probabilité de réussite d'une passe au football

Sébastien Déjean [1] & J avier Lopez Sanchez [2] & Philippe Saint Pierre [1]

[1] *Institut de Mathématiques de Toulouse, UMR 5219, Université de Toulouse et CNRS,* {*sebastien.dejean, philippe.saint-pierre*}*@math.univ-toulouse.fr*
[2] *Toulouse Université Club Tennis et Athletics Coaching Club Ramonville, jlsone1@hotmail.com*

Résumé. Depuis quelques années, le sport offre un environnement très dynamique pour l'analyse statistique de données. Que ce soit pour accompagner le développement de systèmes de paris sportifs, pour agrémenter les diffusions télévisées de rencontres sportives ou en vue d'optimiser les performances d'un athlète ou d'une équipe, l'analyse de données est une aide précieuse dans de tels contextes. Pour le football, plusieurs études ont conduit à populariser la notion d'*expected goal* traduisant la probabilité de marquer un but pour un joueur (ou une équipe) dans une situation donnée. À ce jour, cependant, peu de travaux semblent se consacrer à la probabilité de réussite d'une passe pourtant élément majeur dans tout sport collectif. C'est avec l'objectif de contribuer à ce sujet que nous avons utilisé la régression logistique pour modéliser la probabilité de réussite d'une passe en fonction de différentes variables explicatives. Nos analyses montrent des résultats qui, loin de révolutionner le football (plus une passe est courte et vers l'arrière, plus il est probable de la réussir) mettent en évidence de nombreuses perspectives pour l'optimisation de la performance d'une équipe.

Mots-clés. données sportives, statistique descriptive, régression logistique, visualisation de données

Abstract. In recent years, sport has provided a very dynamic environment for statistical data analysis. Whether it is to accompany the development of sports betting systems, to enhance television broadcasts of sports games or to optimise the performance of an athlete or a team, data analysis is a valuable aid in such contexts. For football, several studies have led to the popularisation of the notion of *expected goal* translating the probability of scoring a goal for a player (or a team) in a given situation. To date, however, few studies seem to focus on the probability of a successful pass, which is a major element in any team sport. With the aim of contributing to this subject, we used logistic regression to model the probability of a successful pass as a function of different explanatory variables. Our analyses show results which, far from revolutionising football (the shorter the pass and the further back it is, the more likely it is to be successful), highlight numerous perspectives for the optimisation of a team's performance.

Keywords. sports data, descriptive statistics, logistic regression, data visualization

1 Introduction

Qu'elle porte le nom de Big Data ou d'intelligence artificielle, l'analyse de données est en plein essor dans le domaine du sport. Cette approche a été popularisée notamment depuis 2011 avec la sortie du film Le Stratège (*Moneyball*) qui s'inspire d'une histoire vraie d'un analyste statisticien recruté par un club de base-ball. Dans la presse sportive, des articles de plus en plus nombreux font état de l'apport de ce domaine de l'IA dans le management des équipes, et de plus en plus de sociétés se positionnent sur ce marché très porteur pour proposer leurs services d'analystes.

À notre échelle, nous nous intéressons depuis quelques années à ce domaine en proposant des sujets de projets à des élèves ingénieurs de l'Institut National des Sciences Appliquées de Toulouse. Pour alimenter ces sujets, nous nous appuyons sur des bases de données disponibles gratuitement sur des sites comme `whoscored.com` ou `statsbomb.com`.

Pour le football, plusieurs études ont conduit à populariser la notion d'*expected goal* traduisant la probabilité de marquer un but pour un joueur (ou une équipe) dans une situation donnée. Les *expected goal* sont maintenant utilisés comme indicateur de la performance d'une équipe et intéressent également le domaine des paris sportifs (Steffen *et al.*, 2019). À ce jour, cependant, peu de travaux semblent se consacrer à la probabilité de réussite d'une passe pourtant élément majeur dans tout sport collectif (Ievoli *et al.*, 2021). Dans le travail présenté ici, nous nous sommes plus particulièrement intéressés aux passes lors d'un match de football. De nombreuses caractéristiques sont disponibles pour représenter une passe : le nom du joueur ou de la joueuse qui l'a effectuée, le moment du match, l'endroit du terrain d'où est partie la passe, l'endroit où elle est arrivée, la partie du corps (pied gauche, pied droit, tête, main si c'est un gardien de but...), etc... et surtout, une information concernant la réussite ou l'échec de la passe. C'est sur cette dernière information que nous nous sommes focalisés pour modéliser la probabilité de réussite d'une passe pendant un match de football. Une fois les données recueillies et pré-traitées, la mise en œuvre d'une régression logistique, déjà utilisée par Pollard and Reed (1997) pour mesurer l'efficacité de certaines stratégies au football, permet d'aboutir à un tel modèle.

Dans cet article, nous présentons d'abord les données qui nous ont servi pour réaliser ce travail. Dans une deuxième partie, nous abordons la partie modélisation pour laquelle nous proposons une visualisation originale des résultats basée sur une représentation schématique d'un terrain de football. Enfin, nous exprimons quelques perspectives de ce travail en lien avec la recherche d'indicateurs de performance susceptibles d'être révélés par des analyses statistiques.

2 Les données

2.1 Récupération des données

Les données que nous analysons dans ce travail proviennent du site StatsBomb ([1]). Elles sont disponibles gratuitement dans le cadre du *Free Data Offerings from StatsBomb Services* et peuvent être récupérées en utilisant le package `StatsBombR` ([2]) en vue d'analyses réalisées avec le logiciel R. Parmi l'ensemble des données disponibles, nous nous sommes restreints à celles relatives aux 64 rencontres de la Coupe du monde de football 2018.

2.2 Description succincte des données

Lors de ces 64 rencontres, **227 886** passes ont été enregistrées. Elles sont caractérisées par différentes variables comme le nom de l'équipe, la phase de jeu (*Regular play, Free kick, throw in...*), le joueur réalisant la passe, le joueur la recevant, la position du ballon au début de la passe, la longueur de la passe, l'angle que fait la passe avec la droite qui relie les 2 buts, la hauteur de la passe (3 modalités : au sol, à mi-hauteur, en hauteur), le fait d'être sous la pression de l'adversaire ou pas et, point important pour notre étude, la réussite (le ballon est arrivé dans les pieds d'un partenaire) ou l'échec (ballon récupéré par l'adversaire ou sorti hors des limites du terrain) de la passe. Le nombre de passes réussies est de **215 172** soit un taux de réussite de l'ordre de 95%.

Pour illustrer ces données, la figure 1 représente sur la localisation du début (à gauche) et de la fin (à droite) des passes avec un codage couleur indiquant la réussite (en vert) ou l'échec (en marron) de la passe.

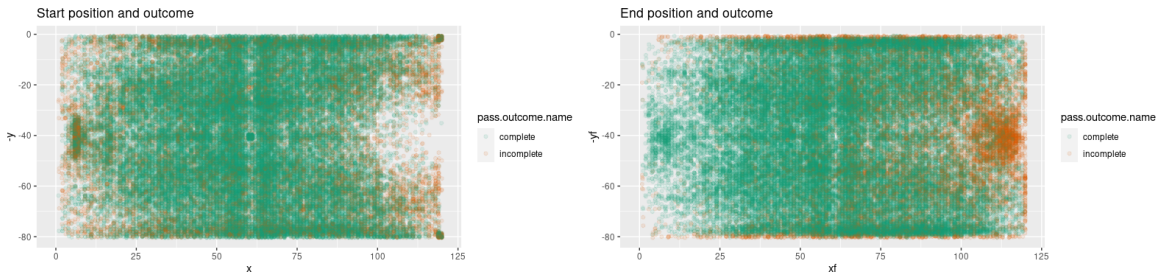


Figure 1: Représentation de la position du début (à gauche) et de la fin (à droite) d'une passe. La couleur indique la réussite (en vert) ou l'échec de la passe (en marron). Les points sont localisés en supposant que l'équipe en possession du ballon défend le but situé à gauche de la figure et attaque vers la droite.

On peut noter sur cette figure que peu de passes ont pour origine la surface de

[1] www.statsbomb.com STATSBOMB
[2] github.com/statsbomb/StatsBombR

réparation adverse (zone proche des buts adverses) et que les passes qui arrivent dans cette zone sont généralement ratées. Ceci est une évidence compte tenu de la forte densité de joueurs adverses dans une telle zone défensive.

Une autre illustration des données est proposée sur la figure 2. Cette fois-ci, la représentation en coordonnées polaires utilisant la longueur et l'angle de la passe, illustre là aussi une évidence footballistique : une passe vers l'arrière ou sur les côtés est plus susceptible d'être réussie qu'une passe vers l'avant.

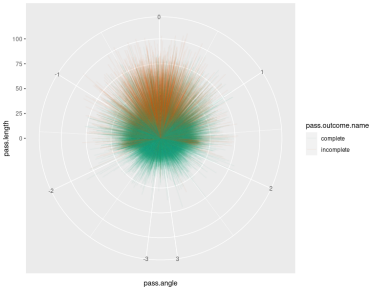


Figure 2: Représentation en coordonnées polaires de la longueur des passes (en yards) en fonction de l'angle de la passe (en radian, 0 indiquant une passe en direction des buts adverses, π ou $-\pi$ une passe en retrait). La couleur indique la réussite (en vert) ou l'échec de la passe (en marron).

3 Modélisation de la réussite d'une passe

Soit Y une variable à valeurs dans $0, 1$ à expliquer par p variables explicatives $\mathbf{X} = (X_1, ..., X_p)$. Le modèle logistique propose une modélisation de la loi de $Y|X = x$ par une loi de Bernoulli de paramètre $p(x) = \mathbb{P}(Y = 1|X = x)$ telle que :

$$\mathbb{P}(Y = 1|\mathbf{X} = \mathbf{x}) = \frac{e^{\alpha+\beta\mathbf{x}}}{1 + e^{\alpha+\beta\mathbf{x}}}$$

où α et $\beta = (beta_1, ..., beta_p)$ sont les coefficients de régressions. Supposons qu'on observe un échantillon $(Y_i, \mathbf{X}_i)_{i=1,...,n}$ de n passes réussies ou non. Les estimateurs de β sont obtenus par maximum de vraisemblance. Les coefficients de régressions sont interprétables en termes d'*odds ratio* ($OR_k = e^{\beta_k}$). Le test du rapport de vraisemblance peut être utilisé pour tester la nullité des coefficients et comparer des modèles emboîtés afin de sélectionner un modèle. La probabilité de réussir une passe dans une situation donnée se déduit aisément du modèle logistique.

4 Application

4.1 Résultats d'un modèle de régression logistique

Nous présentons ici les résultats d'un modèle prenant en compte la position de départ de la passe dans l'axe du terrain (x), la valeur absolue de la position latérale ($abs(y)$) afin de considérer de façon équivalente les parties gauche et droite du terrain, le cosinus de

l'angle de la passe afin là aussi de respecter la symétrie du terrain, et enfin le fait d'être sous la pression de l'équipe adverse.

Table 1: Résultats de la régression logistique modélisant la probabilité de réussite d'une passe en fonction de la position d'origine de la passe, de la longueur de la passe, du cosinus de l'angle de la passe et du fait d'être sous la pression de l'équipe adverse

	Estimate	Std. Error	z value	$\Pr(>\|z\|)$
(Intercept)	4.3090	0.0460	93.67	< 0.00001
x	-0.0164	0.0004	-36.99	< 0.00001
abs(y40)	-0.0190	0.0010	-19.71	< 0.00001
cos(pass.angle)	-1.5683	0.0228	-68.72	< 0.00001
pass.length	-0.0357	0.0007	-52.37	< 0.00001
under_pressurePressure_YES	-0.4048	0.0265	-15.30	< 0.00001

Le tableau 1 indique que toutes les variables du modèle sont significatives. Les estimations des coefficients s'interprètent logiquement en termes footballistiques. Par exemple, l'*odd ratio* associé au fait d'être sous pression de l'équipe adverse ($OR = e^{-0.4048} < 1$) indique qu'une pression diminue la probabilité de réussite d'une passe. Les régressions simples considérant chaque variable explicative séparément présentent des résultats cohérents avec ceux-ci ; ce qui exclut des problèmes de multicolinéarité entre les prédicteurs.

4.2 Visualisation des résultats

Le modèle permet d'estimer la probabilité de réussite d'une passe dans différentes situations. Par exemple, sachant la position d'un joueur qui fait une passe, on peut représenter sur un terrain de football la probabilité de réussite d'une passe en fonction de l'arrivée de la passe. Les probabilités de succès d'une passe peuvent ensuite être comparées dans différentes situations. La figure 3 représente les probabilités de succès dans une situation où le joueur qui fait la passe est dans une situation offensive (but adverse à droite de l'image) en position axiale (figure de gauche) ou décalé sur un côté du terrain (figure de droite).

5 Conclusion et perspectives

Nous n'avons pas révolutionné le football avec un modèle qui indique qu'une passe courte en retrait a plus de chances d'aboutir dans les pieds d'un partenaire qu'une passe longue vers l'avant. Il n'est pas non plus question de remplacer l'expérience et l'expertise d'un entraîneur chevronné qui a passé de longues années sur le bord des terrains. En revanche, nous pensons que l'analyse statistique de données sportives peut apporter un complément d'informations important permettant d'optimiser les performances aussi bien d'une équipe

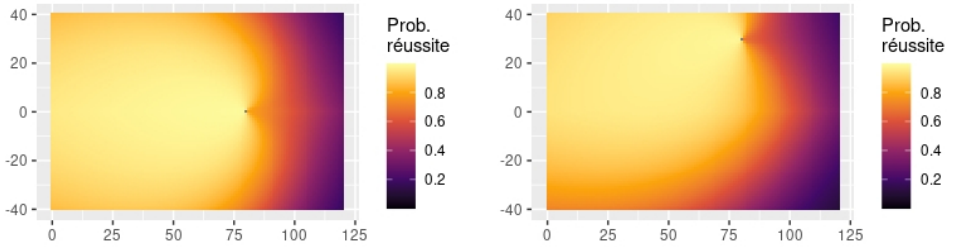


Figure 3: Représentation de la probabilité de réussite d'une passe à partir de 2 positions offensives : en position axiale (à gauche), une position excentrée (à droite). Les dimensions du terrain sont exprimées en yards.

que d'un joueur ou d'une joueuse. L'objectif de ce travail était d'expliquer l'échec d'une passe et de proposer des résultats facilement utilisables pour analyser la performance d'un joueur ou d'une équipe.

D'un point de vue pratique, la prédiction de la réussite d'une passe dans chacune des situations rencontrées dans un match peut être déduite d'un tel modèle. Il est par exemple possible de définir, pour chaque joueur ou pour chaque équipe, un indicateur de performance basé sur la comparaison du nombre de passes prédites et effectivement observées. Cet indicateur peut être calculé sur des intervalles de temps afin de représenter l'évolution des performances au cours d'un match. Une autre perspective est de récupérer des données plus complètes afin de disposer de plus d'informations sur le déroulement du match (la position des autres joueurs par exemple) et ainsi inclure un plus grand nombre de variables dans l'analyse. Des méthodes de *machine learning* peuvent alors être mises en œuvre pour améliorer les qualités prédictives du modèles.

6 Remerciements

Les auteurs remercient Zoé Philippon et Lauriane Kiersnowski étudiantes à l'INSA de Toulouse qui ont contribué à ce travail dans le cadre d'un projet tutoré.

Bibliographie

Steffen, P. and Gerville-Réache, L. Bisoffi, N. (2019). Paris sportifs au football : l'intérêt des expected goals. Journées de Statistique, Nancy, France. ⟨hal-02150047⟩

Ievoli, R., Palazzo, L. and Ragozini, G. On the use of passing network indicators to predict football outcomes (2021), *Knowledge-Based Systems*, 222, 106997

Pollard, R. and Reep, C. (1997). Measuring the Effectiveness of Playing Strategies at Soccer, *Journal of the Royal Statistical Society. Series D (The Statistician)*, 46(4), pp. 541-550

Estimation of race scenarios in standard triathlon according to level density

Thibaut Ledanois[1,2] & Chloé Fouillot[1]

[1] *Institut National du sport de l'exercice et de la performance, 11 avenue du Tremblay Paris 75012*
[2] *Fédération Française de Triathlon, 2 rue de la justice 93213 Saint-Denis la plaine*
thibaut.ledanois@insep.fr
chloe.fouillot@insep.fr

Résumé. Le triathlon est un sport jeune dans les Jeux Olympiques, le mouvement stratégique et le rythme de la compétition sont différents selon les étapes, les saisons et le sexe. L'objectif de cette étude est de prédire les scénarios de course dans le triathlon standard (1,5 km de natation - 40 km de vélo - 10 km de course à pied) avec drafting au niveau international en utilisant des approches de méthode de clustering. Le point de départ était d'estimer le niveau de chaque triathlète par un modèle d'évaluation de Stephenson. Des analyses non supervisées, supervisées et statistiques sur les données historiques des résultats de compétition des hommes et des femmes de l'Union internationale de triathlon depuis 2009 à 2019 sont réalisées. Dans un premier temps, les résultats montrent une différence de niveau et de densité entre le niveau international et le niveau continental. Deuxièmement, nos résultats ont démontré une relation entre le niveau de densité du champ de départ des athlètes et la probabilité de tendre vers la natation groupée (indice de dispersion = 189.06, 176.47), la natation dispersée (indice de dispersion = 227.48, 204.42) et la natation avec échappée (indice de dispersion = 229.52, 219.89). En conclusion, cette étude fournit des informations aux athlètes et à l'encadrement avant la course en fonction de la liste de départ pour mieux préparer leur stratégie.
Mots-clés. Triathlon, classement, Performance, analyse hiérarchique

Abstract. The triathlon is a young sport in the Olympic Games, the move strategic and pacing competition is different by stages, by seasons and by gender. The aim of this study is to predict scenario patterns of race in standard triathlon (1,5 Km swim – 40 Km bike – 10 Km run) with drafting at the international level using clustering method approaches. The starting point was to estimate the level of each triathlete by a Stephenson rating model. Unsupervised, supervised, and statistical analyses on the men's and women's historical data competition results of International Triathlon Union since 2009 to 2019 are carried out. Two level of event and three race patterns for each gender at the swim section exit are defined. In first time results show a difference in levels and the density between the international level and the continental level. Secondly, our results demonstrated a relationship between the level density of the starting field of athletes and the probability to tend towards grouped swimming (*index rating dispersion* = 189.06, 176.47), swimming scattered (*index rating dispersion* = 227.48, 204.42) and swimming with a breakaway (*index rating dispersion* = 229.52, 219.89). In conclusion, this study provides information to the athletes and the coaching staff before the race according to the start list to better prepare their strategy.
Keywords. Triathlon, Rating, Sport Performance, Cluster Analysis

1. Intro

In order to investigate opponents force density at the start of a triathlon, Stephenson's scoring could be useful, based on wins made in a duel over a whole career (Stephenson and Sonas, 2016). Stephenson's rating is inspired by different models already implemented in chess (Elo, 1978) and in tennis (Glickman, 1999) to estimate the probability of an outcome in a confrontation. With the help of this reliable indicator in time, the idea is to answer the question who has the most probability of winning in a future confrontation by considering a triathlete in a duel against each opponent during the race.
The aim of this study was therefore to quantify the level of competition in a triathlon during the swimming part in order to define the physiognomy of the race scenarios.

2. Méthode

2.1.Dataset

This analysis included data from elite athletes competing from 2009 to 2019. All performances in ITU on 708 races, in Standard format (Men: 469, Women: 239). This represents 27 352 performances, realized by 3873 men and 2924 women performers respectively. The ITU race data is publicly available and was extracted from the World Triathlon database (www.triathlon.org).

2.2.Level's estimation with a classification system

In order to identify the best system of probabilistic ratings, three methods are tested (Elo, 1978; Glickman, 1999, Stephenson and Sonas, 2016). The final rankings of the races in the dataset are converted into the results of pairwise duels: an athlete finishing 4th is considered to have lost 3 duels against the top 3 in the race and to have won all duels against competitors ranked lower than him/her. DSQs and DNFs are regarded as having come last and been bested by all other participants.

Three probabilistic systems are being tried: Elo, Glicko and Stephenson. They were all implemented using the PlayerRatings package (Stephenson and Sonas, 2016).The optimization of the parameters of each model was performed using the L-BFGS-B algorithm available in the *optim* function of the R language, in order to minimize the logloss on the validation set.

2.3. Assignment of scenarios

The description of the scenarios is based on the definition of a pack in triathlon. First, two athletes are considered to be in the same pack if they are less than 4 seconds apart. Secondly, a peloton consists of at least two athletes. (Vleck et al., 2008).
In order to objectively define the membership of races to a scenario class, we first preprocessed the dataset, in two steps:
(i) The implementation of an unsupervised clustering, in order to decrease the part of subjectivity in the assignment of race scenarios. This clustering was performed on the 72 races representing the highest international level. The number of clusters (K=6) is set to optimal with the Gap statistic method (Tibshirani & al, 2001). (ii) The idea is to regroup the most similar clusters, using the dendrogram resulting from the hierarchical ascending classification, in order to form only three groups corresponding to the three types of scenarios defined empirically by the experts of the discipline.

(iii) Then labeling using the previously identified scenarios to categorize the remaining 637 races via the Safe Semi-Supervised Support Vector Machine (S4VM) algorithm (Yu-Feng Li and Zhi-Hua Zhou, 2015).

Swimming rating dispersion

The first idea was to used swim rating of athletes registered on races start list, to construct an indicator I which can explain the swim scenario. To build this indicator, we first compute the median swim rating of the race using swim ratings of each athlete registered on the start list. Then, for each athlete which have a swim rating greater than the median swim rating of the race, the difference to this median swim rating is computed. The indicator I is therefore the mean of these differences. We can mathematically express I as follows:

$$I = \frac{1}{m}\sum_{i=1}^{n}(R_i - R)^2 \times 1_{\{R_i > R\}}$$

with:

n the number of athletes registered on the start list

R_i the swim rating of athlete i

R the median rating of the race

$m = \sum_{i=1}^{n} 1_{\{R_i > R\}}$

2.4. Statistical analysis

The statistical analyses were performed separately between men's and women's races. The significance level was set at 5% for all the tests. The probabilities of achieving each scenario were compared by the Z test of comparison of proportions. The Kruskal-Wallis test is used in order to compare the distributions of the dispersion index of the rating previously calculated according to the swimming scenario carried out. If it is significant, Conover's post hoc test of two-to-two comparison of distributions is then used. All data are processed and analyzed R v.1.4.1106

3. Résultats

3.1.Descriptive analysis

The number of male athletes present at the start of the races are mostly similar regardless of the level of competition, however, the continental cups are significantly different with a lower number than the other races types. These results are identical for women with no significant difference in the number of starters for all competitions. Moreover, for the same level of competition, there is a tendency to have relatively fewer women than men at the start of the races, but this difference is not significant. (p-value > 0.05).

The race times have a significant difference between the WChamp - OG/TE - WTS levels compared to the ConCup level races for both men and women (6697 ± 260.12 / 7324 ± 194,56). But the difference in race times varies by stage for men and women, statistically illustrated a difference between WTS and ConChamp for men and not for female triathletes (Table1). However, the comparison of race times between men and women are significantly different independently of the stage. (p-value>0.05).

	Hommes						*Femmes*					
	WChamp	*OG/TE*	*WTS*	*WCup*	*ConChamp*	*ConCup*	*WChamp*	*OG/TE*	*WTS*	*WCup*	*ConChamp*	*ConCup*
N	51.17 ±	52.2 ± 5.67	46.25 ± 8.68[f]	45.15 ±	41.62± 11.14	37.12± 12.52[a,c,]	38.42 ±	45.6 ± 8.08	43.22 ±	38.85 ±	39.17 ±	37.28 ±

	11.84[f]			13.41[f]		[d]	13.10		14.98	13.01	14.33	12.96
Total Time (sec)	6704 ± 314.08 f	6687 ± 174.99 f	6700 ± 278.61 d,e,f	6834 ± 371.29 c,f	6901 ± 514.09 c	7058 ± 484.53 a,b,c,d	7393 ± 333.52 f	7174 ± 502.70 f	7404 ± 349.34 d,f	7576 ± 504.36 c,f	7583 ± 479.53	7881 ± 566.42 a,b,c,d
Athlete gaps total time (sec)	12.87 ± 24.41 d,e,f	13.37 ± 38.12 e,f	13.85 ± 28.99 d,e,f	17.09 ± 39.39 a,c,e,f	33.53 ± 105.98 a,b,c,d, f	35.30 ± 81.04 a,b,c,d, e	16.08 ± 25.25 d,e,f	15.02 ± 19.20 d,e,f	17.67± 27.52 d,e,f	27.01 ± 162.25 a,b,c,e, f	36.73 ± 76.20 a,b,c,d, f	50.82 ± 84.57 a,b,c,d, e
Stephenson rating	1982.8 ± 129.9 d,e,f	1987.8 ± 151.4 e,f	1951.6 ± 102.7 d,e,f	1785.5 ± 109.8 a,c,f	1726.3 ± 120 a,b,c,f	1618.9 ± 64.57 a,b,c,d, e	2124.8 ±206.2 d,e,f	2116.4 ± 217.4 d,e,f	2088.1 ± 178.1 d,e,f	1849.3 ± 155.5 a,b,c,f	1831.3 ± 157.2 a,b,c,f	1646.5 ± 76.2 a,b,c,d, e

Table 1 - Descriptive table of race datas by level stage and gender significant difference (a:with WChamp, b: with JO/TQO, c: with WTS, d: with WCup, e:with ConChamp, f: with ConCup

The analysis of the averages of the time difference to the first allows to observe once again for men and women a significant difference between WTS / OG / TE / WChamp versus Wcup/ConCup/ConChamp, except between OG / WCup for men. Moreover we can notice statistical differences between Wcup, ConCup and ConChamp regardless of gender. The comparison between the female and male races shows that the differences to the first are significantly different with smaller differences in men, this difference increases when the level of the set decreases.

The observation of the Stephenson rating as an indicator of the level of competition shows a higher competitiveness in international races compared to continental races following a decreasing gradient. The rating is an estimator of the level upstream of the race allowing to discriminate the latter between them and in an equivalent or better way than the other tested variables, independently of the gender.

A division into two groups of courses, high level with a constant representativeness of the triathletes force' and the other race group with more variable and lesser level between the races are computable.

The difference in rating between the sexes is statistically different (p-value < 0.05) with a higher average for each level and more for the international races WTS / OG / TE / WChamp.

3.2. Clustering of race scenarios and propagation of labels

The averages resulting from the label propagation allows us to synthesize the number of scenarios and to match the qualitative definitions stated in the introduction. Moreover, the averages of the number of favorites outside the top group remain similar between the two hierarchical clustering and SV4M.

This clustering makes it possible to highlight (*Table 3*) a low number of races with a grouped swimming outing, particularly among women (5.02%). Moreover, we observe that one out of two female races is classified with a breakaway (49,37%) while the highest proportion among men is for the swimming scattered scenario (48,83%). The association with the race level should allow to understand the density of the level on a race and thus the probable scenario.

Cluster	*% pack n°1*	*% pack n°2*	*% pack n°3*	*% pack n°4*	*% pack n°5*	*% after pack n°5*	*Top 10 favorites outside pack n°1*	*N race (%)*
grouped swimming	84.10	4.77	1.69	0.41	0	2.06	0.34	♀ 12 (5,02%) ♂ 50 (10,66%)
swimming scattered	35.69	21.04	13.13	5.64	2.34	7.57	3.04	♀ 109 (45,61%) ♂ 229 (48,83%)
swimming with a breakaway	9.52	30.36	18.73	11.33	5.85	11.82	7.05	♀ 118 (49,37%) ♂ 190 (40,51%)

Table 3 - Semi-Supervised (Classification) by the method S4VM with K=3 (grouped swimming, swimming scattered and swimming with a breakaway)

3.3. Comparison of rating distributions by scenario cluster

The races having been attributed to a swimming scenario, the idea is to present an indicator constructed using the swimming classification which can explain the realization of the scenario in question. The Kruskal-Wallis test demonstrates significant differences between the distributions of the dispersion index according to the created clustered groups for both women's races (p-value < 0.05) and men's races (p-value <0.05) (*Figure 2*).

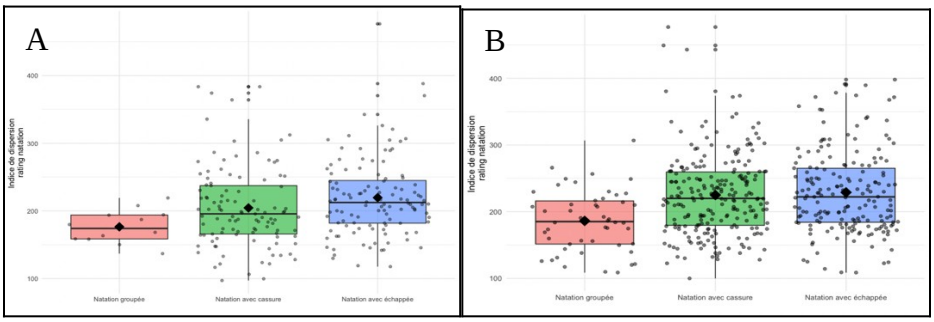


Figure 2 – Dispersion index rating for women's (A) and men's (B) races according to the swimming scenario

For women's races, a significant difference (p-value < 0.05) in the distribution between *grouped races* 176.47 ± 24.66 and breakaway races 219.55 ± 57.92 are observed (Figure 2 B). As well as between races with breaks 204.42 ± 59.49 and *scattered* races 219.55 ± 57.92 (p-value < 0.05) (Figure 1). However races breakaway and grouped races have a test statistic which is not significant but nevertheless remains close to the 5% threshold (p-value = 0.069). The Conover test indicates significant differences for men's races, between grouped races 186.19 ± 45.55 and races with breakaways 228.94 ± 60.87 (p < 0.05) and between grouped races 186.19 ± 45.55 and scattered ones 224.94 ± 60.38 (p < 0.05).

3.4. Scenario race probability

Comparison of proportions tests show that at the end of the first transition, for women, the probability of observing a race *scattered P = 0.472* is not significantly different from the probability of observing a race with a *breakaway P = 0.487* (p-value = 0.752). The probability of observing a *scattered race P = 0.472* is significantly greater than the probability of observing a *grouped race P = 0.04* (p-value < 0.05). The probability of observing a race with a *breakaway P = 0.487* is significantly greater than the probability of observing a grouped race *P = 0.04* (p-value < 0.05). For men's races, the probability of observing a *scattered race P = 0.496* is significantly greater than the probability of observing a race with a *breakaway P = 0.397* (p-value < 0.05). The probability of observing a *scattered race P = 0.496* is significantly higher than the probability of observing a *grouped race P = 0.106* (p-value < 0.05). It is significantly more likely to observe a *breakaway race P = 0.397* than a *grouped race P = 0.106* (p-value < 0.05). As well, the proportion of having a *grouped race* on men's races *P = 0.106* is significantly higher than the probability of having a *grouped race* on women's races P = 0.040 (p-value < 0.05).

4. Discussion

The main results of this study are: (i) the difference in the density of competitiveness between the multiple international and continental competition stages, (ii) the classification between these multiple competitions into three race scenarios and (iii) the prediction after the swim section exit of which race scenario the race will be. Our main practical finding reveals the possibility of changing one's pace and race strategy according to the start list, depending on the profile of the present athletes. Strategy could be adapted in function of the probability of the race cluster: to create gaps quickly to oust certain opponents and in another configuration be in a more passive strategy until the running phase. These race decisions will be directly linked to the final performance which is itself linked to the finality of the sports career, qualifications for the Olympic Games, world champion title or Olympic title (Rüst et al., 2013).

5. Conclusion

This study identifies the different performance scenarios in the standard distance triathlon by estimating the level of each athlete. The results agree with several elements known in the literature, in particular in swimming, where the probability of bringing the race to two possible situations according to the pacing and the positioning is shown by bringing the notion of density of levels at the start of the race. In particular, the probability of having a breakaway in women's races is higher than the same probability in men's races, and in particular that it is victorious at the end. This study presents a novelty solution for the assistance to the coaching staff and the athletes before the competitions and the analyze performance on swimming with the context of the race.

Bibliographie

Elo, A. E. (1978). *The rating of chessplayers, past and present*. New York: Arco Pub.

Glickman, M. E. (1999). Parameter Estimation in Large Dynamic Paired Comparison Experiments. *J R Statist Soc C* 48, 377–394. doi: 10.1111/1467-9876.00159.

Rüst, C. A., Lepers, R., Stiefel, M., Rosemann, T., and Knechtle, B. (2013). Performance in Olympic triathlon: changes in performance of elite female and male triathletes in the ITU World Triathlon Series from 2009 to 2012. *SpringerPlus* 2, 685. doi: 10.1186/2193-1801-2-685.

Vleck, V. E., Bentley, D. J., Millet, G. P., and Bürgi, A. (2008). Pacing during an elite Olympic distance triathlon: Comparison between male and female competitors. *Journal of Science and Medicine in Sport* 11, 424–432. doi: 10.1016/j.jsams.2007.01.006.

Yu-Feng Li and Zhi-Hua Zhou (2015). Towards Making Unlabeled Data Never Hurt. *IEEE Trans. Pattern Anal. Mach. Intell.* 37, 175–188. doi: 10.1109/TPAMI.2014.2299812.

Statistique mathématique

ESTIMATEURS NON PARAMÉTRIQUES À NOYAU GÉNÉRAL BIRNBAUM-SAUNDERS DES DENSITÉS MULTIVARIÉES À SUPPORT NON NÉGATIFS

Smail Adjabi [1] & Nabil Zougab [1] & Lynda Harfouche [1] & Yasmina Ziane [1]
[1] *Unité de Recherche LaMOS, Université de Béjaia, Algérie*
E-mail : adjabi@hotmail.com, nabilzougab@yahoo.fr, lyndaharfouche@hotmail.com, yasmina-ziane@hotmail.com

Résumé

On présente l'estimateur à noyau Général Birnbaum-Saunders (GBS) de la fonction densité multivariée des données non négatives et ses propriétés statistiques (biais, variance et erreur quadratique moyenne intégrée). Pour réduire le biais de cet estimateur, on applique la technique de correction du biais notée MBC(Multiplicative Biais Correction) de Jones-Linton-Nielsen (1995). On obtient ainsi un nouvel estimateur et on donne ses propriétés statistiques. Pour étudier la qualité de l'estimateur obtenu, des études sont menées sur des données simulées à partir de densités cibles bivariées connues et sur des données réelles en utilisant le critère : biais quadratique intégré (ISB).
Mots-clés. Estimation non paramétrique, Noyau Général Birnbaum-Saunders (GBS), Technique multiplicative de correction du biais (MBC), biais quadratique intégré (ISB).

Abstract

We propose the nonparametric General Birnbaum-Saunders (GBS) kernel estimator for the estimation of the multivariate density function of non-megative data. To reduce the bias of the estimator thus obtained, we apply the bias correction technique noted MBC (multiplicative bias correction) of Jones-Linton-Nielsen. We thus obtain a new estimator and we give its statistical propreties (bias, variance and integrated mean square error). To justify the quality of the estimator obtained, studies are carried out on simulated data from know two dimensional target densities and on real data.
Keywords. Nonparametric estimation, General Birnbaum-Saunders (GBS) kernel, Multiplicative bias correction technique (MBC), integrated square bias (ISB).

Introduction

Dans l'estimation de la fonction densité, l'utilisation de noyaux asymétriques est apparue comme une solution aux phénomènes de biais aux bords, qui sont générés par l'emploi de noyaux symétriques (standards) de Parzen (1962) lorsque le support de la densité est limité (par exemple dans le cas de données non négatives). Ce problème est du à l'utilisation de noyaux symétriques qui assignent un poids en dehors du support lorsque le lissage est pris en compte près du bord. Ce problème devient plus sévère dans le cas multivarié car il s'ajoute au problème de dimension du support. Une première solution à ce problème dans le cas multivarié a été donné par Bouezmarni et Rombouts (2010) en utilisant le noyau produit composé des noyaux univariés gamma et beta introduits par Chen (1999, 2000). Dans ce travail (Zougab et al (2018)), on propose d'utiliser le

noyau GBS (General Birnbaum-Saunders) qui comprend les noyaux : BS-classique, BS-Power-Exponentiel (BS-PE) et BS-Student (BS-t). Ce noyau s'adapte à beaucoup de types de données particulièrement les données du type Heavy Tailed (à queues lourdes). Pour corriger le biais de l'estimateur obtenu, on applique la technique MBC (Multiplicative Bias Correction) de Jones, Linton et Nielsen (1995). On obtient un nouvel estimateur de la densité multivarié qu'on notera MBC-GBS et on présente ses propriétés statistiques. Pour comparer l'estimateur MBC-GBS avec l'estimateur GBS, on utilise le critère : biais quadratique intégré (ISB) sur des données simulées à partir de densités cibles connues bivariées et sur des données réelles.

1 Estimateur multivarié à noyau GBS

Soit $\mathbf{X}_1, \ldots, \mathbf{X}_n$ des vecteurs de variables aléatoires indépendantes et identiquement distribuées de densité inconnue f définie sur $\mathbb{T}^d$, où $\mathbb{T} = (0, \infty)$. L'estimateur à noyau produit GBS (General Birnbaum-Saunders) s'écrit

$$\widehat{f}_{GBS}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\prod_{j=1}^{d} K_{GBS(h_j,,x_j;g)}(X_{ij}), \tag{1}$$

où $\mathbf{X}_i = (X_{i1}, \ldots, X_{id})^\top$, $i = 1, \ldots, n$, $\mathbf{x} = (x_1, \ldots, x_d)^\top$ est le vecteur cible et $\mathbf{h} = (h_1, \ldots, h_d)^\top$ est le vecteur des paramètres de lissage. Le noyau GBS proposé par Marchant et al (2013) est de la forme

$$K_{GBS(h,x;g)}(y) = cg\left(\frac{1}{h}\left(\frac{y}{x} + \frac{x}{y} - 2\right)\right)\frac{1}{2\sqrt{h}}\left(\frac{1}{\sqrt{xy}} + \sqrt{\frac{x}{y^3}}\right),\ x > 0,\ y > 0,\ h > 0,$$

où

TABLE 1 – $c = 1/\int_{-\infty}^{\infty} g(y^2)dy$, $c_{g^2} = 1/\int_{-\infty}^{\infty} g^2(y^2)dy$, $u_1(g)$ et $u_2(g)$ moments d'ordre 1 et 2 des distributions

Distribution	c	c_{g^2}	$g = g(u),\ u > 0$	$u_1(g)$	$u_2(g)$
BS-Classique	$\frac{1}{\sqrt{2\pi}}$	$\frac{1}{\sqrt{\pi}}$	$\exp\left(-\frac{1}{2}u\right)$	1	3
BS-PE	$\frac{\nu}{2^{\frac{1}{2\nu}}\Gamma\left(\frac{1}{2\nu}\right)}$	$\frac{\nu}{\Gamma\left(\frac{1}{2\nu}\right)}$	$\exp\left(-\frac{1}{2}u^\nu\right),\ \nu > 0$	$\frac{2^{\frac{1}{\nu}}\Gamma\left(\frac{3}{2\nu}\right)}{\Gamma\left(\frac{1}{2\nu}\right)}$	$\frac{2^{\frac{2}{\nu}}\Gamma\left(\frac{5}{2\nu}\right)}{\Gamma\left(\frac{1}{2\nu}\right)}$
$BS-t$	$\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{\pi}\right)}$	$\frac{\nu\Gamma(\nu)}{\sqrt{\pi}\Gamma\left(\frac{1+2\nu}{\nu}\right)}$	$\left(1+\frac{u}{\nu}\right)^{-\frac{(\nu+1)}{2}},\ \nu > 0$	$\frac{\nu}{\nu-2},\ \nu > 2$	$\frac{3\nu^2}{(\nu-2)((\nu-4))},\ \nu > 4$

1.1 Propriétés de l'estimateur

On suppose que : 1. f est deux fois continue et admet des dérivées partielles.
2. $h_1 = \cdots = h_d = h$ dépendant de n et $\lim_{n\to\infty} h = 0$.
L'erreur quadratique moyenne intégrée (MISE) de l'estimateur est

$$\text{MISE}(\widehat{f}_{GBS}) = \int \text{biais}^2\left(\widehat{f}_{GBS}\right) + \int \text{Var}\left(\widehat{f}_{GBS}\right)$$

$$= \int_{\mathbb{T}^d} \frac{h^2 u_1^2(g)}{4} \left\{ \sum_{j=1}^{d} \left(x_j \frac{\partial f(\mathbf{x})}{\partial x_j} + \frac{\partial^2 f(\mathbf{x})}{\partial x_j^2} x_j^2 \right) \right\}^2 d\mathbf{x} + \int_{\mathbb{T}^d} \frac{f(\mathbf{x})}{nh^{d/2}} \prod_{j=1}^{d} \frac{c^2}{c_{g^2} x_j} d\mathbf{x} + o\left(h^2 + n^{-1}h^{-d/2}\right). \tag{2}$$

Le paramètre de lissage optimal qui minimise le MISE (2) est alors de la forme

$$h_{GBS}^{opt} = \left\{ \frac{d \int_{\mathbb{T}^d} f(\mathbf{x}) \prod_{j=1}^{d} \frac{c^2}{c_{g^2} x_j} d\mathbf{x}}{2 \int_{\mathbb{T}^d} \left\{ \sum_{j=1}^{d} \frac{u_1(g)}{2} \left(x_j \frac{\partial f(\mathbf{x})}{\partial x_j} + \frac{\partial^2 f(\mathbf{x})}{\partial x_j^2} x_j^2 \right) \right\}^2 d\mathbf{x}} \right\}^{2/(4+d)} n^{-2/(4+d)}. \tag{3}$$

2 Estimateur multivarié MBC à noyau GBS

On applique la technique MBC (Multiplicative Bias Reduction) pour l'estimateur de la fonction densité à noyau asymétrique GBS. Cette technique proposée par Jones et al (1995) est basée sur l'expression $f(\mathbf{x}) = \widehat{f}(\mathbf{x})\{f(\mathbf{x})/\widehat{f}(\mathbf{x})\}$, où l'on estime le terme biais correction $f(\mathbf{x})/\widehat{f}(\mathbf{x})$. L'estimateur est alors de la forme :

$$\widehat{f}_{MBC-GBS}(\mathbf{x}) = \widehat{f}_{GBS}(\mathbf{x}) \left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{\prod_{j=1}^{d} K_{GBS(h,x_j;g)}(X_{ij})}{\widehat{f}_{GBS}(\mathbf{X}_i)} \right\}. \tag{4}$$

2.1 Propriétés de l'estimateur

On suppose que 1. f est quatre fois continue et admet des dérivées partielles.
2. $h_1 = \cdots = h_d = h$ dépendant de n, $\lim_{n\to\infty} h = 0$ et $\lim_{n\to\infty} nh^{(3d/2)+2} = \infty$.
L'erreur quadratique moyenne intégrée (MISE) de l'estimateur est

$$\begin{aligned} \text{MISE}(\widehat{f}_{MBC-GBS}) &= \int \text{biais}^2 \left(\widehat{f}_{MBC-GBS} \right) + \int \text{Var} \left(\widehat{f}_{MBC-GBS} \right) \\ &= h^4 \int_{\mathbb{T}^d} f^2(\mathbf{x}) \left\{ \sum_{j=1}^{d} \left[\frac{1}{2} x_j u_1(g) \frac{\partial q(\mathbf{x})}{\partial x_j} + \frac{1}{2} x_j^2 u_1(g) \frac{\partial^2 q(\mathbf{x})}{\partial x_j^2} \right] \right\}^2 d\mathbf{x} \\ &+ \frac{c^{2d}}{c_{g^2}^d nh^{d/2}} \int_{\mathbb{T}^d} \frac{f(\mathbf{x})}{\prod_{j=1}^{d} x_j} d\mathbf{x} + o\left(h^4 + \frac{1}{nh^{d/2}} \right). \end{aligned} \tag{5}$$

où $q(\mathbf{x}) = l_1(\mathbf{x})/f(\mathbf{x})$ avec $l_1(\mathbf{x}) = \frac{1}{2} \sum_{j=1}^{d} x_j u_1(g) \frac{\partial f(\mathbf{x})}{\partial x_j} + \frac{1}{2} \sum_{j=1}^{d} x_j^2 u_1(g) \frac{\partial^2 f(\mathbf{x})}{\partial x_j^2}$ et $l_2(\mathbf{x}) = \frac{1}{4} \sum_{j=1}^{d} x_j^2 u_2(g) \frac{\partial^2 f(\mathbf{x})}{\partial x_j^2} +$

$\frac{1}{4} \sum_{j=1}^{d} x_j^3 u_1(g) \frac{\partial^3 f(\mathbf{x})}{\partial x_j^3} + \frac{1}{24} \sum_{j=1}^{d} x_j^4 u_2(g) \frac{\partial^4 f(\mathbf{x})}{\partial x_j^4} + \frac{1}{4} \sum_{k\neq j} x_j x_k u_1^2(g) \frac{\partial^2 f(\mathbf{x})}{\partial x_j \partial x_k} + \frac{1}{4} \sum_{k\neq j} x_j x_k^2 u_1^2(g) \frac{\partial^3 f(\mathbf{x})}{\partial x_j \partial x_k^2} + \frac{1}{4} \sum_{k\neq j} x_j^2 x_k^2 u_1^2(g) \frac{\partial^4 f(\mathbf{x})}{\partial x_j^2 \partial x_k^2}$

En minimisant le MISE (5) par rapport à h, on obtient le paramètre de lissage optimal

$$h^{opt}_{MBC-GBS} = \left\{ d \frac{c^{2d} \int_{\mathbb{T}^d} \frac{f(\mathbf{x})}{\prod_{j=1}^{d} x_j} d\mathbf{x}}{8 c_{g^2}^d \int_{\mathbb{T}^d} f^2(\mathbf{x}) \left\{ \sum_{j=1}^{d} \left[\frac{1}{2} x_j u_1(g) \frac{\partial q(\mathbf{x})}{\partial x_j} + \frac{1}{2} x_j^2 u_1(g) \frac{\partial^2 q(\mathbf{x})}{\partial x_j^2} \right] \right\}^2 d\mathbf{x}} \right\}^{2/(8+d)} n^{-2/(8+d)}. \tag{6}$$

Remarque. On constate que la technique MBC de réduction du biais améliore l'ordre du MISE de l'estimateur de $o\left(h^2 + \frac{1}{nh^{d/2}}\right)$ à $o\left(h^4 + \frac{1}{nh^{d/2}}\right)$.

3 Résultats numériques

3.1 Etude de Simulation

On examine et on compare sur des données simulées à partir de densités cibles bivariées connues à support non négatif les performances des estimateurs GBS et MBC-GBS en utilisant le critère biais quadratique intégré $ISB := \int\int \left[(\mathbb{E}\widehat{f}(\mathbf{x}))f(\mathbf{x}) \right]^2 dx$. Les noyaux utilisés sont les noyaux $BS - PE(\nu = 2)$ et $BS - t(\nu = 5)$. Le paramètre de lissage est choisi en utilisant la méthode Rule of Thumb (RT) de Siverman (1986) qui consiste à remplacer dans h^{opt} (3) et (6) la densité inconnue f par la loi inverse gamma de paramètres (1,1) qui possède le même support que la densité à estimer. Les densités cibles bivariées sont :
D1 : Densité bivariée de Burr : Produit de deux distributions indépendantes de Burr

$$f(\mathbf{x}) = f(x_1, x_2) = \frac{3x_1^2}{(1+x_1^3)^2} \times \frac{3x_2^2}{(1+x_2^3)^2},\ x_1 > 0,\ x_2 > 0.$$

D2 : Distribution dépendante bivariée Birnbaum-Suanders (BS)

$$\begin{aligned} f(\mathbf{x}) = f(x_1, x_2) &= \frac{1}{16} \phi_2 \left(\left(\sqrt{\frac{x_1}{2}} - \sqrt{\frac{2}{x_1}} \right), \left(\sqrt{\frac{x_2}{2}} - \sqrt{\frac{2}{x_2}} \right); \Gamma \right) \\ &\times \left\{ \left(\frac{2}{x_1} \right)^{\frac{1}{2}} + \left(\frac{2}{x_1} \right)^{\frac{3}{2}} \right\} \left\{ \left(\frac{2}{x_2} \right)^{\frac{1}{2}} + \left(\frac{2}{x_2} \right)^{\frac{3}{2}} \right\},\ x_1 > 0,\ x_2 > 0. \end{aligned}$$

où $\phi_2(u_1, u_2; \Gamma)$ est la loi bivariée gaussienne de matrice de corrélation $\Gamma = \begin{pmatrix} 1 & -0.9 \\ -0.9 & 1 \end{pmatrix}$
A partir des Tables 2 et 3, on constate que :
- La valeur moyenne du biais quadratique intégrée (ISB) diminue avec l'augmentation de la taille de l'échantillon pour les deux estimateurs et ceci quel que soit le modèle.
- L'estimateur MBC-GBS est toujours meilleur au sens du critère ISB pour les deux modèles que l'estimateur GBS et ceci quel que soit la taille de l'échantillon.

TABLE 2 – Valeurs de ISB moyen basées sur 500 replications pour le modèle de Burr.

Noyau	Estimateur	$n = 50$	$n = 100$	$n = 200$
BS-PE($\nu = 2$)	$\widehat{f}_{GBS}$	0.01343	0.001111	0.00926
	$\widehat{f}_{MBC-GBS}$	0.01081	0.00846	0.00781
BS-t($\nu = 5$)	$\widehat{f}_{GBS}$	0.04842	0.04029	0.03420
	$\widehat{f}_{MBC-GBS}$	0.04128	0.03524	0.03073

TABLE 3 – Valeurs de ISB moyen basées sur 500 replications pour le modèle dépendant BS.

Noyau	Estimateur	$n = 50$	$n = 100$	$n = 200$
BS-PE($\nu = 2$)	$\widehat{f}_{GBS}$	0.02499	0.00128	0.00124
	$\widehat{f}_{MBC-GBS}$	0.02041	0.00088	0.00069
BS-t($\nu = 5$)	$\widehat{f}_{GBS}$	0.03239	0.00182	0.00138
	$\widehat{f}_{MBC-GBS}$	0.02810	0.00110	0.00096

3.2 Application sur des données réelles

Nous considérons les données réelles bivariées du minérale osseuse (BMD) mesurées en g/cm^3 pour 24 individus. La première variable X_1 représente la DMO de l'os Dominant Radius avant de commencer l'étude. La deuxième variable X_2 représente la DMO du même os une année après. On observe que les données considérées sont non négatives, par conséquent le noyau $GBS(BS - PE(\nu = 2))$ est approprié pour estimer la distribution de ce couple de données. Le paramètre de lissage est choisi par la méthode Rule of Thumb (RT).
Sur la figure 1 des courbes de surface, on constate que le lissage est meilleur quand on estime ces données par l'estimateur MBC-GBS.

Conclusion

Nous avons présenté l'estimateur à noyau GBS produit pour l'estimation de la fonction densité multivariée à support non négatif et nous avons établi ses propriétés statistiques (Biais, variance et erreur globale : MISE). Nous avons ensuite appliqué la technique MBC de Jone-Linton-Nielsen de réduction du biais sur cet estimateur, ce qui nous a permis de construire un nouvel estimateur de la fonction densité multivariée et nous avons donné ses propriétés statistiques. Nous avons montré que cette technique améliore l'ordre de grandeur du biais de $o(h)$ à $o(h^2)$. Les résultats numériques sur des données simulées à partir de densités cibles bivariées connues et sur des données réelles montrent la bonne performance de l'estimateur MBC-GBS au sens du critère biais quadratique intégrée (ISB).

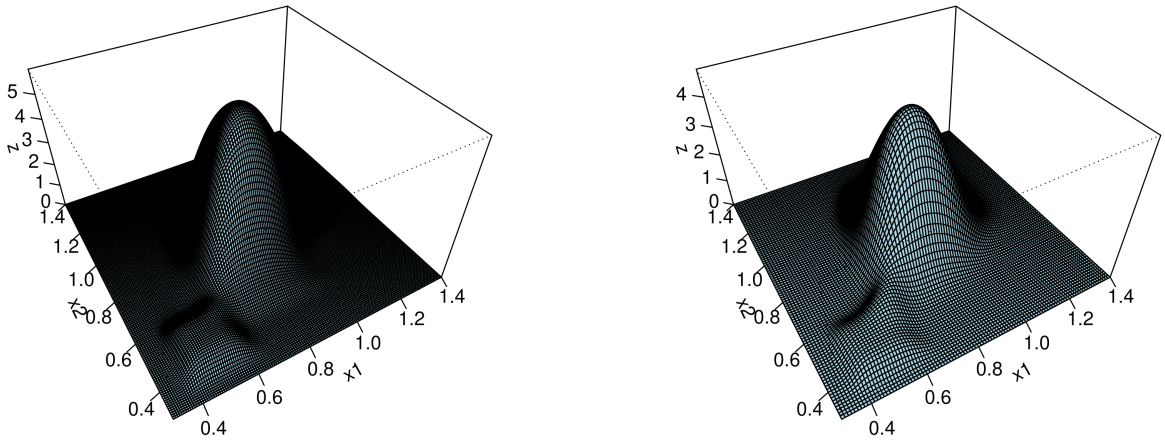


Figure 1 : Courbe des surfaces des estimateurs GBS et MBC-GBS des données réelles BMD

Bibliographie

Bouezmarni, T. and Rombouts, J.V.K. (2010). Nonparametric density estimation for multivariate bounded data, *Journal of Statistical Planning and Inference*, 140, 139–152.

Chen, S. X. (1999). Beta kernels estimators for density functions, *Computational Statistics and Data Analysis*, 31, 131–145.

Chen, S. X. (2000). Gamma kernel estimators for density functions, *Annals of the Institute of Statistical Mathematics*, 52, 471–480.

Funke, B. and Kawka, R. (2015). Nonparametric density estimation for multivariate bounded data using two non-negative multiplicative bias correction methods, *Computational Statistics and Data Analysis*, 92, 148–162.

Jones, M.C. Linton, O. and Nielsen, J.P. (1995). A simple bias reduction method for density estimation, *Biometrika*, 82, 327–338.

Marchant, C. Bertin, K. Leiva, V. and Saulo, H. (2013). Generalized Birnbaum-Saunders kernel density estimators and an analysis of financial data, *Computational Statistics and Data Analysis*, 63, 1–15.

Parzen, E. (1962). On estimation of a probability density function and mode, *Annals of Mathematical Statistics*, 33, 1065–1076.

Silverman, B.W. (1986). Density estimation for statistic and data analysis, *Chapman ans Hall*, New York.

Zougab, N. Harfouche, L. Ziane, Y. and Adjabi, S. (2018). Multiplicative generalized Birnbaum-Saunders kernel density estimation. *Communication in Statistics- Theory and Methods], 47 :18, 4534–4555.*

Integration of bounded monotone functions: Revisiting the nonsequential case, with a focus on unbiased Monte Carlo (randomized) methods

Subhasish Basak[1,2,⋆] & Julien Bect[2] & Emmanuel Vazquez[2]

[1] *ANSES, Maison-Alfort, France.* ⋆*E-mail :* subhasish.basak@centralesupelec.fr
[2] *Université Paris-Saclay, CNRS, CentraleSupélec, Laboratoire des signaux et systèmes, Gif-sur-Yvette, France.*

Résumé. Dans cet article, nous revisitons le problème de l'intégration numérique d'une fonction monotone bornée, en nous concentrant sur la classe des méthodes de Monte Carlo non séquentielles. Nous établissons dans un premier une borne inférieure pour l'erreur maximale dans L^p d'un algorithme non séquentiel, qui généralise pour $p > 1$ un théorème de Novak. Nous étudions ensuite, dans le cas $p = 2$, l'erreur maximale de deux méthodes sans biais—une méthode fondée sur l'utilisation d'une variable de contrôle, et la méthode de l'échantillonnage stratifié.

Mots-clés. Intégration monotone, Monte Carlo, échantillonnage stratifié, échantillonnage hypercube latin, variable de contrôle, complexité

Abstract. In this article we revisit the problem of numerical integration for monotone bounded functions, with a focus on the class of nonsequential Monte Carlo methods. We first provide new a lower bound on the maximal L^p error of nonsequential algorithms, improving upon a theorem of Novak when $p > 1$. Then we concentrate on the case $p = 2$ and study the maximal error of two unbiased methods—namely, a method based on the control variate technique, and the stratified sampling method.

Keywords. Monotone integration, Monte Carlo, stratified sampling, Latin hypercube sampling, control variate, information-based complexity

1 Introduction

We address in this article the problem of constructing a numerical approximation of the expectation $\mathbb{E}(g(Y)) = \int g(y)\,\mathrm{P}_Y(\mathrm{d}y)$, where Y is a real random variable with known distribution P_Y, and g is a real function that is bounded and monotone. Such a problem occurs naturally in applications where one is interested in computing a risk using a model that provides an increasing conditional risk $g(Y)$ with respect to some random variable Y. This situation occurs for instance in the field of food safety, with Y a dose of pathogen and $g(Y)$ the corresponding probability of food-borne illness (see, e.g., Perrin et al., 2014).

Assuming that the cumulative distribution function of Y is continuous, the problem reduces after change of variable and scaling to the computation of $S(f) = \mathbb{E}(f(X)) = \int_0^1 f(x)\,\mathrm{d}x$, where X is uniformly distributed on $[0,1]$ and f belongs to the class F of all non-decreasing functions defined on $[0,1]$ and taking values in $[0,1]$. We work in this article in a fixed sample-size setting, where the number n of evaluations of f to be performed is chosen beforehand.

This problem was first studied by Kiefer (1957), who proved that considering regularly-spaced evaluations at $x_i = i/(n+1)$, $1 \le i \le n$, and then using the trapezoidal integration rule assuming $f(0) = 0$ and $f(1) = 1$, is optimal in the worst-case sense among all deterministic—possibly sequential—methods, with maximal error $1/(2(n+1))$. Novak (1992) later studied Monte Carlo (a.k.a. randomized) methods, and established that sequential methods are better in this setting than nonsequential ones, with a minimax rate of $n^{-3/2}$ over F for the L^1 error. Novak's proof relies on the construction of a particular two-stage algorithm, using stratified sampling in the second stage.

This article revisits the nonsequential setting with a focus on unbiased Monte Carlo methods, which are a key building block for the construction of good (rate-optimal) sequential methods—as can be learned from the proof of Theorem 3 in Novak's article. Section 2 derives a lower bound for the maximal L^p-error of nonsequential methods, for any $p \ge 1$, which is a generalization of a result by Novak (1992) concerning the L^1 error. Sections 3 and 4 then study the maximal L^2 error (variance) of two simple unbiased methods, based respectively on the control variate technique and on stratification. Section 5 concludes the article with a discussion.

2 A lower bound for the maximal L^p error

A nonsequential (also called non-adaptive) Monte Carlo method first evaluates the function at n random points $X_1, \ldots, X_n$ in $[0,1]$, and then approximates the integral $S(f)$ using an estimator

$$\hat{S}_n(f) = \varphi\left(f(X_1), f(X_2), \ldots, f(X_n)\right), \tag{1}$$

where $\varphi : \mathbb{R}^n \to \mathbb{R}$ is a measurable function. A nonsequential method is thus defined by two ingredients: the distribution of $(X_1, \ldots, X_n)$ and the function φ. The worst-case L^p error of such a method over the class F is

$$e_p(\hat{S}_n) \;=\; \sup_{f \in F} \mathbb{E}\left(\left|S(f) - \hat{S}_n(f)\right|^p\right)^{1/p}. \tag{2}$$

Remark 1. The class of nonsequential Monte Carlo methods as usually defined in the literature also allows $\hat{S}_n$ to be randomized (i.e., allows φ to be a random function). We have not considered randomized estimators in our definition, however, since Rao-Blackwell's theorem implies that they do not help in this setting, for any convex loss function.

Novak (1992) proved that for any nonsequential Monte Carlo method with sample size n, the maximal L^1 error $e_1(\hat{S}_n)$ is greater or equal to $1/(8n)$. We generalize this result to the case of the L^p error.

Theorem 2.1. *For any nonsequential Monte Carlo methods with sample size n,*

$$e_p(\hat{S}_n) \;\geq\; \left(\frac{1}{2}\right)^{2+1/p} \frac{1}{n}\,.$$

Observe that Novak's lower bound is recovered for $p = 1$. Using Theorem 2.1 with $p = 2$ we can deduce of lower bound for the variance of unbiased nonsequential methods.

Corollary 2.2. *For any unbiased nonsequential Monte Carlo method with sample size n,*

$$\sup_{f\in F} \operatorname{var}\left(\hat{S}_n(f)\right) \;\geq\; \frac{1}{32n^2}.$$

Proof of Theorem 2.1. Consider a nonsequential Monte Carlo methods with evaluation points $X_1, \ldots, X_n$ and estimator $\hat{S}_n$. Divide the interval $[0,1]$ into $2n$ equal subintervals of length $1/(2n)$: then at least one of the subintervals, call it I, will contain no evaluation point with probability at least $1/2$. Now construct two functions $f_1, f_2 \in F$ that are both equal to zero on the left of I, equal to one on the right, and such that $f_1 = 1$ and $f_2 = 0$ on I. Then $S(f_1) - S(f_2) = 1/(2n)$, and $\hat{S}_n(f_1) = \hat{S}_n(f_2)$ on the event $A = \{\{X_1, X_2, \ldots, X_n\} \cap I = \emptyset\}$, since f_1 and f_2 coincide outside of I. It follows that

$$\begin{aligned}\left(e_p(\hat{S}_n)\right)^p \;&\geq\; \sup_{f\in\{f_1,f_2\}} \mathbb{E}\left(|S(f) - \hat{S}_n(f)|^p\right) \;\geq\; \frac{1}{2}\sum_{j=1}^{2} \mathbb{E}\left(|S(f_j) - \hat{S}_n(f_j)|^p\right) \\ &\geq\; \frac{1}{2}\sum_{j=1}^{2} \mathbb{E}(|S(f_j) - \hat{S}_n(f_j)|^p \cdot \mathbb{1}_A) \;=\; \frac{1}{2}\sum_{j=1}^{2} \mathbb{E}(|S(f_j) - T|^p \cdot \mathbb{1}_A),\end{aligned}$$

where T denotes the common value of $\hat{S}_n(f_1)$ and $\hat{S}_n(f_2)$ on A. We conclude that

$$\left(e_p(\hat{S}_n)\right)^p \;\geq\; \frac{1}{2}\,\frac{|I|^p}{2^{p-1}}\,P(A) \;\geq\; \left(\frac{1}{2}\right)^{2p+1} \cdot \frac{1}{n^p},$$

using the fact that, for any $a, b, x \in \mathbb{R}$ and $p \geq 1$, $|a - x|^p + |b - x|^p \geq |a - b|^p/2^{p-1}$. □

3 Uniform i.i.d. sampling

The simple Monte Carlo method is the most common example of a nonsequential method: the evaluation points $X_1, \ldots, X_n$ are drawn independently, uniformly in $[0,1]$, and then the integral is estimated by $\hat{S}_n^{\mathrm{MC}}(f) = \frac{1}{n}\sum_{i=1}^{n} f(X_i)$. The estimator is clearly unbiased,

and it follows from Popoviciu's inequality—i.e., $\mathrm{var}(Z) \leq 1/4$ for any random variable Z taking values in $[0,1]$—that

$$\left(e_2\left(\hat{S}_n^{\mathrm{MC}}\right)\right)^2 \;=\; \max_{f\in F} \mathrm{var}\left(\hat{S}_n^{\mathrm{MC}}(f)\right) \;=\; \frac{1}{4n}.$$

The maximal error is attained when f is a unit step function jumping at $x_0 = 1/2$. It turns out that a smaller error can be achieved, for the same (uniform i.i.d.) sampling scheme, using the control variate technique. More specifically, we consider the control variate $\tilde{f}(X_i) = X_i$ and set

$$\hat{S}_n^{\mathrm{cv}}(f) \;=\; \frac{1}{n}\sum_{i=1}^{n}\left(f(X_i) - \tilde{f}(X_i)\right) + \frac{1}{2}.$$

Theorem 3.1. *The estimator $\hat{S}_n^{\mathrm{cv}}(f)$ is unbiased, and satisfies*

$$\left(e_2\left(\hat{S}_n^{\mathrm{cv}}\right)\right)^2 \;=\; \max_{f\in F} \mathrm{var}\left(\hat{S}_n^{\mathrm{cv}}(f)\right) \;=\; \frac{1}{12n}.$$

The maximal error is attained for any unit step function.

Proof. The estimator is unbiased since $\mathbb{E}(\tilde{f}(X_i)) = 1/2$, and therefore the mean-squared error is equal to $\mathrm{var}(\hat{S}_n^{\mathrm{cv}}(f)) = \frac{1}{n}\mathrm{var}(f(X) - X)$. For a unit step function $f = \mathbb{1}_{[x_0,1]}$ with a jump at $x_0 \in [0,1]$, the random variable $f(X) - X$ is uniformly distributed over $[-x_0, 1-x_0]$, which yields $\mathrm{var}(\hat{S}_n^{\mathrm{cv}}(f)) = 1/(12n)$ as claimed. It remains to show that $\mathrm{var}(f(X) - X) \leq \frac{1}{12}$ for all $f \in F$.

Let $F_m \subset F$ denote the class of all non-decreasing staircase functions of the form $f = \sum_{k=1}^{m} \alpha_k \cdot \mathbb{1}_{(\frac{k-1}{m}, \frac{k}{m}]}$, with $0 \leq \alpha_1 \leq \alpha_2 \leq \ldots \leq \alpha_m \leq 1$. For any $f \in F$, consider the piecewise-constant approximation $f_m \in F_m$ defined by averaging f over each subinterval of length $1/m$. Then, $\mathbb{E}(f_m(X)) = \mathbb{E}(f(X))$ and $|\mathrm{var}\left(f_m(X) - X\right) - \mathrm{var}\left(f(X) - X\right)| \leq \frac{1}{m}$. Thus,

$$\sup_{f\in F} \mathrm{var}\left(f(X) - X\right) \;=\; \lim_{m\to\infty} \sup_{f\in F_m} \mathrm{var}\left(f(X) - X\right). \tag{3}$$

Let us now show that $\mathrm{var}\left(f(X) - X\right)$ is maximized over F_m when f is a unit step function. Pick any $f = \sum_{k=1}^{m} \alpha_k \cdot \mathbb{1}_{(\frac{k-1}{m}, \frac{k}{m}]} \in F_m$. Set $\alpha_0 = 0$ and $\alpha_{m+1} = 1$. If f is not a unit step function, then there exist $k_1, k_2 \in \{1, \ldots, m\}$ such that $k_1 \leq k_2$ and $\alpha_{k_1-1} < \alpha_{k_1} = \ldots = \alpha_{k_2} < \alpha_{k_2+1}$. Denote by $f_u \in F_m$ the function obtained by changing the common value of $\alpha_{k_1}, \ldots, \alpha_{k_2}$ in f to $u \in [\alpha_{k_1-1}, \alpha_{k_2+1}]$. The variance $\mathrm{var}(f_u(X) - X)$ is a convex function of u, since it can be expanded as $au^2 + bu + c$ with $a = \frac{k_2-k_1+1}{m}(1 - \frac{k_2-k_1+1}{m}) > 0$. Consequently, we have $\mathrm{var}(f_u(X) - X) > \mathrm{var}(f(X) - X)$ at one of the two endpoints of $[\alpha_{k_1-1}, \alpha_{k_2+1}]$. Note that the corresponding staircase function f_u has one fewer step than f. Iterating as necessary, we conclude that for any $f \in F_m$ there exists a unit step function $g \in F_m$ such that $\mathrm{var}\left(f(X) - X\right) \leq \mathrm{var}\left(g(X) - X\right) = \frac{1}{12}$. Therefore $\sup_{f\in F_m} \mathrm{var}\left(f(X) - X\right) = \frac{1}{12}$, which completes the proof. □

4 Stratified sampling

Consider now a stratified sampling estimator with K strata:

$$\hat{S}_n^{\text{str}}(f) \;=\; \sum_{k=1}^{K} w_k \cdot \frac{1}{n_k} \sum_{i=1}^{n_k} f\left(X_{k,i}\right), \tag{4}$$

where the k-th stratum is $I_k = [x_{k-1}, x_k]$, $0 = x_0 < x_1 < \cdots < x_{K-1} < x_K = 1$, the weight $w_k = |x_{k-1} - x_k|$ is the length of the k-th stratum, the allocation scheme $(n_1, \ldots, n_K)$ is such that $n_k \geq 0$ for all k and $\sum_k n_k = n$, and the random variables $X_{k,i}$ are independent, with the $X_{k,i}$s uniformly distributed in I_k. Note that the sampling points are no longer identically distributed here. The estimator $\hat{S}_n^{\text{str}}(f)$ is unbiased, with variance

$$\operatorname{var}\left(\hat{S}_n^{\text{str}}(f)\right) \;=\; \sum_{k=1}^{K} \frac{w_k^2}{n_k} \operatorname{var}\left(f\left(X_{k,1}\right)\right). \tag{5}$$

Theorem 4.1. *For any $K \leq n$, any choice of strata and any allocation scheme, the stratified sampling estimator* (4) *satisfies*

$$\left(e_2\left(\hat{S}_n^{\text{str}}\right)\right)^2 \;=\; \max_{f \in F} \operatorname{var}\left(\hat{S}_n^{\text{str}}(f)\right) \;=\; \frac{1}{4} \max_k \frac{w_k^2}{n_k}, \tag{6}$$

The maximal error is attained for a unit step function with a jump at the middle of I_{k^*}, where $k^* \in \operatorname{argmax} w_k^2/n_k$. The minimal value of the maximal error (6) *is $\frac{1}{4n^2}$, and is obtained with $K = n$ strata of equal lengths ($w_k = n^{-1}$ and $n_k = 1$ for all k).*

The optimal stratified sampling method can be seen as a one-dimensional special case of the Latin Hypercube Sampling (LHS) method (McKay et al., 1979). (On a related note, McKay et al. (1979) prove that, in any dimension, the LHS method is preferable to the simple Monte Carlo method if the function is monotone in each of its arguments.)

Proof. For all $K \in \mathbb{N}^*$, let $\mathbf{\Delta}_K = \left\{(\Delta_1, \ldots, \Delta_K) \in \mathbb{R}_+^K \mid \sum_{k=1}^K \Delta_k \leq 1\right\}$. For a given stratified sampling method with K strata, for all $\underline{\Delta} \in \mathbf{\Delta}_K$, define

$$F_{\underline{\Delta}} \;=\; \left\{ f \in F \;\middle|\; \forall k \in \{1, \ldots, K\},\, f(x_k) - f(x_{k-1}) = \Delta_k \right\}.$$

Then it follows from (5) and Popoviciu's inequality that

$$\max_{f \in F_{\underline{\Delta}}} \operatorname{var}\left(\hat{S}_n^{\text{str}}(f)\right) \;=\; \frac{1}{4} \sum_{k=1}^{K} \frac{w_k^2 \Delta_k^2}{n_k}, \tag{7}$$

where the maximum is attained for a non-decreasing staircase function with jumps of height Δ_k at the middle of the strata. Note that $\sum_{k=1}^K \Delta_k^2 \leq \sum_{k=1}^K \Delta_k \leq 1$. Therefore, the right-hand side of (7) is upper-bounded by $\frac{1}{4} \max_k w_k^2/n_k$, which is indeed the value of

the variance (5) when f is a unit step function with a jump at the middle of the stratum where w_k^2/n_k is the largest.

In order to prove the second part of the claim, observe that any stratum with $n_k \geq 2$ can be further divided into n_k sub-strata of equal lengths without increasing the upper bound. Considering then the case where $K = n$ and $n_k = 1$ for all k, the upper bound reduces to $\frac{1}{4} \max_k w_k^2$, which is minimal when $w_1 = \cdots = w_n = n^{-1}$ since $\sum_k w_k = 1$. □

5 Discussion

The stratified sampling (LHS) method provides the best-known variance upper bound over the class F for an unbiased nonsequential method as soon as $n \geq 3$, but is outperformed by the control variate method of Section 3 when $n \leq 2$. We do not know at the moment if these results are optimal in the class of unbiased nonsequential methods. (The ratio between the best variance upper bound and the lower bound of Corollary 2.2 is $\frac{8}{3} \approx 2.67$ for $n = 1$, $\frac{16}{3} \approx 5.33$ for $n = 2$ and 8 for $n \geq 3$.)

Relaxing the unbiasedness requirement, it turns out that both methods are outperformed for all n by the (deterministic) trapezoidal method discussed in the introduction, which has a worst-case squared error of $1/(4(n+1)^2)$. The ratio of worst-case mean-squared errors, however, is never very large—at most $\frac{16}{9} \approx 1.78$—and goes to 1 when n goes to infinity.

Directions for future work include constructing better sequential methods using the findings of this article, and extending our results to mutivariate integration with partial monotonicity (see, e.g., McKay et al., 1979, Section 2.1).

References

F. Perrin, F. Tenenhaus-Aziza, V. Michel, S. Miszczycha, N. Bel, and M. Sanaa. Quantitative risk assessment of haemolytic and uremic syndrome linked to O157:H7 and non-O157:H7 shiga-toxin producing escherichia coli strains in raw milk soft cheeses. *Risk Analysis*, 35(1):109–128, 2014.

J. Kiefer. Optimum sequential search and approximation methods under minimum regularity assumptions. *Journal of the Society for Industrial and Applied Mathematics*, 5 (3):105–136, 1957.

E. Novak. Quadrature formulas for monotone functions. *Proceedings of the American Mathematical Society*, 115(1):59–68, 1992.

M. D. McKay, R. J. Beckman, and Conover W. J. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 21(2):239–245, 1979.

Subsampling under constraint

Florian COMBES[1] Ricardo Fraiman[2] Badih Ghattas[3]

[1] *Aix Marseille University, CNRS, I2M, Marseille, France*
Renault Group, Customer usage, Guyancourt, France
florian.combes@centraliens.net
[2] *Centro de Matemática, Universidad de la República, Uruguay*
fraimanricardo@gmail.com
[2] *Aix Marseille University, CNRS, I2M, Marseille, France*
badih.ghattas@univ-amu.fr

Abstract. In our setting we have an input X in a general space, and an output $Y = f(X)$ where f is a very complicated function, whose computational cost for every new input is very high. We are given two sets of observations of X, S_1 and S_2 of different sizes such that only $f(S_1)$ is available. We tackle the problem of selecting a subsample $S_3 \in S_2$ of smaller size on which to run the complex model f, and such that distribution of $f(S_3)$ is close to that of $f(S_1)$. We suggest two algorithms to solve this problem and show their efficiency using simulated datasets.

Keywords. Optimal sampling, numerical models, nearest neighbours, Kolmogorov–Smirnov.

1 Introduction

Suppose we have an iid sample $S_1 = \{X_1, \ldots, X_{n_1}\}$, with the same distribution as X, of data defined on a space, that we will assume to be just a complete separable metric space $(\mathcal{E}, \rho)$. We apply to each of these observations a very complicated, expensive and deterministic smooth function $f : \mathcal{E} \to \mathbb{R}$, which we will consider as black box. The pair $(S_1, f(S_1))$ may also be seen as the result of a large establishment survey.

Next, another large iid sample S_2 of size n_2 is provided with the same distribution as the first one, but for which the values $f(S_2)$ are not provided.

The main problem we address is how to provide a subsample $S_3 \subset S_2$ of size n_3 smaller than n_2 and such that the distribution of $f(S_3)$ will be close to that of $f(S_1)$. The idea is that in the future the values of $f(X)$ will be only computed for S_3.

This problem appears quite often in practice, in particular in some industrial applications, semi-supervised learning, neuroscience, big data regression and clustering, among many other problems. At first glance this problem is that of a classical subsampling. It can be approached by sampling techniques used in surveys, or more recent techniques

adapted to unsupervised ([1]) or supervised ([2], [3],[4]) situations. These approaches are tailored to sample from within a population(a large sample, accessible or not). They may be used partially to solve our problem which is semi supervised. In our context, we need to sample from S_2 with constraints related to S_1.

2 The problem setting

Let $\mathcal{S}_1 = \{X_1, \ldots, X_{n_1}\}$ be a set of n_1 iid random elements in a complete separable metric space $(\mathcal{E}, \rho)$, with the same common distribution μ as X, and $\mathcal{S}_2 = \{X'_1, \ldots, X'_{n_2}\}$ a second iid sample of size n_2 with the same distribution μ. Let $f : \mathcal{E} \to \mathbb{R}$, a deterministic function which is very complicated and hard to compute (which we may think as a regular black box). The unknown distribution of $f(X)$ will be denoted by F. We have a sample

$$\mathcal{Y}_1 =: \{Y_i = f(X_i) \text{ for } i = 1, \ldots n_1\}$$

of the images of the first sample $\mathcal{S}_1$. Images for S_2 are not available.

With this information at hand, we want to find a subsample $\mathcal{S}_3 \subset \mathcal{S}_2$, with size $n_3 << n_2$, such that the empirical distribution of $f(S_3) := \{f(X_j) : X_j \in S_3\}$ will be close to the distribution of $f(X_1)$.

We will consider several different approaches to this problem throughout the manuscript, with different complexities. Some of them do not make use of the sample $\mathcal{Y}_1$, while others do.

3 Some Algorithms

We first propose a simple solution which does not make use of $\mathcal{Y}_1$. Then we will introduce another algorithm that make use of the output sample $\mathcal{Y}_1$ in different ways. The idea is to get a subsample $\mathcal{Y}_3$ from $\mathcal{Y}_1$ whose distribution is close to that of $\mathcal{Y}_1$, consider its inverse image $f^{-1}(\mathcal{Y}_3)$, which is a subsample of S_1, and look for its neighbours in S_2.

3.1 A simple extended nearest neighbours approach

Consider $S_1 = \{X_1, ...X_{n_1}\}$ and $S_2 = \{X'_1, ...X'_{n_2}\}$ with $n_2 > n_1$. Compute the nearest neighbours of S_1 in S_2, let $d_1, ..., d_{n_1}$ be their ordered distances and $j(1), ...j(n_1)$ their indices.

If two observations from S_1, X_i and X_j, have the same nearest neighbour, say X'_l, at distances d_i and d_j, such that $d_i < d_j$, than X'_l will be kept as a neighbour of X_i and for X_j we take its second nearest neighbour from S_2. If more than two observations have the same nearest neighbour, we will need to explore further away neighbours.

We end with the set $\mathcal{S}_3 = \{X'_1, \ldots, X'_{n_1}\}$ and its Prokhorov distance to μ_1 will be smaller

than $d^{(n_1)}$ which will be small if n_2 is large and $n_2 >> n_1$. Indeed, $d^{(n_1)}$ will converge to 0 over any compact set $K \subset \mathcal{E}$. This simple approach is based on the idea that observations from S_2 close to S_1, should have images through f close to the images of S_1.

3.2 A histogram based approach

In the previous section we did not make use of $\mathcal{Y}_1 = f(S_1)$. Suppose now that we are interested in $\mathbb{P}(f(X) \in I)$, where $I \in \mathbb{R}$ is an interval, or a finite union of disjoint intervals in $\mathbb{R}$, say $I_1, \ldots, I_k$. We look for a subsample $\mathcal{S}_3 \subset \mathcal{S}_2$, with size $n_3 << n_2$, such that we can approach $\mathbb{P}(f(X) \in I)$. We start by considering the set

$$A_{n_1}(I) := \{X_j \in \mathcal{S}_1 : f(X_j) \in I\}.$$

Next, given $\epsilon > 0$, define

$$B_{n_1}(I) := B(A_{n_1}(I), \epsilon) := \bigcup_{X_j \in A_{n_1}(I)} B(X_j, \epsilon),$$

and

$$\mathcal{S}_3(\epsilon) = \{X_i \in \mathcal{S}_2 : X_i \in B(A_{n1}(I), \epsilon)\}.$$

The heuristic idea in this case is to look for a subsample S_3 such that the histogram of $f(S_3)$ is close to the one built up with the intervals I of the distribution of $f(X)$, assuming that the function f is smooth.

In this case, the size of $\mathcal{S}_3(\epsilon)$ is random and depends on ϵ. From an asymptotic point of view, we will need that $\epsilon \to 0$ slowly enough, since we can think of the problem as estimating the distribution of $X_1 | f(X_1) \in I$. Some theory can be derived along this line. An alternative is to fix n_3 and choose ϵ in order to have a subsample of size approximately n_3. A special case of this approach is implemented in Algorithm 2 and consists in using bins obtained by adjusting a histogram to $f(S_1)$.

The partitions in this algorithm may be obtained at random or using a randomized clustering algorithm like k-means. Note that the size of the obtained subsamples in this case are fixed and equal to m.

4 Some experiments

In this section we provide the results of some simulations for one algorithms proposed; histogram based (algorithm 2). As mentioned above, our objective is to find a subsample $S_3 \subset S_2$ for which the distribution of the set $f(S_3)$ will be close to that of $f(S_1)$, without using the values of $f(S_2)$.
To do this, we will apply each algorithm to data generated from various distributions, varying sample sizes (n_1 and n_2) as well as the dimension d of the inputs. The values of

these parameters are fixed as follows: $n_1 = 100$, $n_2 \in \{200, 1000, 5000\}$, $d \in \{1, 2, 20\}$. We use the following distributions; the coordinates are independent for $d > 1$:

- beta distribution $\mathcal{B}(0.5, 0.5)$.
- Gaussian distribution, $\mathcal{N}(0, 1)$.
- uniform distribution, $\mathcal{U}[-1, 1]$.
- truncated Gaussian distribution, where the interval of truncation is $[-1, 1]$. We denote it by $\mathcal{TN}(1)$.

Once we get the output subset S_3 from any algorithm we use the Kolmogorov–Smirnov test to compare the observed empirical distributions of $f(S_1)$, with that of $f(S_3)$ which is not available in general. We report the values of the test statistic as well as the corresponding p-values, averaged over $K = 100$ runs for each configuration.

Tables 1 give the results for Algorithm 2 and we can notice that we obtain good results for almost all distributions and values of n_2. The only case where the obtained subsamples are not satisfactory was for the Gaussian distribution in dimension 20. Using the truncated Gaussian distributions improved the results. We think that the actual version of our algorithms might fail for tailed distributions in high dimensions.

Distribution	n_2	$d = 1$			$d = 2$			$d = 20$		
		n_3	Stat	Pvalue	n_3	Stat	Pvalue	n_3	Stat	Pvalue
$\mathcal{B}(0.5, 0.5)$	200	48	0.06	0.99	48	0.08	0.96	47	0.16	0.43
	1000	47	0.06	0.99	47	0.06	0.99	47	0.16	0.42
	5000	47	0.06	0.99	48	0.06	0.99	47	0.16	0.43
$\mathcal{N}(0, 1)$	200	48	0.08	0.92	48	0.11	0.80	47	0.30	0.03
	1000	48	0.08	0.93	47	0.09	0.93	48	0.32	0.01
	5000	48	0.08	0.92	47	0.08	0.96	47	0.33	0.01
$\mathcal{U}[-1, 1]$	200	48	0.06	0.99	48	0.07	0.97	48	0.14	0.53
	1000	47	0.06	0.99	48	0.06	0.99	47	0.15	0.47
	5000	47	0.06	0.99	48	0.06	0.99	47	0.15	0.45
$\mathcal{TN}(1)$	200	47	0.07	0.99	47	0.09	0.89	48	0.26	0.11
	1000	48	0.06	0.99	47	0.07	0.97	48	0.30	0.03
	5000	48	0.07	0.98	48	0.07	0.99	47	0.29	0.04

Table 1: Average size of n_3, Kolmogorov-Smirnov test between S_1 and S_3 for beta, Gaussian, distribution and truncated gaussian distribution (Algorithm 2).

4.1 Comparing to other sampling approaches

Our algorithms aim to construct a subsample S_3 from S_2 using information or constraints related to S_1. Existing subsampling approaches cannot achieve this task but may be used as an alternative to some steps in our algorithms. We will use such approaches to obtain a subsample S_3 directly form S_2 without considering neither S_1 nor $\mathcal{Y}_1$ in algorithm 1, and as an alternative in algorithms 2 to select a subsample from S_1.
To do that, we consider two recent unsupervised approaches; the "support points" [1] and the "D-optimality" sampling [2]. We will use Support Points (SP) and D-optimality as alternatives in the histogram based approach (algorithm 2). Recall that in algorithm 2 we use a histogram to partition $\mathcal{Y}_1$, and sample 50% of the observations within each bin; we consider then their inverse image by f which is a subsample from S_1. We replace this process by either sampling from $\mathcal{Y}_1$ using support points and D-optimality, either sampling directly from S_1.

Distribution	Selection	n_2	$d=1$			$d=2$			$d=20$		
			n_3	Stat	Pvalue	n_3	Stat	Pvalue	n_3	Stat	Pvalue
$\mathcal{B}(0.5,0.5)$	SP	200	50	0.04	1	50	0.06	0.99	50	0.13	0.61
		1000	50	0.03	1	50	0.05	1	50	0.13	0.64
		5000	50	0.02	1	50	0.04	1	50	0.13	0.63
	D-opt	200	50	0.26	0.02	48	0.13	0.63	40	0.14	0.58
		1000	50	0.26	0.02	48	0.12	0.65	40	0.15	0.57
		5000	50	0.26	0.02	48	0.13	0.63	40	0.14	0.59
$\mathcal{N}(0,1)$	SP	200	50	0.05	1	50	0.08	0.95	50	0.32	0.02
		1000	50	0.03	1	50	0.06	0.99	50	0.36	0.01
		5000	50	0.03	1	50	0.06	1	50	0.37	0
	D-opt	200	50	0.31	0.01	48	0.29	0.01	40	0.29	0.06
		1000	50	0.32	0	48	0.31	0.01	40	0.30	0.04
		5000	50	0.31	0	48	0.32	0.01	40	0.29	0.06
$\mathcal{U}[-1,1]$	SP	200	50	0.04	1	50	0.07	0.99	50	0.13	0.61
		1000	50	0.03	1	50	0.05	1	50	0.13	0.60
		5000	50	0.02	1	50	0.05	1	50	0.13	0.59
	D-opt	200	50	0.26	0.02	48	0.15	0.46	40	0.15	0.53
		1000	50	0.26	0.02	48	0.15	0.44	40	0.14	0.59
		5000	50	0.26	0.02	48	0.15	0.44	40	0.14	0.59
$\mathcal{TN}(1)$	SP	200	50	0.04	1	50	0.07	0.98	50	0.18	0.34
		1000	50	0.03	1	50	0.05	1	50	0.21	0.18
		5000	49.98	0.02	1	50	0.04	1	50	0.21	0.21
	D-opt	200	50	0.26	0.02	50	0.24	0.04	50	0.18	0.35
		1000	50	0.26	0.02	50	0.26	0.02	50	0.20	0.21
		5000	50	0.26	0.02	50	0.26	0.02	50	0.21	0.19

Table 2: Average size of n_3, Kolmogorov-Smirnov test between S_1 and S_3 for beta, Gaussian, uniform and truncated gaussian distribution (Algorithm 2 on X).

Tables 2 give the results obtenu when using SP over X in our histogram based algorithm. Using SP gives better results when compared to the other approaches for all the simulations models mainly in dimension lower than 5. D-optimality is less efficient except

for large dimensions (20) where its performance is close to that of SP. Moreover, the SP and the D-optimality behave like the other algorithms with the normal distribution, i.e. they are less good on the normal distribution than on the truncated normal distributions. For the uniform distribution, SP gives very good results, while the D-optimality is poor except for dimension 20 where both give the same results.

5 Conclusion

We have considered the problem of selecting a new subsample S_3 to use for inference with a model f whose output is available only for a small subsample S_1. The subsample should be small, and the distribution of $f(S_3)$ should be close to that of $f(S_1)$. Two algorithms were considered. The first one, an extended nearest neighbours approach, makes no use of the sample $f(S_1)$, while the other (histogram based approache) makes use of it in different ways. All the algorithms showed a good behaviour when analysed through simulation using different distributions and different dimensions for the input, except for the Gaussian case in high dimensions. Our algorithms are currently extended to the case where the input data are time series and the output is multidimensional.

References

[1] R. Joseph V., S. Mak, Support points, The Annals of Statistics 46 (6A) (2018) 2562–2592. doi:10.1214/17-AOS1629.
URL https://doi.org/10.1214/17-AOS1629

[2] W. HaiYing, Y. Min, S. John, Information-based optimal subdata selection for big data linear regression, Journal of the American Statistical Association 114 (525) (2019) 393–405. arXiv:https://doi.org/10.1080/01621459.2017.1408468, doi:10.1080/01621459.2017.1408468.
URL https://doi.org/10.1080/01621459.2017.1408468

[3] W. HaiYing, Z. Rong, M. Ping, Optimal subsampling for large sample logistic regression, Journal of the American Statistical Association 113 (522) (2018) 829–844, pMID: 30078922. arXiv:https://doi.org/10.1080/01621459.2017.1292914, doi:10.1080/01621459.2017.1292914.
URL https://doi.org/10.1080/01621459.2017.1292914

[4] V. R. Joseph, S. Mak, Supervised compression of big data, Statistical Analysis and Data Mining: The ASA Data Science Journal 14 (3) (2021) 217–229. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/sam.11508, doi:10.1002/sam.11508.
URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sam.11508

Recursive conditional distribution estimation using kernel method and Bernstein polynomials

Helali Salima [1] and Yousri Slaoui[2]

[1]Laboratoire Angevin de Recherche en Mathématiques - Angers – angers – France
[2]Laboratoire de Mathématiques et Applications – Université de Poitiers, Centre National de la Recherche Scientifique : UMR7348 – Téléport 2 - BP 30179 Boulevard Marie et Pierre Curie 86962 Futuroscope Chasseneuil Cedex, France

Résumé

The estimation of the conditional distribution function is a fundamental issue in non parametric estimation. From this perspective, a recursive estimator using kernel method and Bernstein polynomials, is introduced. We derive the asymptotic properties of the proposed estimator such as its asymptotic bias, variance and mean squared error, which strongly depend on the choice of some parameters. The asymptotic normality of the estimator is also established. Under some conditions, the proposed recursive estimator will be very competitive to other estimators, in terms of estimation error. Eventually, the performance of the proposed estimator is explored through a few simulation examples.

Explainability as statistical inference

Hugo Senetaire[1], Jes Frellsen[2] & Pierre-Alexandre Mattei[3]

[1] *Technical University of Denmark, Denmark, hhjs@dtu.dk*
[2] *Technical University of Denmark, Denmark, jefr@dtu.dk*
[3] *Université Côte d'Azur, Inria, LJAD, CNRS, France, pierre-alexandre.mattei@inria.fr*

Résumé. Une grande variété d'approches pour l'explication de modèle en apprentissage profond a été proposée au cours de ces dernières années. Dans ce papier, nous nous intéresserons aux approches basées sur la suppression de variables. Habituellement, ces approches quantifient l'importance d'un ensemble de caractéristiques via les variations d'une fonction d'évaluation dues à l'absence ou la présence des dites caractéristiques. Plus précisément, nous nous intéressons ici aux modèles d'interprétabilité amortis, où un réseau de neurones est utilisé comme sélecteur d'un ensemble minimal de variables sans trop endommager la fonction d'évaluation. L'utilisation d'un réseau secondaire pour la selection doit permettre une interprétation rapide du modele en inférence. Les méthodes traditionnelles reposent sur l'apprentissage séparé du sélecteur et du predicteur. Nous proposons ici une méthode qui repose sur la maximisation de la vraisemblance d'un modèle probabiliste. Cette éthode permet un entrainement rapide et efficace du sélecteur et du prédicteur dans une même descente de gradient. Cette méthode peut s'appliquer a tout type de problème et architecture.

Mots-clés. Interprétabilité, explication de modéle, inférence statistique,

Abstract. A wide variety of model explanation approaches have been proposed in the recent years. Here, we are interested in the removal-based approach, which quantifies how much a prediction change when only a subset of features is shown to the model. More precisely, we look into amortized interpretability models, where a neural network is used as a selector to allow for fast interpretation at inference time. Traditional methods rely on separate training of the predictor and the selector models. Here, we propose a single probabilistic model with a single maximum likelihood based loss. This allows for fast and efficient training of the selector and predictor at the same time. Our proposed method is independent of the predictor network architecture and can be apply to any problem.

Keywords. Interpretability, model explanation, statistical inference, instance wise feature selection.

1 Introduction

Fueled by the many advances in deep learning, machine learning models are omnipresent in society. Their wide-spread use for decision making or predictions in critical fields

lead to a growing need for transparency and interpretability of these methods. Machine learning models are getting more complex with a growing number of parameters used. While Rudin (2019) argues that we should always favor interpretable models for high stake decisions, in practice, a lot of black-box methods remains.

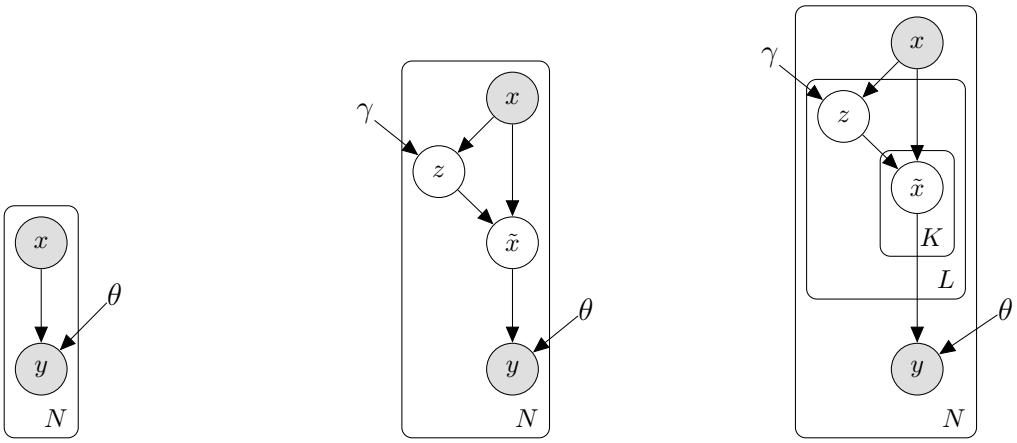
Researchers have proposed a variety of model explanations approaches for black-box models. Wachter et al. (2017) uses counterfactuals to find limitations in the prediction, Ribeiro et al. (2018) describes sufficient conditions to stay within range of a prediction, Ribeiro et al. (2016) fit an interpretable linear regression locally around a given instance, Seo et al. (2018) creates saliency map by back-propagating gradient through the model. Finding interpretable method is hard and many of the previous methods lack fidelity. The lack of proper evaluation methods (Afchar et al., 2021; Jethani et al., 2021a; Liu et al., 2021) makes it difficult to assess the qualities of the different methods.

In this paper, we will focus on methods that offer an understanding of which features is important for the generation of the target for a given instance. This type of methods are called instance-wise feature selection and are part of the removal based selection framework proposed by Covert et al. (2021). The idea of such methods is to quantify how much a prediction change, when only a subset of features is shown to the model.

The first methods in that fashion were quite slow, as they required evaluating the performance anew for each instances. For instance, Ribeiro et al. (2016) proposed to fit a linear regression model locally to every instances, and Lundberg and Lee (2017) calculate Shapley values for every instance.

Over the recent years, there has been an increased research interested in finding amortized explanation methods. The idea with such method is to train a *selector* network that will choose the important feature for the prediction. While there is a higher cost of entry due to training an extra network, the interpretation at test time is very fast. Jethani et al. (2021b) proposed an extension of SHAP in order to get fast explanation for a dataset. Chen et al. (2018), Yoon et al. (2018) both proposed to train a selector that select a minimum subset of features while maximizing a information theoretical threshold. Jethani et al. (2021a) showed that the selector was not really constrained to select the important features, but would encode target information to facilitate the prediction and proposed a method that train a surrogate predictor to alleviate this issue. In all these methods, we could either have unfaithful interpretations or need to separate the training of predictor and selector.

Here we propose a specific method to train both selector and predictor that derives from a single model likelihood hence which has a statistical learning interpretation with guarantees of control selection. This method can be used either post-hoc, to explain a given predictor without the need of a surrogate or ad-hoc, the selector and predictor are trained jointly. The complete model can be considered an explainable model. We also provide some theoretical guarantees that the selection is faithful in our model.



(a) (b) (c)

Figure 1: The different graphical models for a simple predictor (1a), the complete model with selection (1b) and the IWAE model (1c).

2 Method: An explainable statistical model

Let $\mathcal{X} = \prod_{d=1}^{D} \mathcal{X}_i$ a D-dimensional feature space and $\mathcal{Y}$ the target space. We consider $\mathbf{X} = (X_1, \ldots, X_D)$ and $Y \in \mathcal{Y}$, two random variables following the true distribution $p(\mathbf{X}, Y)$. We suppose we have access to N realisations of these two random variables, $\mathbf{x}_1, \ldots, \mathbf{x}_N \in \mathcal{X}$ and labels $y_1, \ldots, y_N \in \mathcal{Y}$. We want to find the distribution $p(Y|X)$.

To find this distribution, we consider a distribution $p_\theta(y|x) = \Phi(y|f_\theta(x))$ where $f_\theta : \mathbb{R}^d \to H$ is a neural network and $(\Phi(.|\eta))_{\eta \in H}$ is a parametric family, usually categorical distribution for the output density, here parameterized by the output of f_θ. We then maximise the likelihood of the model or equivalently the log-likelihood given by

$$\mathcal{L}(\theta) = \sum_{n=1}^{N} \log p_\theta(y_n|x_n). \tag{1}$$

Remembering that interpretability is our goal, we want to obtain a score that tells if a feature is useful for the prediction, and we want this interpretation to be easily readable by a human, to that end we propose to have a score per feature. We propose to create a latent $Z \in \{0,1\}^D$ that correspond to a subset of selected feature. The idea is that if $z_d = 1$, then feature d is used by the predictor, hence it is not turned off and conversely. We endow this latent variable with a distribution $p_\gamma(z|x)$. For instance, to simplify and get easily readable score, we model the latent variable Z as a product of independent Bernoulli variables $Z_1, \ldots, Z_D \in \{0,1\}^D$ such that $p_\gamma(z|x) = \prod_{d=1}^{D} \text{Bern}(z_d|g_\gamma(x)_d)$. g_γ is a neural

network denoted the *selector*, and $g_\gamma(x)_d$ designate the importance allocated to feature d for a given instance x. For a given selection $z \in \{0,1\}^D$, we consider the random variable $X_z \in \mathbb{R}^{|z|}$ where only the selected part of X are observed. X_z is the marginalisation over the unobserved feature of X. This allows us to consider $\forall y \in \mathcal{Y}, \forall x_z \in \mathcal{X}^z$,

$$p_\theta(y|x_z) = \mathbb{E}_{X_{1-z}\sim p(.|x_z)}[p_\theta(Y|x_z, X_{1-z})] = \int_{x_{1-z}\in\mathcal{X}^{1-z}} p_\theta(y|x)p(x_{1-z}|x_z)d_{x_{1-z}} \tag{2}$$

We now need to model how the prediction $p_\theta(y_i|x_n, z)$ will handle a feature that is turned off. To guide the modelisation, we want to define $p_\theta(y|x, z)$ such that $p_\theta(y|x, z) = p_\theta(y|x_z)$. In particular, we would want $p_\theta(y|x, \mathbf{1}) = p_\theta(y|x)$. According to Covert et al. (2021), there is only one distribution fulfilling this criterion, given by

$$p_\theta(y|x, z) = \mathbb{E}_{\tilde{X}\sim p(.|x,z)} p_\theta(y|\tilde{X}). \tag{3}$$

Here, $\tilde{X}$ is defined by the generative process: first, we sample $\hat{X} \sim p(X_{Z-1}|X_Z)$, then, we define

$$\tilde{X}_j = \begin{cases} X_j & \text{if } Z_j = 1 \\ \hat{X}_j & \text{if } Z_j = 0. \end{cases} \tag{4}$$

The complete model is shown in Fig. 1b and has the log-likelihood function

$$\mathcal{L}(\theta, \gamma) = \sum_{n=1}^{N} \log[\mathbb{E}_{Z\sim p_\gamma(.|x_n)} \mathbb{E}_{\tilde{X}\sim p(.|x_Z)} p_\theta(y_n|\tilde{X})]. \tag{5}$$

Constraints on the selection If $p_\gamma(z|x)$ can be any conditional distribution, then the model can just refuse to learn something explainable by setting $p_\gamma(z|x) = \mathbf{1}_{z=\mathbf{1}_d}$ (which means all features are always turned on). To prevent the model from doing that, we can add an additional constraint on p_γ. Following the idea of Yoon et al. (2018), we add an L1-regularisation on the selection to the complete log-likelihood with a control parameter λ. We then get the optimisation problem

$$\max_{\theta,\gamma} \mathcal{L}(\theta, \gamma) + \lambda \sum_{n=1}^{N} \sum_{d=1}^{D} ||g_\gamma(x_n)_d|| \tag{6}$$

or equvivalently

$$\max_{\theta,\gamma} \sum_{n=1}^{N} \left[\log \mathbb{E}_{Z\sim p_\gamma(.|x_n)} \mathbb{E}_{\tilde{X}\sim p(.|x,Z)} p_\theta(y|\tilde{X}) + \lambda \sum_{d=1}^{D} ||g_\gamma(x_n)_d|| \right]. \tag{7}$$

An importance weighted lower bound We want to maximise Eq. (5) but the difficulty resides in the sum inside the log. We propose to use importance weighted variational inference Burda et al. (2015) so we can have a tighter lower bound than with a simple Jensen inequality.

Optimisation of θ, γ We want to train the maximum likelihood model with stochastic gradient descent. Using Monte Carlo gradient estimator for θ is straightforward. Finding Monte Carlo Gradient estimator for γ is more complicated because the expectation on masks depends on this parameter. We proposed to use a continuous relaxation of the Bernoulli distribution following Jang et al. (2017) that allows us to calculate Monte Carlo Gradient estimator for γ the same way as θ.

Sampling $\tilde{X}$ In most cases, sampling $\tilde{X}$ is not an easy task because we do not have access to the conditional distribution of X. This issue led Jethani et al. (2021a) to train a surrogate classifier model that would approximate $p_\theta(\tilde{X}|Z, X)$. Here we propose instead to approximate the conditional distribution of X using a generative models. While this looks like an expensive task just to then train a model to classify, it appears that even a simple model can achieve some good accuracy in both selection and prediction. More precisely, we advocate the use of a multivariate Gaussian distribution with diagonal covariance matrix. Compared to Jethani et al. (2021a), this allows us to use directly the classifier p_θ in post-hoc and to have a single loss when training both predictor and selector from scratch.

3 Experiment

We propose to evaluate our method on a dataset created following the next steps. We select randomly two images from MNIST (Deng (2012)) and FashionMNIST (Xiao et al., 2017). The two images are put next to each other in a random order and the target is given by the MNIST instance. The optimal selection should select a maximum number of pixels in the MNIST half and a minimum in the FashionMNIST region. We use a 3 layer fully connected neural network with 50 hidden units and ELU activation as a classifier and a 4 layers fully connected neural network with 50 hidden units as a selector. We sample from $\tilde{X}$ using a multivariate Gaussian distribution with diagonal covariance matrix with 20 components.

In Fig. 2, one can observe that our model properly select the correct half space in the images. Even a simple multivariate Gaussian with diagonal covariance is enough to approximate correctly the conditional distribution of X. The accuracy in prediction of the complete model (93 %) is similar to the accuracy of the predictor alone (95 %).

4 Conclusion

We proposed in this paper an instance wise selection model that derives from a single model likelihood. As opposed to previous methods, this model allows for faithful feature selection while training both selector and predictor at the same time. We tested our model

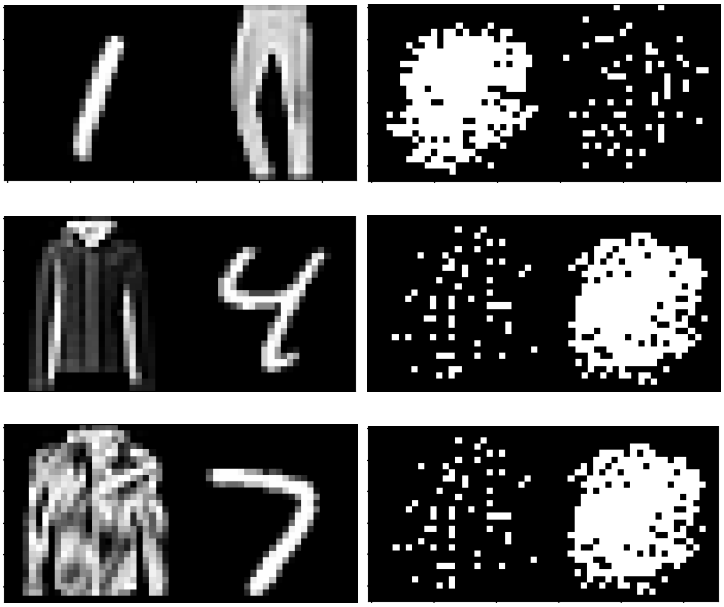


Figure 2: Input image on the left with two halves images and resulting interpretation on the right. We can see that the most selected part of the image corresponds to the MNIST number as expected.

with an artificial dataset consisting of MNIST and FashionMNIST images that gave good results in both prediction accuracy and feature selection accuracy.

Bibliographie

Darius Afchar, Vincent Guigue, and Romain Hennequin. Towards rigorous interpretations: a formalisation of feature attribution. In *International Conference on Machine Learning*, pages 76–86. PMLR, 2021.

Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*, 2015.

Jianbo Chen, Le Song, Martin Wainwright, and Michael Jordan. Learning to explain: An information-theoretic perspective on model interpretation. In *International Conference on Machine Learning*, pages 883–892. PMLR, 2018.

Ian Covert, Scott Lundberg, and Su-In Lee. Explaining by removing: A unified framework for model explanation. *Journal of Machine Learning Research*, 22(209):1–90, 2021.

Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparametrization with gumble-softmax. In *International Conference on Learning Representations (ICLR 2017)*. OpenReview. net, 2017.

Neil Jethani, Mukund Sudarshan, Yindalon Aphinyanaphongs, and Rajesh Ranganath. Have we learned to explain?: How interpretability methods can learn to encode predictions in their interpretations. In *International Conference on Artificial Intelligence and Statistics*, pages 1459–1467. PMLR, 2021a.

Neil Jethani, Mukund Sudarshan, Ian Connick Covert, Su-In Lee, and Rajesh Ranganath. Fastshap: Real-time shapley value estimation. In *International Conference on Learning Representations*, 2021b.

Yang Liu, Sujay Khandagale, Colin White, and Willie Neiswanger. Synthetic benchmarks for scientific research in explainable machine learning. *arXiv preprint arXiv:2106.12543*, 2021.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

Junghoon Seo, Jeongyeol Choe, Jamyoung Koo, Seunghyeon Jeon, Beomsu Kim, and Taegyun Jeon. Noise-adding methods of saliency map as series of higher order partial derivative. *arXiv preprint arXiv:1806.03000*, 2018.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31: 841, 2017.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Jinsung Yoon, James Jordon, and Mihaela van der Schaar. Invase: Instance-wise variable selection using neural networks. In *International Conference on Learning Representations*, 2018.

La Géométrie pour l'Unicité, la Parcimonie et le Regroupement des Estimateurs Pénalisés

Patrick J.C. Tardivel [1] & Ulrike Schneider [2]

[1] *Institut de Mathématiques de Bourgogne, UMR 5584 CNRS, Université Bourgogne Franche-Comté, F-21000 Dijon, France, patrick.tardivel@u-bourgogne.fr*
[2] *TU Wien, ulrike.schneider@tuwien.ac.at*

Résumé. Durant la présentation nous donnerons une condition nécessaire et suffisante pour l'unicité d'un estimateur des moindres carrés pénalisés dont le terme de pénalité est une norme polyèdrique. Nos résultats couvrent de nombreuses méthodes incluant les estimateurs OSCAR, SLOPE et LASSO ainsi que la méthode de poursuite de base qui est une extension du LASSO. Cette condition d'unicité est géométrique et implique l'espace vectoriel engendré par les lignes de la matrice de planification ainsi que les faces de la boule unité de la norme duale coupées par cet espace. Des résultats théoriques sur la parcimonie des estimateurs LASSO et poursuite de base sont déduit de cette condition via la caractérisation des vecteurs signes accessibles pour ces deux méthodes.

Mots-clés. Unicité, LASSO, poursuite de base, vecteur signe accessible.

Abstract. During the talk we will give a necessary and sufficient condition for the uniqueness of a penalized least squares estimator whose penalty term is a polyhedral norm. Our results cover many methods including the OSCAR, SLOPE and LASSO estimators as well as the related method of basis pursuit. The geometrical condition for uniqueness involves how the row span of the design matrix intersects the faces of the dual norm unit ball. Theoretical results on sparsity by LASSO and basis pursuit estimators are deduced from this condition via the characterization of accessible sign vectors for these two methods.

Keywords. Uniqueness, LASSO, basis pursuit, accessible sign vector.

1 Condition nécéssaire et suffisante pour l'unicité

De nombreux estimateurs des moindres carrés pénalisés tels que le LASSO (Least Absolute Shrinkage and Selection Operator), clustered LASSO, OSCAR (Octogonal Shrinkage and Clustering Algorithm for Regression) ou SLOPE (Sorted L One Penalized Estimation)... sont solutions du problème (1) donné ci-dessous. Soit $X \in \mathbb{R}^{n\times p}$, $y \in \mathbb{R}^n$ et $\|.\|$ une norme sur $\mathbb{R}^p$ dont la boule unité est un polytope. L'ensemble $S_{X,\|.\|}(y)$ des solutions du problème des moindres carrés pénalisé est défini par

$$S_{X,\|.\|}(y) := \underset{b\in\mathbb{R}^p}{\text{Argmin}}\ \frac{1}{2}\|y - Xb\|_2^2 + \|b\|. \tag{1}$$

On peut remarquer que $S_{X,\|.\|}(y)$ n'est jamais vide mais $S_{X,\|.\|}(y)$ n'est pas toujours un singleton. En statistique y représente la réponse aléatoire d'un modèle de régression linéaire ainsi il est pertinent de donner une condition sur la matrice X garantissant l'unicité du problème (1) pour tout $y \in \mathbb{R}^n$.

Le Théorème 1 fournit une condition sur X pour que l'ensemble $S_{X,\|.\|}(y)$ soit un singleton pour tout élément $y \in \mathbb{R}^n$. Rappelons que pour une norme $\|.\|$ sur $\mathbb{R}^p$, la norme duale $\|.\|^*$ est définie par $\|x\|^* = \sup_{s\in\mathbb{R}^p:\|s\|\leq 1} s^t x$.

Théorème 1 *Soit $X \in \mathbb{R}^{n\times p}$ et $\|.\|$ une norme pour laquelle la boule unité est un polytope. Il existe $y \in \mathbb{R}^n$ pour lequel l'ensemble $S_{X,\|.\|}(y)$ n'est pas un singleton si et seulement si l'espace vectoriel $\mathrm{im}(X^t)$ coupe une face de la boule unité de la norme duale B^* dont la dimension est strictement inférieure à $dim(\ker(X))$.*

Le Théorème 1 prouvé dans l'article Schneider et Tardivel (2020) généralise le Theorème 15 de Ewald et Schneider (2020) traitant de l'unicité du LASSO. Le Théorème 1 s'étend également au cas où le terme de pénalité est une gauge polyèdrique à valeurs finies (Tardivel et al. (2021)). Les Figures 1 et 2 illustrent le Théorème 1.

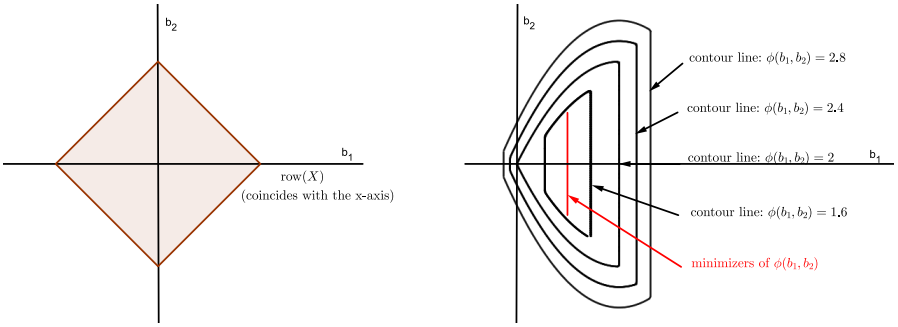


Figure 1: Lorsque $X = (1\ 0)$, les courbes de niveau de la fonction $\phi(b_1, b_2) = 0,5(2 - b_1)^2 + \max\{|b_1|, |b_2|\}$ illustrent que l'ensemble $S_{X,\|.\|_\infty}(2)$ n'est pas un singleton.

Notons que cette condition d'unicité est en pratique très faible. En effet, en posant μ la mesure de Lebesgue sur $\mathbb{R}^{n\times p}$ et $\|.\|$ une norme sur $\mathbb{R}^p$ dont la boule unité est un polytope alors, l'ensemble suivant est négligeable :

$$\mu\left(\left\{X \in \mathbb{R}^{n\times p} : \exists y \in \mathbb{R}^n \text{ tel que } S_{X,\|.\|}(y) \text{ n'est pas un singleton}\right\}\right) = 0.$$

Néanmoins, dans certains cas le problème LASSO n'a pas une unique solution. Par exemple, lorsque $X \in [-1, 1]^{n\times p}$ est une matrice où $p > n$ et ayant une ligne avec au

moins $n+1$ composantes dans $\{-1,1\}$ alors, d'après le Théorème 1, il existe $y \in \mathbb{R}^p$ pour lequel $S_{X,\lambda\|.\|_1}(y)$ n'est pas un singleton. Une telle matrice est fournie dans le jeu de données "gisette". En particuliers, Dupuis et Vaiter (2019) ont montré que pour la matrice X et le vecteur y de ce jeu de données $S_{X,\lambda\|.\|_1}(y)$ n'est pas un singleton dès que $\lambda > 0$ est suffisamment petit.

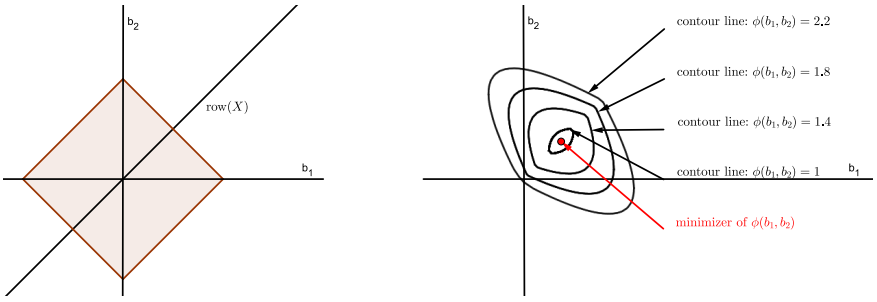


Figure 2: Lorsque $X = (1\ 1)$, les courbes de niveau de la fonction $\phi(b_1,b_2) = 0,5(2-b_1-b_2)^2 + \max\{|b_1|,|b_2|\}$ illustrent que l'ensemble $S_{X,\|.\|_\infty}(2)$ est un singleton.

Le problème poursuite de base introduit ci-dessous peut être vu comme une extension du problème d'optimisation LASSO où le terme de pénalité est $\lambda\|.\|_1$ avec $\lambda > 0$ infiniment petit. L'ensemble $S_{X,\mathrm{pb}}(y)$ des solutions du problème poursuite de base est défini par

$$S_{X,\mathrm{pb}}(y) = \text{Argmin } \|b\|_1 \text{ sous la containte } Xb = y.$$

Une condition nécéssaire et suffisante pour l'unicité du problème poursuite de base est donnée au Théorème 2.

Théorème 2 *Soit $X \in \mathbb{R}^{n\times p}$. Il existe $y \in \mathrm{im}(X)$ pour lequel $S_{X,\mathrm{pb}}(y)$ n'est pas un singleton si et seulement si $\mathrm{im}(X^t)$ coupe une face du cube unité $[-1,1]^p$ dont la dimension est strictement inférieure à $dim(\ker(X))$.*

Notons qu'une matrice en position générale (condition introduite par Dossal (2012) et Tibshirani (2013) garantissant l'unicité des problèmes LASSO et poursuite de base) est une matrice satisfaisant la condition donnée au Théorème 2. Les Figures 3 et 4 illustrent le Théorème 2.

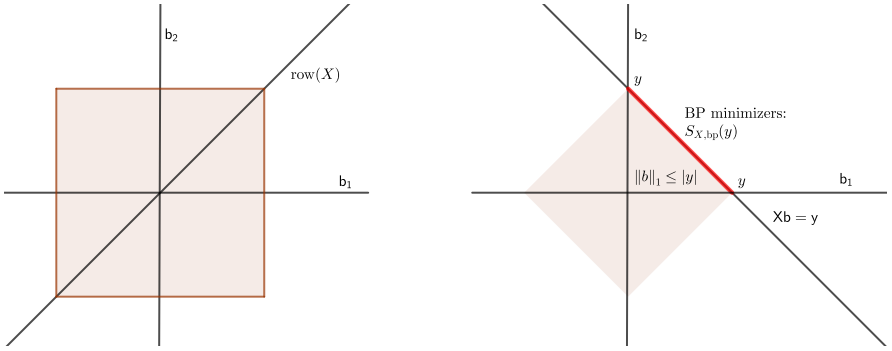


Figure 3: Lorsque $X = (1\ 1)$, la figure montre que l'ensemble $S_{X,\mathrm{pb}}(y)$ n'est pas un singleton pour un certain $y \in \mathbb{R}$.

2 Vecteur signe accessible pour le LASSO et pour la méthode poursuite de base

La notion de vecteur signe accessible a été introduite pour la première fois par Sepehri and Harris (2017) pour l'estimateur LASSO sous l'hypothèse d'unicité. La définition suivante généralise légèrement la notion de vecteur signe accessible donnée dans Sepehri and Harris.

Définition 1 *Soit $X \in \mathbb{R}^{n\times p}$, $\sigma \in \{-1,0,1\}^p$ et $\lambda > 0$. On dit que σ est un vecteur signe accessible pour LASSO (ou poursuite de base), s'il existe $y \in \mathbb{R}^n$ et $\hat{\beta} \in S_{X,\lambda\|\cdot\|_1}(y)$ (ou s'il existe $y \in \mathrm{im}(X)$ et $\hat{\beta} \in S_{X,\mathrm{pb}}(y)$), tels que* $\mathrm{signe}(\hat{\beta}) = \sigma$.

On pose $F(\sigma)$ la face du cube unité définie par

$$F(\sigma) = E_1 \times \cdots \times E_p \text{ avec } E_j = \begin{cases} \{\sigma_j\} & \text{si } |\sigma_j| = 1 \\ [-1,1] & \text{si } \sigma_j = 0 \end{cases}.$$

Théorème 3 *Soit $X \in \mathbb{R}^{n\times p}$, $\lambda > 0$ et $\sigma \in \{-1,0,1\}^p$.*

Caractérisation géométrique : *Le vecteur de signe σ est accessible pour LASSO (ou poursuite de base) si et seulement si* $\mathrm{im}(X^t)$ *coupe la face $F(\sigma)$.*

Caractérisation analytique : *Le vecteur de signe σ est accessible pour LASSO (ou poursuite de base) si et seulement si l'implication suivante est vraie : $Xb = X\sigma \Rightarrow \|b\|_1 \geq \|\sigma\|_1$.*

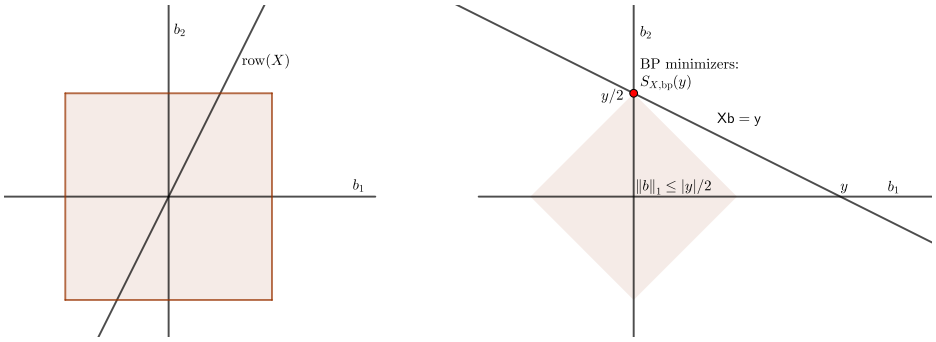


Figure 4: Lorsque $X = (1\ 2)$, la figure montre que l'ensemble $S_{X,\mathrm{pb}}(y)$ est un singleton.

Dans le cadre du modèle de régression linéaire, la condition d'accessibilité est nécessaire et suffisante pour recouvrir les signes des coefficients de régression (positif, nul ou négatif) par le LASSO seuillé, la méthode poursuite de base seuillée (Tardivel et Bogdan (2020)) et par la méthode poursuite de justice seuillée (Descloux et al. (2020)). Un résultat similaire au Théorème 3, donné dans l'article de Schneider et Tardivel (2020), illustre les propriétés de regroupement et de parcimonie des estimateurs SLOPE et OSCAR.

Bibliographie

Descloux, P. and Boyer, C. and Josse, J. and Sportisse, A. and Sardy, S (2020). Robust Lasso-Zero for sparse corruption and model selection with missing covariates, *arXiv*.
Dossal, C. (2012). A necessary and sufficient condition for exact sparse recovery by ℓ_1 minimization, *Comptes Rendus Mathematique*, 350, 1-2, 117–120.
Dupuis, X. Vaiter, S. (2019). The geometry of sparse analysis regularization, *arXiv*.
Ewald, K. and Schneider, U. (2020). On the distribution, model selection properties and uniqueness of the Lasso estimator in low and high dimensions, *Electronic Journal of Statistics*, 14, 1, 944–969.
Schneider, U. and Tardivel, P.J.C. (2020). The geometry of uniqueness, sparsity and clustering in penalized estimation, *arXiv*.
Sepehri, A. and Harris, N. (2017). *The accessible lasso models*, *Statistics*, 51, 4, 711–721.
Tardivel, P.J.C. and Bogdan, M. (2020). On the sign recovery by LASSO, thresholded LASSO and thresholded Basis Pursuit Denoising, *Scandinavian Journal of Statistics*.
Tardivel, P.J.C and Skalski. T and Graczyk, P. and Schneider, U. (2021). The Geometry of Model Recovery by Penalized and Thresholded Estimators, *HAL*.

Tibshirani, R. (2013). The lasso problem and uniqueness, *Electronic Journal of statistics*, 7, 1456–1490.

PLENIERE : Li-Chun Zhang

Descriptive inference of big-data statistics

Li-Chun Zhang

Univ. Southampton

We consider descriptive inference where the targets of interest can in principle be observed in a 'perfect census', in contrast to analytic inference where such a 'perfect census' does not exist even conceptually. For instance, which ones among the residents of France (as a given population) are infected by a certain virus on a given day is a descriptive inference problem, whereas the 'true' regression relationship between a scaler response and a given set of explanatory variables is a problem of analytic inference.

A fundamental challenge for descriptive inference based on supervised (machine) learning is to 'extrapolate' the model learned from the available observations (as a sample) to the unobserved ones, without which the learning would have little use. No matter how learning is organised within the sample, one cannot ensure the adopted model is valid outside it unless the sample is selected from the population in some controlled manner.

We shall consider, in two situations particularly, how probability sampling (or design-based) methods can be combined with supervised learning, such that the validity of descriptive inference is ensured with respect to hypothetical repeated probability sampling, regardless the adopted model is 'true' or not. In the first situation, supervised learning is based on a probability sample such that one can obtain an estimator of the total errors from applying the learned model to the out-of-sample units, which is unbiased over repeated sampling. In the second situation, we assume that the adopted model is learned from a very large 'convenience sample' (or big data), such that it is necessarily misspecified to some extent for the out-of-sample units, due to problems of incomplete coverage, imperfect measurement or informative selection, and the bias of any resulting statistic overwhelms the associated variance. A so-called auditing sample can now be used to provide accuracy measures that are valid over repeated sampling and unaffected by the failure of the assumptions underlying the big-data statistics themselves.

PLENIERE : Emilie Kaufmann

Algorithmes de bandits non-paramétriques: optimalité et robustesse

Emilie Kaufmann

CNRS CRIStAL Lille

Dans un modèle de bandit, un agent sélectionne de manière séquentielle des “bras”, qui sont des lois de probabilité initialement inconnues de l’agent, dans le but de maximiser la somme des échantillons obtenus, qui sont vus comme des récompenses. Les algorithmes de bandits les plus populaires sont basés sur la construction d’intervalles de confiance ou l’échantillonnage d’une loi a posteriori, mais ne peuvent atteindre des performances optimales qu’un ayant une connaissance a priori sur la famille de distributions des bras. Dans cet exposé nous allons présenter des approches alternatives basées sur du ré-échantillonnage de l’historique de chaque bras. De tels algorithmes peuvent s’avérer plus robustes en deux sens. Nous verrons qu’ils peuvent être optimaux pour plusieurs classes de distributions, et être aisément adaptés à des situations où le critère de performance n’est pas lié à la récompense moyenne de l’agent, mais prend en compte une mesure de risque.

16 juin 2022

Programme

PLENIERE : Samory Kpotufe **771**

PLENIERE : Pierre Chainais **772**

SFdS - Statistique mathématique **773**

Profondeur de Tukey: Ensembles de niveau empiriques et théoriques, Victor-Emmanuel Brunel 774

Statistical depth functions: Characterization and reconstruction problems, Stanislav Nagy 780

SFdS - BFA **786**

Multi-normex for evaluating the distribution of aggregated heavy tailed risks, Marie Kratz [et al.] 787

Lundberg-Aumann-Serrano dangerousness risk index, Etienne Marceau 792

ANNULÉ : Three-Step Risk Inference In Insurance Ratemaking, Liang Peng . . . 794

Optimal pooling and distributed inference for the tail index and extreme quantiles, Gilles Stupfler [et al.] 798

Clustering parcimonieux pour extrêmes multivariés, Olivier Wintenberger [et al.] 804

Applications **810**

Votes par évaluation avec des fonctions de profondeur, Jean-Baptiste Aubin [et al.] 811

Impact of the energy transition on long-term factor productivity, Guillaume Flament 818

Pl@ntNet-300K: une base de données d'images de plantes avec une ambiguïté entre classes importante et une distribution à longue traîne, Camille Garcin [et al.] 824

Variance estimators for weighted and stratified linear dose-response function estimators using generalized propensity score, Valerie Gares [et al.] 830

Topological period counting method for reparametrized functions, Wojciech Reise 836

Extrêmes & Prediction **842**

Extremal cluster inference for heavy-tailed time series, Gloria Buritica [et al.] . . 843

Étude de la dépendance spatiale dans les extrêmes : un point de vue géostatistique, Marine Demangeot [et al.] 849

Maximum Mean Discrepancy invariant par translation et convergence en loi, Thibault Modeste [et al.] 855

Vitesse de Convergence Minimax pour la Régression Distributionnelle basée sur le Continuous Ranked Probability Score, Romain Pic [et al.] 860

Non-Asymptotic Analysis of Stochastic Approximation Algorithms for Streaming Data, Nicklas Werge 866

Survie **872**

Machine learning for survival data prediction: Application of the super learner on pseudo-observations, Ariane Cwiling [et al.] 873

Comparaisons de méthodes pour données de survie en grande dimension sur de petits échantillons : optimisation des hyperparamètres et validation., Audrey Lavenu [et al.] 879

Modèle de Cox avec des données hétérogènes, Eliz Peyraud 885

Cox regression with linked data, Thanh Huan Vo [et al.] 891

SFdS - AMIES **897**

Lights: a generalized joint model for high-dimensional multivariate longitudinal data and censored durations, Van Tuan Nguyen 898

L'agence pour les mathématiques en interaction avec l'entreprise et la société (AMIES), Véronique Maume-Deschamps 899

Prédiction des niveaux de risque pollinique à partir de données historiques multi-sources, Esso-Ridah Bleza [et al.] 900

Deep Calibration of Interest Rates Models, Djibril Sarr 906

ML - Graphes - Réseaux **912**

The deep latent position model for node partitioning in graphs, Dingge Liang [et al.] 913

Etude de la Pénalisation GraphNet en Analyse de Données Multi-blocs, Annabelle Beaudoin [et al.] 921

Don't fear the unlabelled: safe deep semi-supervised learning via simple debiasing, Hugo Schmutz [et al.] 927

Generalised Mutual Information Maximisation for Deep Clustering, Louis Ohl [et al.] 933

Méta-modèles – Analyse de sensibilité **939**

Heterogeneous Treatment Effects Estimation: When Machine Learning meets multiple treatments regime, Naoufal Acharki [et al.] 940

Risk bounds for Sobol'-based Global Sensitivity Analysis Indices when using meta-models., Clément Bénesse 941

Test d'indépendance basé sur les indices HSIC-ANOVA d'ordre total, Gabriel Sarazin [et al.] 944

Sélection de variables lors de la généralisation d'un effet causal, Bénédicte Colnet [et al.] 950

Calage conditionnel bayésien d'un modèle numérique, Oumar Baldé [et al.] 956

ML & Regret **962**

Multitask Online Mirror Descent, Pierre Laforgue [et al.] 963

Identification de Meilleur Bras à Budget Fixé, Antoine Barrier [et al.] 969

On Best Policy Identification with Instance-specific sample complexity, Aymen Al Marjani 975

Bayesian Actor for Deep Reinforcement Learning, Léo Grill 980

ANNULÉ : Apprentissage robuste pour la résolution d'EDPs, Applications aux marchés de l'énergie, Carl Remlinger 986

Statistique mathématique **992**

Estimation confidentielle de quantiles en présence d'atomes, Clément Lalanne [et al.] 993

Généralisation et estimation de la Q-distribution gaussienne, Oumaima Ben Mrad [et al.] . . . 999
Projection de mesures de probabilité sous contraintes de quantile par distance de Wasserstein et approximation monotone polynomiale, Marouane Il Idrissi [et al.] 1005
Sparse regression to anticipate the declarations of natural disaster in France, Thi Thanh Yen Nguyen [et al.] . . . 1011
Estimation adaptative optimale de fonctions moyennes et covariances irrégulières, Steven Golovkine [et al.] . . . 1017
Réduction d'une fonction aléatoire cyclostationnaire définie sur R, Sylvie Viguier-Pla [et al.] . . . 1023
PLENIERE : Frédéric Chazal . . . **1029**
PLENIERE : Gersende Fort . . . **1030**
Séries Temporelles . . . **1031**
Une nouvelle méthode pour les propriétés asymptotiques de modèles à seuil autoexcités, Guy Mélard [et al.] . . . 1032
Barycentres de séries temporelles : une nouvelle approche basée sur la méthode de la signature, Raphael Mignot [et al.] . . . 1038
Réduction de dimension pour les séries temporelles avec des auto-encodeurs variationnels, William Todo [et al.] . . . 1044
ML et Extrêmes . . . **1050**
In-store traffic flow forecasting using high dimensional time series methods, Siham Alaoui Belghiti [et al.] . . . 1051
Estimation de paramètres pour un modèle de propagation de la fièvre typhoïde à Mayotte, Ibrahim Bouzalmat [et al.] . . . 1057
Crowdsourcing label noise simulation on image classification tasks, Tanguy Lefort [et al.] . . . 1063
Evaluating the Impact of Parenthood on Career Progression in STEM jobs, across Gender, Mira Rahal [et al.] . . . 1069
Profondeur à noyau multivariée et mesure de risque extrême associée, Sara Armaut 1075
Kalman Recursions Aggregated Online, Eric Adjakossa . . . 1081
Méthodes Génératives - ML - Deep . . . **1088**
Etude du comportement de l'algorithme DbAS pour l'optimisation de propriétés de séquences nucléiques, Teddy Ardouin [et al.] . . . 1089
Limite d'échelle pour les réseaux de neurones résiduels, Pierre Marion . . . 1095
Can Push-forward Generative Models Fit Multimodal Distributions?, Antoine Salmona [et al.] . . . 1101
Generating new crystal structures with statistical methods, Arsen Sultanov [et al.] 1107
Comment imposer des pré-activations gaussiennes dans un réseau de neurones ?, Pierre Wolinski [et al.] . . . 1111
Processus - Statistique mathématique . . . **1117**
Vitesses de contraction du posterior pour les processus gaussiens profonds contraints en classification et estimation de densité, Francois Bachoc . . . 1118
Sélection de variables en régression SIR par seuillage doux ou seuillage dur de la matrice d'intérêt, Hadrien Lorenzo [et al.] . . . 1124
Liens entre les décompositions de Hoeffding–Sobol et de Möbius, Cécile Mercadier 1129
Mélange de modèles d'analyse factorielle longitudinale pour l'analyse de données longitudinales multivariées, Amine Ounajim [et al.] . . . 1133

Estimation adaptative dans un modèle de régression-convolution en base d'Hermite, Ousmane Sacko 1139
A local version of R-hat for MCMC convergence diagnostic, Théo Moins [et al.] . 1145
Statistique mathématique – Valeurs Manquantes **1151**
Vitesse de convergence pour une prediction linéaire en présence de données manquantes, Alexis Ayme [et al.] 1152
Imputation multiple dans le modèle linéaire fonctionnel avec une covariable partiellement observée et des valeurs manquantes dans la réponse, Christophe Crambes [et al.] 1158
Usefulness of surrogate outcomes on the regression model and the PaO2 prediction, Firas Ibrahim [et al.] 1164
Apprentissage semi-supervisé avec réponses manquantes informatives, Aude Sportisse [et al.] 1169
Lights: a generalized joint model for high-dimensional multivariate longitudinal data and censored durations, Herbert Susmann 1174
Rencontre Jeunes statisticiens **1181**
Café de la Statistique **1182**

PLENIERE : Samory Kpotufe

Adaptivity in Domain Adaptation and Friends

Samory Kpotufe

Columbia University

Domain adaptation, transfer, multitask, meta, few-shots, representation, or lifelong learning ... these are all important recent directions in ML that all touch at the core of what we might mean by 'AI'. As these directions all concern learning in heterogeneous and ever-changing environments, they all share a central question: what information a data distribution may have about another, critically, in the context of a given estimation problem, e.g., classification, regression, bandits, etc.

Our understanding of these problems is still rather fledgeling. We plan to present both some recent positive results and also some negative ones. On one hand, recent measures of discrepancy between distributions, fine-tuned to given estimation problems (classification, bandits, etc) offer a more optimistic picture than existing probability metrics (e.g. Wasserstein, TV) or divergences (KL, Renyi, etc) in terms of provable rates. On the other hand, when considering seemingly simple extensions to choices between multiple datasets (as in multitask), or multiple prediction models (as in Structural Risk Minimization), it turns out that minimax oracle rates are not always adaptively achievable, i.e., using just the available data without side information.

The talk will be based on joint work with collaborators over the last few years, namely, G. Martinet, S. Hanneke, J. Suk.

PLENIERE : Pierre Chainais

Echantillonner efficacement grâce à l'approximation de distributions

Pierre Chainais

ECLille CRIStAL Lille

Les méthodes bayésiennes pour les problèmes inverses en traitement du signal et des images ont l'avantage de donner accès à la distribution a posteriori des paramètres à estimer. Ainsi, on accède non seulement à une solution au problème, mais aussi à des intervalles de crédibilité précieux. Par exemple, en astrophysique ou en médecine, il n'existe en général pas de vérité terrain. Fournir des prédictions assorties d'intervalles de confiance est essentiel : la lecture de l'image reconstruite se fait avec un niveau de confiance contrôlé. Néanmoins, les méthodes de Monte Carlo utilisées pour ces simulations de lois a posteriori sont réputées gourmandes en temps de calcul et limitées quant au passage à l'échelle en grande dimension ou pour une grand nombre de paramètres à estimer. Nous présenterons une famille d'approches appelées « augmentation de données asymptotiquement exacte » (AXDA). Cette approche, inspirée du splitting en optimisation, permet de construire de façon systématique une distribution approchée moins coûteuse à échantillonner que la loi cible du modèle initial, dans le cadre d'un compromis efficacité numérique/qualité de l'approximation. Ces méthodes ouvrent la voie à de nombreuses déclinaisons que nous évoquerons et illustrerons par des applications à la résolution de problèmes inverses.

SFdS - Statistique mathématique

Profondeur de Tukey: Ensembles de niveau empiriques et théoriques

Victor-Emmanuel Brunel [1]

[1] *CREST-ENSAE*
5, Av. Le Chatelier, 91120 Palaiseau
victor.emmanuel.brunel@ensae.fr

Résumé. La profondeur de Tukey est une notion statistique qui suscite beaucoup d'intérêt en statistique multivariée, car elle permet d'ordonner les données en généralisant, d'une certaine manière, la notion de quantile. Etant donnés une loi de probabilité et des données i.i.d. suivant cette loi, nous étudions la convergence des ensembles de niveau de la profondeur empirique vers les ensembles de niveau théoriques. Ces ensembles de niveau peuvent être interprétés comme des quantiles multivariés. Sous des hypothèses raisonnables, nous montrons la concentration des ensembles de niveau empiriques à la vitesse paramétrique, en utilisant des outils de géométrie convexe, de processus empiriques et de programmation linéaire semi-infinie.

Mots-clés. Pronfondeur de Tukey, ensembles de niveau, quantiles multivariés, convexité, fonction de support, distance de Hausdorff, programmation linéaire semi-infinie.

Abstract. Tukey depth has attracted much attention in multivariate statistics, because it allows to order the data, while extending the notion of univariate quantiles. Given a probability distribution and i.i.d. data from this distribution, we study the convergence of the level sets of the empirical depth towards those of the population depth. These level sets can be interpreted as multivariate quantiles. Under reasonable assumptions, we prove that the empirical level sets concentrate at the parametric rate, borrowing some tools from convex geometry, empirical process theory and semi-infinite linear programming.

Keywords. Tukey depth, level sets, multivariate quantiles, convexity, support function, Hausdorff distance, semi-infinite linear programming.

1 Introduction

Pour des données multivariées, il n'existe pas d'ordre canonique, contrairement au cas univarié, où beaucoup de procédures statistiques sont basées sur les statistiques d'ordre (tests de rang, valeurs extrêmes, etc.). C'est pourquoi plusieurs notions de *profondeur statistique* ont été proposées, afin d'ordonner les données de sorte que, dans un nuage de point multivarié, les points les plus centraux soient considérés comme profonds, et que les points périphériques soient considérés comme peu profonds. La notion de profondeur

statistique a été définie de manière rigoureuse par [14]. Le cas particulier de la profondeur de Tukey, définie par [12] dans un but général de décrire les données multivariées, est celui auquel nous nous intéressons ici. Cette notion de profondeur a été le sujet de beaucoup de travaux de recherche, avec des applications à la visualisation des données [7], la statistique robuste [3, 1], l'inférence non paramétrique [8], le bootstrap [13], la classification supervisée [5, 6], etc. Certaines propriétés asymptotiques de la profondeur de Tukey empirique ont été établies par [10], et la définition même de cette profondeur a été généralisée au cas de données fonctionnelles, ou non Euclidiennes [4]. Ici, nous nous intéressons aux ensembles de niveau de la profondeur de Tukey empirique et à leurs propriétés non asymptotiques, notamment leur concentration, au sens de la distance de Hausdorff, autour des ensembles de niveau théoriques. Nous montrons aussi pourquoi ces ensembles peuvent être interprétés comme des quantiles multivariés, et établissons un lien avec un objet important en géométrie convexe, appelé *corps flottant* ("floating body").

Cette présentation est basée sur le travail [2].

1.1 Définitions

Dans toute la suite, μ une mesure de probabilité sur $\mathbb{R}^d$ ($d \geq 1$ est fixé dans toute la suite). La profondeur de Tukey associée à μ est la fonction D_μ définie par

$$D_\mu(x) = \inf_{H \in \mathcal{H}: H \ni x} \mu(H), \quad \forall x \in \mathbb{R}^d,$$

où $\mathcal{H}$ est l'ensemble des demi-espaces fermés de $\mathbb{R}^d$. La profondeur d'un point est la plus petite masse d'un demi-espace fermé contenant ce point. Lorsque $X_1, \ldots, X_n$ sont des données dans $\mathbb{R}^d$ (où $n \geq 1$), la profondeur empirique est la fonction de profondeur $D_{\hat{\mu}_n}$, associée à la mesure empirique $\hat{\mu}_n$. Elle se réécrit de manière plus simple

$$D_{\hat{\mu}_n}(x) = \inf_{H \in \mathcal{H}: H \ni x} n^{-1} \#\{i = 1, \ldots, n : X_i \in H\}, \quad \forall x \in \mathbb{R}^d,$$

et s'interprète comme le nombre minimal de données incluses dans un demi-espace fermé contenant x.

On fixe un nombre $\alpha \in (0, 1)$ et on s'intéresse aux ensembles de niveau

$$G_\mu = \{x \in \mathbb{R}^d : D_\mu(x) \geq \alpha\} \quad \text{et} \quad G_{\hat{\mu}_n} = \{x \in \mathbb{R}^d : D_{\hat{\mu}_n}(x) \geq \alpha\}.$$

En particulier, que peut-on dire sur ces deux ensembles, et sur leur distance ? La distance que nous considérons est la distance de Hausdorff, définie par

$$\mathsf{d_H}(A, B) = \max\left(\max_{a \in A} \mathsf{d}(a, B), \max_{b \in B} \mathsf{d}(b, A)\right),$$

pour tous ensembles compacts non vides A et B dans $\mathbb{R}^d$: il s'agit de la plus grande distance d'un point d'un des deux ensembles à l'autre ensemble (ici, $\mathsf{d}(a, B)$ est la distance Euclidienne du point a à l'ensemble B).

La fonction de support d'un ensemble convexe compact (non vide) G est la fonction

$$h_G(u) = \max_{x \in G} u^\top x, \quad \forall u \in \mathbb{R}^d.$$

Lorsque $u \neq 0$, il s'agit simplement de la plus grande distance signée de l'origine à un hyperplan tangent à G orthogonal à u. Une propriété essentielle de la distance de Hausdorff est que si les ensembles A et B sont convexes, alors

$$\mathsf{d_H}(A, B) = \max_{u \in \mathbb{S}^{d-1}} |h_A(u) - h_B(u)|, \tag{1}$$

où $\mathbb{S}^{d-1}$ est la sphère Euclidienne unité de $\mathbb{R}^d$.

Pour tout $u \in \mathbb{R}^d \setminus \{0\}$, on note F_u la fonction de répartition de la projection de μ le long de u, i.e., $F_u(t) = \mathbb{P}[u^\top X \leq t]$, pour tout $t \in \mathbb{R}$. On note les quantiles gauche et droit d'ordre $1 - \alpha$ de la projection de μ le long de u comme suit:

$$q_u^\flat = \inf\{t \in \mathbb{R} : \mathbb{P}[u^\top X \leq t] \geq 1 - \alpha\} \quad \text{et} \quad q_u^\sharp = \sup\{t \in \mathbb{R} : \mathbb{P}[u^\top X \geq t] \geq \alpha\}.$$

On peut à présent définir les quantiles multivariés gauche et droit d'ordre $(1 - \alpha)$ de μ:

$$G^\eta_{\mathsf{MQ}} = \{x \in \mathbb{R}^d : \langle u, x \rangle \leq q_u^\eta, \forall u \in \mathbb{S}^{d-1}\}, \quad \eta \in \{\flat, \sharp\}. \tag{2}$$

Enfin, on rappelle la définition du corps flottant d'ordre α de μ:

$$G_{\mathsf{FB}} = \bigcap_{H \in \mathcal{H}: \mu(H) \geq 1-\alpha} H.$$

Cette définition apparaît essentiellement en géométrie convexe, dans la cas où μ est la mesure uniforme sur un ensemble convexe compact K: dans ce cas, G_{FB}, aussi appelé *corps flottant d'ordre α de K*, consiste en l'intersection de toutes les calottes de K (intersection de K avec un demi-espace fermé) de volume au moins égal à une fraction $1 - \alpha$ du volume total de K [11].

1.2 Le cas univarié

En dimension 1, la profondeur de Tukey se décrit de manière très simple. Pour tout $x \in \mathbb{R}$, $D_\mu(x) = \min\left(\mu\left((-\infty, x]\right), \mu\left([x, \infty)\right)\right)$. En particulier, si μ est absolument continue par-rapport à la mesure de Lebesgue, $D_\mu(x) = \min\left(F(x), 1 - F(x)\right)$, où F est la fonction de répartition de μ. L'ensemble G_μ n'est autre que $[q_\alpha^\flat, q_{1-\alpha}^\sharp]$, où, pour tout $\beta \in (0, 1)$, $q_\beta^\flat$ (resp. $q_\beta^\sharp$) est le quantile gauche (resp. droit) d'ordre β de μ. Remarquons que le quantile gauche d'ordre α de μ coïncide avec l'opposé du quantile droit d'ordre $1 - \alpha$ de $-X$, où $X \sim \mu$. Si $X_1, \ldots, X_n$ sont des données réelles, $G_{\hat{\mu}_n} = [X_{(\lceil n\alpha \rceil)}, X_{(\lfloor n(1-\alpha) \rfloor + 1)}]$, où $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$ correspond à l'échantillon réordonné. Dans ce cas, on observe immédiatement le lien entre l'ensemble de niveau de la profondeur de Tukey et la notion de quantiles.

1.3 Premières propriétés des ensembles de niveau

Le premier résultat que nous exposons ici montre le lien entre les quantiles multivariés, le corps flottant et l'ensemble de niveau de la profondeur de Tukey.

Lemme 1 $G_{FB} = G_{MQ}^{\flat} \subseteq G_{MQ}^{\sharp} = G_\mu$.

En particulier, si μ satisfait des conditions de continuité, ces quatre ensembles coïncident. Une autre conséquence de ce lemme est que l'ensemble G_μ (et, de même pour $G_{\hat{\mu}_n}$) est convexe, puisque, vu comme $G_{\mathsf{MQ}}^{\sharp}$, il est donné par des contraintes linéaires. Ainsi, et comme on le verra plus tard, sa fonction de support est donnée comme la valeur maximale d'une forme linéaire sous un nombre infini de contraintes linéaires:

$$h_{G_\mu}(u) = \max\left\{u^\top x : v \top x \leq q_v^{\sharp}, \forall v \in \mathbb{S}^{d-1}\right\}.$$

En particulier, $h_{G_\mu} \leq q_\cdot^{\sharp}$, avec égalité garantie dès lors que la fonction $q_\cdot^{\sharp}$ est convexe, ce qui n'est pas le cas en général (cf. [2, Proposition 1]).

1.4 Concentration des ensembles de niveau empiriques

Dans la suite, on considère un échantillon de données i.i.d. de loi μ, noté $X_1, \ldots, X_n$ et on cherche à borner, avec grande probabilité, $\mathsf{d}_{\mathsf{H}}(G_{\hat{\mu}_n}, G_\mu)$. Pour $u \in \mathbb{R}^d$, on note $\hat{q}_u$ le quantile empirique droit des données projetées $u^\top X_1, u^\top X_2, \ldots, u^\top X_n$.

On fait les hypothèses suivantes sur μ. Les nombres strictement positifs ε, L, r et R sont fixés et satisfont $\varepsilon < r \leq R$.

Hypothèse 1

- *Pour tout* $u \in \mathbb{S}^{d-1}$, F_u *est continue sur* $[q_u^{\sharp} - \varepsilon, q_u^{\sharp} + \varepsilon]$.
- $\left|F_u(t) - F_u(q_u^{\sharp})\right| \geq L\left|t - q_u^{\sharp}\right|$, *pour tout* $u \in \mathbb{S}^{d-1}$ *et* $t \in [q_u^{\sharp} - \varepsilon, q_u^{\sharp} + \varepsilon]$.
- *Il existe* $a \in \mathbb{R}^d$ *tel que* $B(a, r) \subseteq G_\mu \subseteq B(a, R)$.

La seconde hypothèse assure que $q_u^{\flat} = q_u^{\sharp}$ et que F_u n'est pas trop plate autour de $q_u^{\sharp}$, ce qui assure la concentration du quantile empirique à la vitesse paramétrique, pour tout $u \in \mathbb{S}^{d-1}$. Les deux premières hypothèses sont vérifiées pour une très grande classe de distributions. Par exemple, si μ admet une densité strictement positive par-rapport à la mesure de Lebesgue.

La troisième hypothèse implique, en particulier, que G_μ n'est pas vide. Il est bien connu que dès lors que $\alpha \leq 1/(d+1)$, G_μ ne peut pas être vide, mais que si $\alpha > 1/(d+1)$, il existe des distributions μ pour lesquelles $G_\mu = \emptyset$ [3]. En revanche, si μ est log-concave, alors on a la garantie que $G_\mu \neq \emptyset$ lorsque $\alpha \leq 1/e$ [9, Lemme 5.12].

Notre résultat principal est le suivant.

Théorème 1 *Supposons les hypothèses ci-dessus satisfaites et soit* $x \geq 0$ *tel que* $\frac{10\sqrt{5(d+1)}}{L} \leq x < \varepsilon\sqrt{n}$, *où* $C = \frac{R}{r}\frac{1+\varepsilon/r}{1-\varepsilon/r}$ *et soit* $A = e^{-250(d+1)}$. *Avec probabilité au moins* $1 - Ae^{-L^2x^2/2+10\sqrt{5(d+1)}x}$,

$$\mathsf{d_H}(G_{\hat{\mu}_n}, G_\mu) \leq \frac{Cx}{\sqrt{n}}.$$

En particulier, cela prouve que $\mathsf{d_H}(G_{\hat{\mu}_n}, G_\mu) = O_{\mathbb{P}}(n^{-1/2})$. La démonstration de ce théorème repose essentiellement sur les deux arguments suivants.

Lemme 2 *Pour tout* $z \in \mathbb{R}$ *satisfaisant* $\frac{10\sqrt{5(d+1)}}{L\sqrt{n}} \leq z < \varepsilon$,

$$\mathbb{P}\left[\sup_{u\in\mathbb{S}^{d-1}} |\hat{q}_u - q^\sharp_u| \leq z\right] \geq 1 - A\exp\left(-L^2z^2n/2 + 10\sqrt{5(d+1)}Lz\sqrt{n}\right).$$

Ce premier lemme repose sur la théorie des processus empiriques et sur le fait que, grâce à la seconde hypothèse,

$$\mathbb{P}\left[\sup_{u\in\mathbb{S}^{d-1}} |\hat{q}_u - q^\sharp_u| \leq z\right] \geq \mathbb{P}\left[\sup_{H\in\mathcal{H}_0} |\mu_n(H) - \mu(H)| \leq Lz\right],$$

où $\mathcal{H}_0$ est un sous-ensemble de la famille de tous les demi-espaces fermés de $\mathbb{R}^d$, dont la dimension de Vapnik-Chervonenkis est égale à $d+1$.

Le second lemme permet de contrôler $\sup_{u\in\mathbb{S}^{d-1}} |h_{G_{\hat{\mu}_n}}(u) - h_{G_\mu}(u)|$ à partir du contrôle de $\sup_{u\in\mathbb{S}^{d-1}} |\hat{q}_u - q^\sharp_u|$. C'est l'argument clef de la démonstration, qui s'appuie sur la programmation linéaire semi-infinie. En effet, d'après le Lemme 1, $h_{G_\mu}(u)$ (et de même pour $h_{G_{\hat{\mu}_n}}(u)$) est la valeur maximale de la forme linéaire $u^\top x$ sous les contraintes linéaires $v^\top x \leq q^\sharp_v, v \in \mathbb{S}^{d-1}$.

Pour toute fonction $\zeta : \mathbb{S}^{d-1} \to \mathbb{R}$, on note $K_\zeta = \{x \in \mathbb{R}^d : u^\top x \leq \zeta(u), \forall u \in \mathbb{S}^{d-1}\}$.

Lemme 3 *Soient* ϕ *et* $\hat{\phi}$ *deux fonctions continues sur* $\mathbb{S}^{d-1}$. *Supposons que* K_ϕ *and* $K_{\hat{\phi}}$ *sont d'intérieurs non vides et que* $B(a,r) \subseteq K_\phi \subseteq B(a,R)$, *pour un certain* $a \in \mathbb{R}^d$. *Soit* $\eta = \max_{u\in\mathbb{S}^{d-1}} |\hat{\phi}(u) - \phi(u)|$. *Alors, si* $\eta < r$, *on a*

$$\mathsf{d_H}(K_{\hat{\phi}}, K_\phi) \leq \frac{\eta R}{r}\frac{1+\eta/r}{1-\eta/r}.$$

On applique enfin ce lemme déterministe à $G_\mu = K_{q^\sharp_\cdot}$ et $G_{\hat{\mu}_n} = K_{\hat{q}^\sharp_\cdot}$, en remarquant que la fonction $\hat{q}^\sharp_\cdot : u \mapsto \hat{q}^\sharp_u$ est μ-presque sûrement continue, et que la fonction $q^\sharp_\cdot : u \mapsto q^\sharp_u$, quant à elle, est continue grâce aux deux premières hypothèses ci-dessus.

References

[1] Miguel A. Arcones, Zhiqiang Chen, and Evarist Giné. Estimators related to U-processes with applications to multivariate medians: asymptotic normality. *Ann. Statist.*, 22(3):1460–1477, 1994.

[2] Victor-Emmanuel Brunel. Concentration of the empirical level sets of tukey's halfspace depth. *Probability Theory and Related Fields*, 173(3-4):1165–1196, 2019.

[3] David L. Donoho and Miriam Gasko. Breakdown properties of location estimates based on halfspace depth and projected outlyingness. *Ann. Statist.*, 20(4):1803–1827, 1992.

[4] Subhajit Dutta, Anil K. Ghosh, and Probal Chaudhuri. Some intriguing properties of Tukey's half-space depth. *Bernoulli*, 17(4):1420–1434, 2011.

[5] Anil K. Ghosh and Probal Chaudhuri. On data depth and distribution-free discriminant analysis using separating surfaces. *Bernoulli*, 11(1):1–27, 2005.

[6] Anil K. Ghosh and Probal Chaudhuri. On maximum depth and related classifiers. *Scand. J. Statist.*, 32(2):327–350, 2005.

[7] Regina Y. Liu, Jesse M. Parelius, and Kesar Singh. Multivariate analysis by data depth: descriptive statistics, graphics and inference. *Ann. Statist.*, 27(3):783–858, 1999. With discussion and a rejoinder by Liu and Singh.

[8] Regina Y. Liu and Kesar Singh. A quality index based on data depth and multivariate rank tests. *J. Amer. Statist. Assoc.*, 88(421):252–260, 1993.

[9] László Lovász and Santosh Vempala. The geometry of logconcave functions and sampling algorithms. *Random Structures Algorithms*, 30(3):307–358, 2007.

[10] Jean-Claude Massé. Asymptotics for the Tukey depth process, with an application to a multivariate trimmed mean. *Bernoulli*, 10(3):397–419, 2004.

[11] Carsten Schütt and Elisabeth Werner. The convex floating body. *Math. Scand*, 66:275–290, 1990.

[12] John W. Tukey. Mathematics and the picturing of data. *Proceedings of the International Congress of Mathematicians (Vancouver, B. C., 1974), Vol. 2*, pages 523–531, 1975.

[13] Arthur B. Yeh and Kesar Singh. Balanced confidence regions based on Tukey's depth and the bootstrap. *J. Roy. Statist. Soc. Ser. B*, 59(3):639–652, 1997.

[14] Yijun Zuo and Robert Serfling. General notions of statistical depth function. *Ann. Statist.*, 28(2):461–482, 2000.

Statistical depth functions: Characterization and reconstruction problems

Stanislav Nagy[1]

[1] *Charles University, Faculty of Mathematics and Physics, Prague, Czech Republic, nagy@karlin.mff.cuni.cz*

Résumé. La profondeur est un concept de statistique non paramétrique qui généralise les rangs, ordonnancements et quantiles aux données multivariées. Comme pour les quantiles, il est souhaitable qu'une fonction de profondeur appropriée caractérise pleinement les mesures de probabilité, et qu'une mesure puisse être reconstruite efficacement à partir de sa profondeur. Nous considérons deux profondeurs importantes de la littérature — la profondeur demi-espace et la profondeur simpliciale — et explorons la théorie concernant leurs propriétés de caractérisation et de reconstruction.

Mots-clés. Caractérisation, profondeur du demi-espace, reconstruction.

Abstract. The depth is a concept of nonparametric statistics that generalizes ranks, orderings, and quantiles to multivariate data. Just as for quantiles, it is desired that a proper depth function fully characterizes probability measures, and that a measure is possible to be reconstructed efficiently from its depth. We consider two important depth functions from the literature — the halfspace and the simplicial depth — and explore the theory regarding their characterization and reconstruction properties.

Keywords. Characterization, halfspace depth, reconstruction.

1 Statistical depth function

The statistical depth function, or shortly just depth, is a remarkable tool of nonparametric statistics that extends the notions of orderings, rankings, and quantiles to multivariate datasets. Let $\mathcal{P}\left(\mathbb{R}^d\right)$ denote the collection of Borel probability measures on the Euclidean space $\mathbb{R}^d$. A *depth* is, formally speaking, a mapping

$$D\colon \mathbb{R}^d \times \mathcal{P}\left(\mathbb{R}^d\right) \to [0,1]\colon x \mapsto D(x;P)$$

that is intended to assess the centrality of the point $x \in \mathbb{R}^d$ with respect to (w.r.t.) the measure $P \in \mathcal{P}\left(\mathbb{R}^d\right)$. Unlike a density, a depth should be global, and should quantify the degree of fit of x within the shape of the whole distribution P. Points x that are positioned "centrally" inside P attain high depth values; points x on the periphery of P or outlying points x get $D(x;P)$ close to zero. The point (or a representative of the set of points) in $\mathbb{R}^d$ that maximizes $D(\cdot;P)$ is called the *depth median* of P, and it is expected

that this location parameter is robust, properly equivariant, and represents a plausible extension of the standard univariate median to multivariate data. The upper level sets of the depth $\left\{x \in \mathbb{R}^d \colon D(x;P) \geq \delta\right\}$ are called the *δ-central regions* of P. They represent multivariate analogues of the inter-quantile regions from $\mathbb{R}$.

The history of the depth in statistics dates to Tukey (1975), even though in other fields of mathematics equivalent notions have been explored since much earlier, for references see Nagy et al. (2019). Currently, there are dozens of viable depth functions with many applications in statistics (Zuo and Serfling, 2000). We mention two important depths.

Halfspace depth. The classical and arguably the most important is the *halfspace depth* (also called the *Tukey depth*, due to Tukey, 1975). For $x \in \mathbb{R}^d$ and $P \in \mathcal{P}\left(\mathbb{R}^d\right)$ it is defined as the minimum P-mass on any side of a hyperplane containing x

$$HD(x;P) = \inf_{u \in \mathbb{S}^{d-1}} \mathsf{P}\left(\langle X, u\rangle \leq \langle x, u\rangle\right),$$

where $\mathbb{S}^{d-1}$ is the unit sphere in $\mathbb{R}^d$, and $X \sim P$. The central regions of HD are convex sets; for P an empirical measure they are convex polytopes (Rousseeuw and Ruts, 1999).

Simplicial depth. The second prominent depth we mention is the *simplicial depth* of Liu (1990), defined for $x \in \mathbb{R}^d$ and $P \in \mathcal{P}\left(\mathbb{R}^d\right)$ as

$$SD(x;P) = \mathsf{P}\left(x \in \mathsf{S}\left(X_1, \ldots, X_{d+1}\right)\right),$$

where $\mathsf{S}\left(x_1, \ldots, x_{d+1}\right)$ is the closed convex hull of the points $x_1, \ldots, x_{d+1}$, and $X_1, \ldots, X_{d+1}$ are independent random variables sampled from P. The set $\mathsf{S}\left(X_1, \ldots, X_{d+1}\right)$ is a simplex in $\mathbb{R}^d$ with random vertices. The depth SD is thus the probability of x falling into a random P-simplex. The central regions of SD are not necessarily convex. Unlike the halfspace depth, the simplicial depth also does not have to be maximized in a convex set, and its value does not necessarily decrease monotonically as one moves from a simplicial median along a straight line.

For an application of HD and SD to a bivariate dataset see Fig. 1.

2 The characterization problem

We are interested in two problems from the theory of the statistical depth:

Characterization. Can two different measures $P, Q \in \mathcal{P}\left(\mathbb{R}^d\right)$ share the same depth function $D(x;P) = D(x;Q)$ for all $x \in \mathbb{R}^d$?

Reconstruction. Can P be reconstructed from its depth function $x \mapsto D(x;P)$?

In this contribution, we provide an overview of the theoretical results relevant to these two questions. For the case of the halfspace depth, many authors have already dealt with these problems, for a list of references see Nagy (2020) and Laketa and Nagy (2021). For the simplicial depth, as far as we know, these problems appear to be unexplored.

Our main findings can be summarized as:

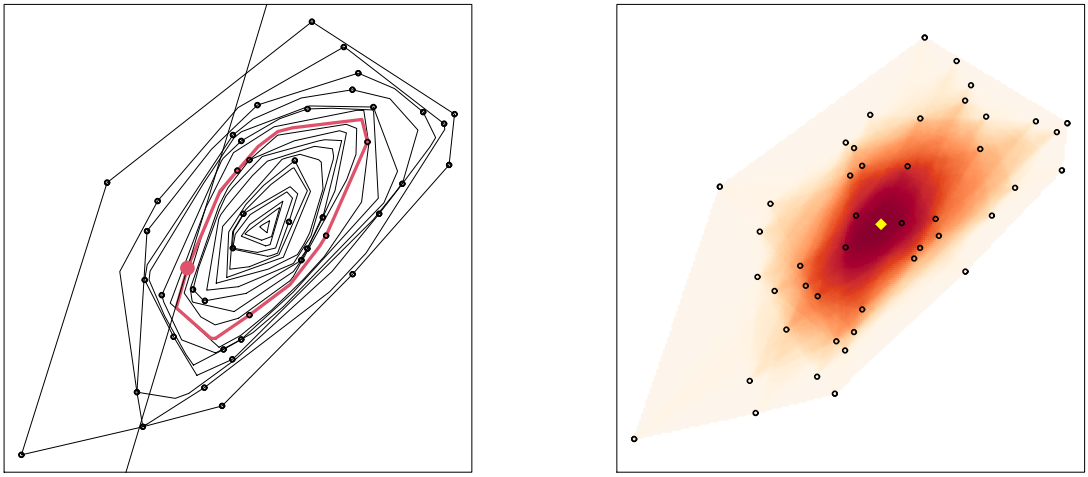


Figure 1: A dataset P_n of $n = 54$ points in the plane (black dots) and the contours of the corresponding halfspace (left panel) and simplicial (right panel) depth. For HD we see the complete collection of contours. The halfspace median set of P_n is the smallest polygon in the centre of the figure. For the pink point $x \in \mathbb{R}^2$ we display also a halfplane from $\mathcal{H}(x)$ whose P_n-mass equals the halfspace depth of x, which is $HD(x; P_n) = 7/54$. In the right panel, the simplicial median is plotted as a yellow diamond. Darker shades of brown represent regions of higher simplicial depth $SD(\cdot; P_n)$.

- The halfspace depth does not characterize $\mathcal{P}\left(\mathbb{R}^d\right)$ in dimension $d > 1$.
- Both halfspace and simplicial depth characterize empirical measures $P_n \in \mathcal{P}\left(\mathbb{R}^d\right)$.
- The n sample points of P_n can be reconstructed efficiently from its depth function.

In what follows, we outline an interesting argument regarding the characterization problem for the halfspace depth. The other issues, including several major open problems related to our questions, will be discussed during the talk.

2.1 Spherically symmetric measures

Recall that $P \in \mathcal{P}\left(\mathbb{R}^d\right)$ is called *spherically symmetric* if the distribution of $X \sim P$ is invariant w.r.t. orthogonal transformations. Examples of spherically symmetric distributions are the standard normal distribution, the uniform distribution on the unit ball in $\mathbb{R}^d$, or the uniform distribution on the surface of the unit sphere $\mathbb{S}^{d-1}$. A measure $P \in \mathcal{P}\left(\mathbb{R}^d\right)$ with $X = (X_1, \ldots, X_d) \sim P$ is spherically symmetric if and only if

$$\langle X, u \rangle \stackrel{d}{=} X_1 \text{ for all } u \in \mathbb{S}^{d-1}, \tag{1}$$

where $\stackrel{d}{=}$ means that the two random variables are equal in distribution.

The halfspace depth of any spherically symmetric measure $P \in \mathcal{P}\left(\mathbb{R}^d\right)$ takes a particularly simple form. Let $X = (X_1, \ldots, X_d) \sim P$, and denote by $F_1 \colon \mathbb{R} \to [0,1]$ the cumulative distribution function of X_1. Direct computation gives

$$\begin{aligned} HD(x;P) &= \inf_{u \in \mathbb{S}^{d-1}} \mathsf{P}\left(\langle X, u\rangle \leq \langle x, u\rangle\right) = \inf_{u \in \mathbb{S}^{d-1}} \mathsf{P}\left(X_1 \leq \langle x, u\rangle\right) \\ &= \inf_{u \in \mathbb{S}^{d-1}} F_1\left(\langle x, u\rangle\right) = F_1\left(\inf_{u \in \mathbb{S}^{d-1}} \langle x, u\rangle\right) = F_1\left(-\left\|x\right\|\right). \end{aligned} \tag{2}$$

We have used the characterization (1), the fact that the distribution function F_1 is non-decreasing, and the Cauchy-Schwarz inequality that gives $\langle x, u\rangle \geq -\left\|x\right\| \left\|u\right\|$, with equality for $u = -x/\left\|x\right\|$ provided that $\left\|x\right\| \neq 0$. We obtain that the halfspace depth of any spherically symmetric measure depends only on the Euclidean norm of x and the distribution function of (any) univariate marginal of X.

2.2 α-symmetric measures

A useful extension of the previous result is obtained within the realm of α-symmetric measures. For $\alpha \in (0, \infty]$ and $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$ set

$$\left\|x\right\|_\alpha = \begin{cases} \left(\sum_{i=1}^d \left|x_i\right|^\alpha\right)^{1/\alpha} & \text{if } \alpha < \infty, \\ \max_{i=1,\ldots,d} \left|x_i\right| & \text{for } \alpha = \infty. \end{cases}$$

A distribution $P \in \mathcal{P}\left(\mathbb{R}^d\right)$ with $X \sim P$ is said to be *α-symmetric* if its characteristic function takes the form

$$\psi_X(t) = \mathsf{E}\, \exp\left(i\, \langle t, X\rangle\right) = \int_{\mathbb{R}^d} \exp\left(i\, \langle t, x\rangle\right)\, \mathrm{d}\, P(x) = \xi(\left\|t\right\|_\alpha) \text{ for all } t \in \mathbb{R}^d$$

for some function $\xi \colon [0, \infty) \to \mathbb{R}$. It is easy to show that for $\alpha = 2$, the 2-symmetric measures are exactly the spherically symmetric distributions. An important generalization of the property (1) from $\alpha = 2$ is the fact that a measure $P \in \mathcal{P}\left(\mathbb{R}^d\right)$ is α-symmetric if and only if for $X = (X_1, \ldots, X_d) \sim P$

$$\langle X, u\rangle \stackrel{d}{=} \left\|u\right\|_\alpha X_1 \text{ for all } u \in \mathbb{S}^{d-1}.$$

Denote again by F_1 the distribution function of X_1. For the halfspace depth of an α-symmetric measure P we can write

$$\begin{aligned} HD(x;P) &= \inf_{u \in \mathbb{S}^{d-1}} \mathsf{P}\left(\langle X, u\rangle \leq \langle x, u\rangle\right) = \inf_{u \in \mathbb{S}^{d-1}} \mathsf{P}\left(\left\|u\right\|_\alpha X_1 \leq \langle x, u\rangle\right) \\ &= \inf_{u \in \mathbb{S}^{d-1}} F_1\left(\frac{\langle x, u\rangle}{\left\|u\right\|_\alpha}\right) = F_1\left(\inf_{u \in \mathbb{S}^{d-1}} \frac{\langle x, u\rangle}{\left\|u\right\|_\alpha}\right) = F_1\left(-\left\|x\right\|_\beta\right). \end{aligned} \tag{3}$$

We have proceeded as in (2), except for the last step where we used a generalized version of the Hölder inequality $\langle x, u\rangle \geq -\|u\|_\alpha \|x\|_\beta$ for $\beta = \alpha/(\alpha-1)$ the conjugate exponent to α. This generalized Hölder inequality holds also for $\alpha \leq 1$ in which case $\beta = \infty$. This can be proved using arguments from convexity theory (Schneider, 2014, Section 1.6). The inequality follows by considering the polar set to the unit ball $B_\alpha = \left\{x \in \mathbb{R}^d \colon \|x\|_\alpha \leq 1\right\}$, which turns out to be its dual ball B_β. The general Hölder inequality becomes an equality for a specially selected $u \in \mathbb{S}^{d-1}$, which gives the last equality in (3).

2.3 A counterexample

We now know that the halfspace depth of an α-symmetric distribution P with $\alpha \leq 1$ depends only on its first marginal distribution function F_1, and its argument $x \in \mathbb{R}^d$ via its L^∞-norm $\|x\|_\infty$. Two α-symmetric distributions $P, Q \in \mathcal{P}\left(\mathbb{R}^d\right)$ with $\alpha \leq 1$ have therefore the same halfspace depth if their first marginal distributions coincide. Interestingly, examples of α-symmetric distributions with this property have been known since the 1930s and the work of Lévy and Schoenberg on positive definite functions, for references see Nagy (2021). Consider the function

$$\psi_{d,\alpha,\gamma} \colon \mathbb{R}^d \to \mathbb{R} \colon t \mapsto \exp\left(-\|t\|_\alpha^\gamma\right).$$

For any integer $d \geq 2$, $\alpha \in (0, 2]$ and $\gamma \in (0, \alpha]$ is this function known to be positive definite. Bochner's theorem thus gives that $\psi_{d,\alpha,\gamma}$ corresponds to a characteristic function $\mathsf{E} \exp\left(i \langle t, X\rangle\right) = \psi_{d,\alpha,\gamma}(t)$, $t \in \mathbb{R}^d$, of a random vector $X = (X_1, \ldots, X_d) \sim P \in \mathcal{P}\left(\mathbb{R}^d\right)$. The measure P is certainly α-symmetric. The characteristic function ψ_{X_1} of X_1 is

$$\psi_{X_1}(t_1) = \mathsf{E} \exp\left(i \langle (t_1, 0, \ldots, 0), X\rangle\right) = \psi_{d,\alpha,\gamma}\left((t_1, 0, \ldots, 0)\right) = \exp\left(-|t_1|^\gamma\right) \text{ for } t_1 \in \mathbb{R}.$$

The function ψ_{X_1} determines the distribution of X_1 uniquely, and thus for any $\alpha \in (0, 2]$ the marginal distribution function $F_{1,\gamma}$ of X_1 depends only on the exponent $\gamma \in (0, \alpha]$.

For any $d \geq 2$ and $\gamma \in (0, 1)$ let $P_\alpha \in \mathcal{P}\left(\mathbb{R}^d\right)$ correspond to the characteristic function $\psi_{d,\alpha,\gamma}$. Taking the collection $\{P_\alpha\}_{\alpha \in [\gamma, 1]} \subset \mathcal{P}\left(\mathbb{R}^d\right)$ we obtain

$$HD(x; P_\alpha) = F_{1,\gamma}\left(-\|x\|_\infty\right) \text{ for all } x \in \mathbb{R}^d \text{ and } \alpha \in [\gamma, 1].$$

For any integer $d > 1$ and $\gamma \in (0, 1)$ we have found an uncountable collection of measures $\{P_\alpha\}_{\alpha \in [\gamma, 1]}$ sharing the same halfspace depth. Two such distributions and their halfspace depth contours are visualised in Fig. 2.

Acknowledgement. The work was supported by Czech Science Foundation (EXPRO project n. 19-28231X).

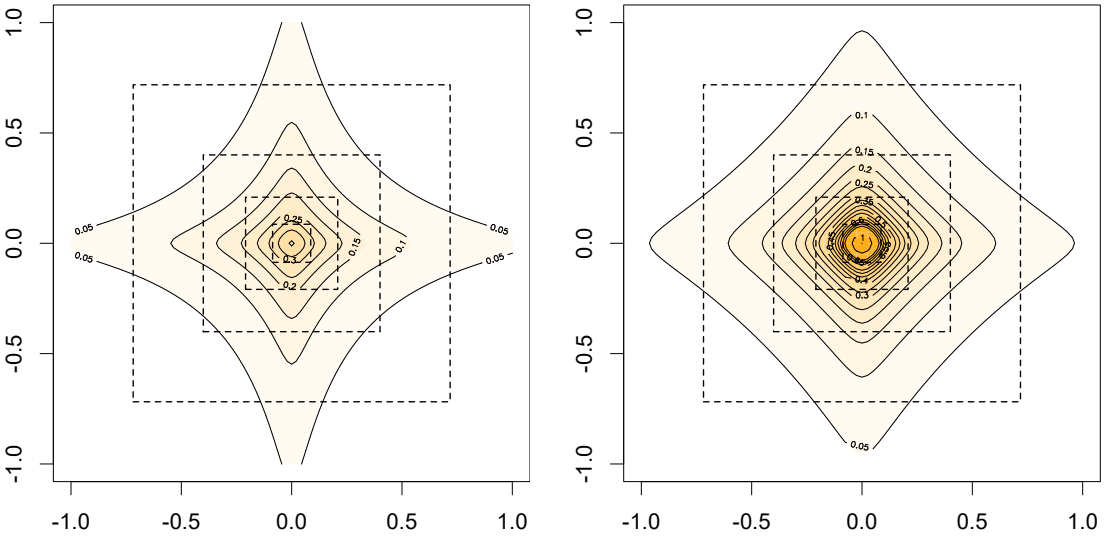


Figure 2: Several numerically computed contours (solid lines) of the densities of distributions $P_{1/2}$ (left panel) and P_1 (right panel) with $d = 2$ and $\gamma = 1/2$. Their exact contours of HD at levels 0.3, 0.35, 0.4, and 0.45 are plotted in dashed lines in both panels. The two densities are clearly different, but their halfspace depths coincide in all $\mathbb{R}^2$.

Bibliography

Laketa, P. and Nagy, S. (2021). Reconstruction of atomic measures from their halfspace depth. *J. Multivariate Anal.*, 183:104727.

Liu, R. Y. (1990). On a notion of data depth based on random simplices. *Ann. Statist.* 18 405–414.

Nagy, S. (2021). Halfspace depth does not characterize probability distributions. *Statist. Papers*, 62, 1135–1139.

Nagy, S. (2020). The halfspace depth characterization problem. In La Rocca, M., Liseo, B., and Salmaso, L., editors, *Nonparametric Statistics*, pages 379–389, Springer, Cham.

Nagy, S., Schütt, C., and Werner, E. M. (2019). Halfspace depth and floating body. *Stat. Surv.*, 13:52–118.

Rousseeuw, P. J. and Ruts, I. (1999). The depth function of a population distribution. *Metrika*, 49(3):213–244.

Schneider, R. (2014). *Convex bodies: the Brunn-Minkowski theory*, volume 151 of Encyclopedia of Mathematics and its Applications. Cambridge University Press, Cambridge.

Tukey, J. W. (1975). Mathematics and the picturing of data. In Proceedings of the *International Congress of Mathematicians* (Vancouver, B. C., 1974), Vol. 2, pages 523–531. Canad. Math. Congress, Montreal, Que.

Zuo, Y. and Serfling, R. (2000). General notions of statistical depth function. *Ann. Statist.*, 28(2):461–482.

SFdS - BFA

Multi-normex for evaluating the distribution of aggregated heavy tailed risks

Marie Kratz [1] & Evgeny Prokopenko [2]

[1] *ESSEC Business School, CREAR; kratz@essec.edu*
[2] *Sobolev Institute of Mathematics, Novosibirsk; prokopenko@math.nsc.ru*

Abstract. We build a sharp approximation of the whole distribution of the sum of iid heavy-tailed random vectors, combining mean and extreme behaviors. It extends the so-called 'normex' approach from a univariate to a multivariate framework. We propose two possible multi-normex distributions, named d-Normex and MRV-Normex. Both rely on the Gaussian distribution for describing the mean behavior, via the CLT, while the difference between the two versions comes from using the exact distribution or the EV theorem for the maximum. The main theorems provide the rate of convergence for each version of the multi-normex distributions towards the distribution of the sum, assuming second order regular variation property for the norm of the parent random vector when considering the MRV-normex case. Numerical illustrations and comparisons are proposed with various dependence structures on the parent random vector, using QQ-plots based on geometrical quantiles.

Keywords. aggregation; central limit theorem; dependence; extreme value theorem; geometrical quantiles; multivariate extremes; multivariate regular variation; (multivariate) Pareto distribution; ordered statistics; QQ-plots; rate of convergence; second order regular variation; sum of random vectors

Introduction

Motivation. Looking for the most accurate possible evaluation of the distribution of the sum of random variables, or vectors, or processes, with unknown distributions, has always been a classical problem in the probabilistic and statistical literature, with various answers depending on the given framework and on the specific application in view. On one hand, (uni- or multivariate) Central Limit Theorems (CLT) or Functional ones prove, under finite variance for the sum components or/and additional conditions, the asymptotic Gaussian behavior of the sum with some rate of convergence, focusing on the 'body' of the distribution. When considering heavy-tailed marginal distributions, Generalized CLT with the convergence to stable distributions, handle the case of infinite variance (see e.g. Samorodnitsky and Taqqu (1994) and Petrov (1995), and references therein), while, in the case of finite variance, an alternative way is to consider trimmed sums, removing extremes from the sample, to improve the rate of convergence; see e.g. Mori (1984) and

Hahn (1991) and references therein.

When interested in tail distributions, CLTs may give poor results, especially when considering heavy tails. That is why different approaches have been developed, among which: (i) large deviation theorems (see e.g. Petrov (1975) and Borovkov (2020) for light tails, and Mikosch and Nagaev (1998), Foss et al. (2013), and Lehtomaa (2017) for heavy tails), (ii) extreme value theorems (EVT) focusing on the tail only (see e.g. Embrechts et al. (1997), de Haan and Ferreira (2006), and Resnick (2007)), (iii) hybrid distributions combining (asymptotic) distributions for both the main and extreme behaviors when considering independent random variables (see e.g. Csörgö et al. (1988), Zaliapin et al. (2005), Kratz (2014), and Müller (2019); we use the name given in Kratz (2014) for this type of hybrid distribution/ method/approach, namely Normex distribution/method/approach.

Recall briefly the idea of Normex (for Norm(al)-Ex(tremes)) method. It consists of rewriting the sum of n random variables as the sum of their ordered statistics, and splitting it into two main parts, a trimmed sum removing the extremes, and the extremes. Using that the trimmed sum of the first $n-k-1$ ordered statistics is conditionnally independent of the k largest order statistics, given the $(n-k)$-th order statistics, we can express the distribution of the sum, integrating w.r.t. to the $(n-k)$-th order statistics and using a CLT for the conditional trimmed sum, and an EVT one for the k largest order statistics. Note that a benefit of Normex approach is that it does not require any condition on the existence of moments, as the CLT applies on truncated random variables.

The following example (as developed in Kratz (2014)), simulating identically distributed and independent (iid) Pareto(α) (such that $\overline{F}(x) = x^{-\alpha}$ for $x > 1$) random variables (rv), with $\alpha = 2.3$ (finite variance, but no third moment), illustrates perfectly the adding value of combining main and extreme behaviors.

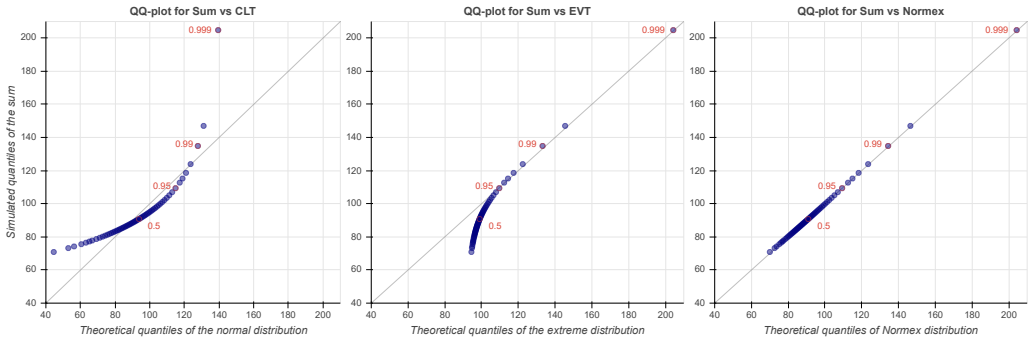


Figure 1: QQ-plots for the empirical distribution (sample size $= 10^7$) of the sum of 52 iid Pareto($\alpha = 2.3$) rv and three different approximations of the sum distribution using: CLT, EVT and Normex, respectively. The red circles and numbers denote the extreme quantiles and their levels.

Indeed, on the QQ-plots given in Figure 1, we can compare the fit of the distribution of the empirical sample with the following three distributions: the Gaussian one (CLT

approach), the Fréchet one (EVT approach), the hybrid one (*i.e.* Normex, combining the CLT and the distribution of the maximum). For the hybrid Normex distribution, we may consider either the exact distribution of the maximum, or its asymptotic approximation, the Fréchet distribution. Since both provide the same plot, we display it in the third plot on the right. Note that we choose a rather small number of components in the sum, $n = 52$, also to illustrate the speed of convergence when using asymptotic theorems. We observe that the CLT approach does not provide a sharp evaluation, even in the body of the distribution, due to this choice of n, which cannot compensate yet the fact that the distribution of the rv is asymmetric and skewed; increasing n will of course improve the fit in the body of the distribution. Given the fact that the Pareto distribution belongs to the Fréchet maximum domain of attraction, using the Fréchet distribution for the distribution of the sum of Pareto rv's gives a very sharp approximation in the tail (from the 93% quantile), but not for the average behavior, as expected. Finally, a perfect match between empirical quantiles and Normex ones is observed for the whole distribution in the right plot, even for a small number of summands.

Goal of the study. It is natural to extend the normex approach to a multivariate framework. With this goal of proposing a multi-normex method and distribution, we consider iid random vectors $\mathbf{X}_1, \ldots, \mathbf{X}_n$, with parent random vector $\mathbf{X}$ having a heavy-tailed d-dimensional distribution $F_{\mathbf{X}}$ and density $f_{\mathbf{X}}$ (when existing). Note that there are different ways to define multivariate extremes (see e.g. Chap.8 in Beirlant et al. (2004)). The chosen way in this paper is w.r.t. the norm $\|\cdot\|$ in $\mathbb{R}^d$, meaning that the ordered (w.r.t. the norm) vector of $(\mathbf{X}_1, \ldots, \mathbf{X}_n)$, denoted by $(\mathbf{X}_{(1)}, \ldots, \mathbf{X}_{(n)})$, satisfies

$$\|\mathbf{X}_{(1)}\| \leq \|\mathbf{X}_{(2)}\| \leq \cdots \leq \|\mathbf{X}_{(n)}\|.$$

We propose two versions of multi-normex. The first one, named *d-Normex*, is a natural extension to any dimension d of the univariate ($d = 1$) normex method as developed in Kratz (2014): We approximate the distribution of the trimmed sum via the CLT and consider the distribution of the maximum $\mathbf{X}_{(n)}$. This latter distribution is approximated via the Extreme Value (EV) theorem in the second multi-normex version, named *MRV-Normex*.

Aiming at proving the benefit of using a multi-normex distribution for a better fit of the whole (unknown) distribution F, assuming F heavy-tailed in the sense of $\|\mathbf{X}\| \in \mathcal{RV}_{-\alpha}$, *i.e.* regularly varying rv with $\alpha > 0$, we focus analytically on the case $\alpha \in (2; 3]$ (when $\|\mathbf{X}\|$ has a finite second moment, but no third moment), to compare the rates of convergence when using the CLT and the multi-normex approach, respectively.

Note our focus on heavy tailed distributions (i.e. distributions belonging to the max domain of attraction of Fréchet), where the impact of using Normex distribution will be much stronger than in the light tail case (because of the one big jump principle), in particular for risk analysis and management. We prove that the normex approach leads, as

expected, to a better speed of convergence for evaluating the distribution of the sum than the CLT does, for such type of heavy-tailed distributions. When varying the fatness of the tail measured by $\alpha > 0$, we draw this comparison numerically, using e.g. geometrical multivariate quantiles (Chaudhuri (1996), Dhar et al. (2014)).

Bibliographie

Beirlant J., Y. Goegebeur, J. Segers and J.L. Teugels (2004). Statistics of extremes: theory and applications, *Wiley Series in Probability and Statistics*

Borovkov, A. A. (2020). Asymptotic Analysis of Random Walks: Light-Tailed Distributions. *Encyclopedia of Mathematics and its Applications*. Cambridge University Press

Chaudhuri, P. (1996). On a Geometric Notion of Quantiles for Multivariate Data, *Journal of the American Statistical Association*, 91, 862–872

Csörgö, S., E. Haeusler, and D. M. Mason (1988). A probabilistic approach to the asymptotic distribution of sums of independent, identically distributed random variables. *Advances in Applied Mathematics*, 9, 259-333

Dhar, S., B. Chakraborty and P. Chaudhuri (2014). Comparison of multivariate distributions using quantile-quantile plots and related tests, *Bernoulli*, 3, 1484–1506

Embrechts, P., C. Kluppelberg, and T. Mikosch (1997). Modelling Extreme Events for Insurance and Finance. *Stochastic Modelling and Applied Probability*, Springer-Verlag

Foss, S., D. Korshunov, and S. Zachary (2013). An Introduction to Heavy-Tailed and Subexponential Distributions. *Springer Series in Operations Research and Financial Engineering*

de Haan, L. and A. Ferreira (2006). Extreme Value Theory: An Introduction. *Springer Series in Operations Research and Financial Engineering*, Springer

Hahn, M. G. (1991). Sums, Trimmed Sums and Extremes. Ed. by M. G. Hahn, D. M. Mason, and D. C. Weiner. *Progress in Probability*, 23, Birkhäuser Boston

Kratz, M. (2014). Normex, a new method for evaluating the distribution of aggregated heavy tailed risks. Application to risk measures. *Extremes*, 17, 661-691

Kratz, M. and E. Prokopenko (2021). Multi-Normex Distributions for the sum of random vectors. Rates of Convergence. *ESSEC Working Paper 2102* and *arXiv:2107.09409v1*

Lehtomaa, J. (2017). Large deviations of means of heavy-tailed random variables with finite moments of all orders. *Journal of Applied Probability*, 54, 66-81

Mikosch, T. and A. Nagaev (1998). Large Deviations of Heavy-Tailed Sums with Applications in Insurance. *Extremes* 1, 81-110

Mori, T. (1984). On the limit distributions of lightly trimmed sums. *Mathematical Proceedings of the Cambridge Philosophical Society*, 96, 507-516

Müller, U. K. (2019). Refining the central limit theorem approximation via extreme value theory. *Statistics and Probability Letters*, 155, 1–7

Petrov, V. V. (1975). Sums of Independent Random Variables. *Ergebnisse der Mathematik und ihrer Grenzgebiete*, Springer
Petrov, V.V. (1995). Limit Theorem of Probability Theory: Sequences of Independent Random Variables. *Oxford Studies in Probability*
Resnick, S.I. (2007). Heavy Tail Phenomena: Probabilistic and Statistical Modeling. *Springer Series in Operations Research and Financial Engineering*, Springer
Samorodnitsky, G. and M. S. Taqqu (1994). Stable non-Gaussian Random Processes: Stochastic Models with Infinite Variance. *Chapman and Hall*
Zaliapin, I. V., Y. Y. Kagan, and F. P. Schoenberg (2005). Approximating the Distribution of Pareto Sums. *Pure and Applied Geophysics*, 162, 1187-1228

Lundberg-Aumann-Serrano dangerousness risk index

Etienne Marceau, Professeur titulaire en actuariat, PhD, ASA [1]

[1] *École d'actuariat, Faculté des Sciences et de Génie, Université Laval, Pavillon Paul-Comtois, 2425, rue de l'Agriculture, Bureau 4177, Québec (Québec), G1V 0A6, Canada et etienne.marceau@act.ulaval.ca*

Résumé. Considérons une entité exposée à un risque dont on modélise l'évolution temporelle à l'aide d'un processus aléatoire en temps discret. Par exemple, en mathématiques actuarielles, cette entité est une compagnie d'assurance et le risque correspond aux pertes nettes globales d'un portefeuille de cette compagnie. Une gestion efficace nécessite une mesure de risque appropriée en regard de l'exposition à ce risque. Une mesure de risque vise à atteindre un ou plusieurs des objectifs suivants : (1) décrire le risque d'un événement le plus précisément possible ; (2) communiquer le niveau de risque ; (3) établir des comparaisons entre des risques ; (4) identifier le risque le plus grave ; (5) traduire l'évolution du risque dans le temps ; (6) mesurer/évaluer l'efficacité de stratégies pour réduire/prévenir le risque ; (7) recommander et prendre des décisions selon un niveau de risque donné. Dans le cadre du modèle classique de risque (en mathématiques actuarielles ; proposé par De Finetti), le processus de risque est une marche aléatoire avec dérive négative et avec accroissements indépendants et identiquement distribués. Dans un premier temps, on introduit l'indice de risque de Lundberg-Aumann-Serrano comme une fonctionnelle de la processus de risque en temps discret et prenant valeurs dans les réels positifs. L'indice de risque de Lundberg-Aumann-Serrano est défini comme l'inverse multiplicatif du coefficient d'ajustement de Lundberg. Ce coefficient se retrouve dans plusieurs applications des probabilités appliquées. Via cet indice de risque, on vise à quantifier et analyser l'exposition au risque de l'entité en question sous une perspective qui se trouve à la frontière de la théorie des mesures de risque et de la théorie de la ruine. On démontre que l'indice de risque de Lundberg-Aumann-Serrano satisfait des propriétés souhaitables pour une mesure de risque en regard des objectifs mentionnés, telles que la monotonicité, l'homogénéité, la convexité et la sous-additivité. On prouve aussi que l'indice de risque possède les propriétés d'être cohérent par rapport aux ordres en dominance stochastique et convexe croissant pour les processus. On discute de l'application de l'indice de risque de Lundberg-Aumann-Serrano pour des processus de risque définis à partir de marches aléatoires dont les accroissement peuvent être dépendants. On démontre l'impact de cette dépendance temporelle sur l'indice de Lundberg-Aumann-Serrano. On utilise le principe d'Euler pour aborder la question du calcul des contributions des composantes des modèles à l'indice Lundberg-Aumann-Serrano. On conclut l'exposé avec des exemples numériques pour illustrer l'applicabilité de l'indice de risque de Lundberg-Aumann-Serrano.

Mots-clés. mesures de risque, théorie du risque, processus aléatoires en temps discret, marches aléatoires, coefficient de Lundberg, principe des déviations larges, ordres stochastiques, probabilités appliquées.

Abstract. Consider an entity exposed to a risk whose temporal evolution is modeled using a random process in discrete time. For example, in actuarial mathematics, this entity is an insurance company and the risk corresponds to the overall net losses of a portfolio of this company. Effective management requires appropriate risk measurement with respect to exposure to that risk. A risk measure aims to achieve one or more of the following objectives: (1) describe the risk of an event as precisely as possible; (2) communicate the level of risk; (3) making comparisons between risks; (4) identify the most serious risk; (5) reflect the evolution of risk over time; (6) measure/evaluate the effectiveness of strategies to reduce/prevent risk; (7) recommend and make decisions based on a given level of risk. In the classical risk model (in actuarial mathematics; proposed by De Finetti), the risk process is a random walk with negative drift and with independent and identically distributed increments. First, we introduce the Lundberg-Aumann-Serrano risk index as a functional of the risk process in discrete time and taking values in positive reals. The Lundberg-Aumann-Serrano risk index is defined as the multiplicative inverse of the Lundberg adjustment coefficient. This coefficient is found in several applications of applied probabilities. Via this risk index, we aim to quantify and analyze the risk exposure of the entity in question from a perspective that is at the border of the theory of risk measures and the theory of ruin. It is demonstrated that the Lundberg-Aumann-Serrano risk index satisfies desirable properties for a risk measure with respect to the mentioned objectives, such as monotonicity, homogeneity, convexity and sub-additivity. We also prove that the risk index has the properties of being consistent with respect to orders in stochastic and convex increasing dominance for the processes. We discuss the application of the Lundberg-Aumann-Serrano risk index for risk processes defined from random walks whose increments can be dependent. We demonstrate the impact of this temporal dependence on the Lundberg-Aumann-Serrano index. We use Euler's principle to address the issue of calculating the contributions of model components to the Lundberg-Aumann-Serrano index. We conclude the presentation with numerical examples to illustrate the applicability of the Lundberg-Aumann-Serrano risk index.

Keywords. risk measures, risk theory, random processes in discrete time, random walks, Lundberg coefficient, large deviations, stochastic orders, applied probability

Annulé

Three-Step Risk Inference In Insurance Ratemaking

Yanxi Hou[1], Seul Ki Kang[2], Chia Chun Lo[2] and Liang Peng[2]

[1]*School of Data Science, FuDan University, China*
[2]*Department of Risk Management and Insurance, Georgia State University, Atlanta, GA 30303, USA*

Abstract. As catastrophic events happen more and more frequently, accurately forecasting risk at a high level is vital for the financial stability of the insurance industry. This paper proposes an efficient three-step procedure to deal with the semicontinuous property of insurance claim data and forecast extreme risk. The first step uses a logistic regression model to estimate the nonzero claim probability. The second step employs a quantile regression model to select a dynamic threshold for fitting the loss distribution semiparametrically. The third step fits a generalized Pareto distribution to exceedances over the selected dynamic threshold. Combining these three steps leads to an efficient risk forecast. Furthermore, a random weighted bootstrap method is employed to quantify the uncertainty of the derived risk forecast. Finally, we apply the proposed method to an automobile insurance data set.

Keywords. Generalized Pareto distribution, Insurance loss, Logistic regression, Quantile regression, Random weighted bootstrap

1 Introduction

Consider an insurance dataset $\{\mathbf{X}_i, N_i, \{L_{i,j}\}_{j=1}^{N_i}\}_{i=1}^{n}$ in insurance ratemaking in a given year, where $\mathbf{X}_i$ is an explanatory vector representing some characteristics of the ith policyholder (e.g., age of the policyholder and age of car), N_i is the number of claims, and $\{L_{i,j}\}_{j=1}^{N_i}$ are the corresponding observed losses. Then $S_i = \sum_{j=1}^{N_i} L_{i,j}$ is the aggregate loss of the ith policyholder. A practical question in risk management is to forecast the risk of the aggregate loss S of a future policyholder with characteristic vector $\mathbf{x}$.

Two widely employed conditional risk measures of S given $\mathbf{x}$ in the financial industry and insurance business are the conditional Value-at-Risk (VaR) and conditional Expected Shortfall (ES), defined as

$$\mathrm{VaR_S}(\alpha|\mathbf{x}) = \inf\{\mathrm{s} : \mathrm{P}(\mathrm{S} \le \mathrm{s}|\mathbf{x}) \ge \alpha\} \text{ and } \mathrm{ES_S}(\alpha|\mathbf{x}) = \mathrm{E}(\mathrm{S}|\mathrm{S} > \mathrm{VaR_S}(\alpha|\mathbf{x}), \mathbf{x}), \text{ respectively.}$$

Practically, regulators often require the risk level α to be high such as 0.99 for VaR and 0.975 for ES, making nonparametric inference inefficient. On the other hand, a parametric

inference may lead to an unstable risk forecast due to the higher risk level and the fact that the parametric inference mainly employs the information around the distribution center.

To better appreciate the proposed study, we describe the recent two-step inference procedure in [9] for predicting the Value-at-Risk of the aggregate loss. The first step uses the logistic regression model to estimate the probability of having no claim, i.e., $p_i = P(N_i = 0|X_i)$. When the semicontinuous property of S_i admits the following decomposition

$$P(S_i \leq s|\mathbf{X}_i) = P(N_i = 0|\mathbf{X}_i) + P(N_i > 0|\mathbf{X}_i)P(S_i \leq s|\mathbf{X}_i, N_i > 0) \text{ for } s > s_0, \quad (1)$$

the Value-at-Risk of S_i at level α is equal to that of the conditional loss S_i given $N_i > 0$ at the adjusted level $\tilde{\alpha}_i = \frac{\alpha - p_i}{1 - p_i}$ as long as the VaR is above s_0. Therefore, the second step in [9] applies quantile regression to those $(S_i, \mathbf{X}_i)$ with positive N_i at the adjusted risk level $\tilde{\alpha}_i$, which depends on the estimator for p_i in the first step. To quantify the uncertainty of this two-step risk forecast, [13] develops a random weighted bootstrap method when the second step uses empirical quantile estimation rather than quantile regression. [12] develops another two-step procedure using weighted quantile regression. A drawback of quantile regression is the infeasible application to other risk measures such as Expected Shortfall or Expectile.

For applications to more general risk measures at a high risk level, this paper proposes to model the conditional excess function of S_i over a dynamic threshold $u_i = u(\mathbf{X}_i)$ given $\mathbf{X}_i$ and $N_i > 0$ parametrically. That is, we need a model for the dynamic threshold and a parametric model for the excess function of S_i (i.e., a semiparametric model for the distribution of S_i). When $F_{i,\mathbf{x}}(s) = P(S_i \leq s|\mathbf{X}_i = \mathbf{x}, N_i > 0)$ is in the domain of attraction of extreme value distribution with index $\xi_{i,\mathbf{x}}$ and right endpoint z_0 (see [18] and [6] for an overview about Extreme Value Theory), there exists a function $\sigma_{i,\mathbf{x}}(u) > 0$ such that

$$\lim_{u \to z_0} \sup_{0 \leq z < z_0 - u} |F_{i,\mathbf{x},u}(z) - G_{\xi_{i,\mathbf{x}}, \sigma_{i,\mathbf{x}}(u)}(z)| = 0, \quad (2)$$

where

$$F_{i,\mathbf{x},u}(z) = 1 - \frac{1 - F_{i,\mathbf{x}}(u + z)}{1 - F_{i,\mathbf{x}}(u)}$$

is the excess function, and

$$G_{\xi,\sigma}(z) = 1 - (1 + \xi z/\sigma)^{-1/\xi} \quad (3)$$

for $1 + \xi z/\sigma > 0$ with $\sigma > 0$ and $\xi \in \mathrm{R}$ is called the generalized Pareto distribution (GPD); see [1]. Therefore, we propose to model the conditional excess function of S_i given $\mathbf{X}_i$ by a GPD with parameters ξ and σ in (3) depending on the covariate vector $\mathbf{X}_i$. Because we do not model losses below the threshold, the resulted risk forecast is robust against high risk levels.

It is not new to model the parameters in the GPD as parametric functions of some covariates; see [2], [14], [3], and [17] for financial returns and [8] for climate data. However, the threshold in these papers is independent of the covariates. Because of using a dynamic GPD, we employ a quantile regression model to estimate the dynamic threshold and study the following three-step procedure for forecasting risk at a high risk level: i) using logistic regression to model the probability $p_i = P(N_i = 0|\mathbf{X}_i)$ at the first stage, ii) using quantile regression to model the dynamic threshold u_i at the 90% or 95% level at the second stage as a rule of thumb in fitting a generalized Pareto distribution (see [10]), iii) and fitting a dynamic generalized Pareto distribution to exceedances over the selected dynamic threshold u_i based on those S_i's and $\mathbf{X}_i$'s with $N_i > 0$. Combining these three steps leads to a robust forecast for a given risk measure at a high risk level because we model the distribution of S_i semiparametrically, i.e., we model losses over the threshold parametrically and below the threshold nonparametrically. In contrast, the two-step inference in [9] only works for Value-at-Risk. To quantify the uncertainty of the derived risk forecast, we adopt the random weighted bootstrap method in Jin, Ying and Wei (2001) and Zhu (2016).

References

[1] Balkema, A. A., and De Haan, L. (1974). Residual lifetime at great age. *Annals of Probability, 2(5), 792-804.*

[2] Chavez-Demoulin, V., Davison, A. C., and McNeil, A. J. (2005). A point process approach to value-at-risk estimation. *Quantitative Finance, 5(2), 227-234.*

[3] Chavez-Demoulin, V., Embrechts, P., and Sardy, S. (2014). Extreme-quantile tracking for financial time series. *Journal of Econometrics, 181(1), 44-52.*

[4] Chiang, C. T., James, L. F., and Wang, M. C. (2005). Random weighted bootstrap method for recurrent events with informative censoring. *Lifetime Data Analysis, 11(4), 489-509.*

[5] De Jong, P., and Heller, G. Z. (2008). Generalized Linear Models for Insurance Data. *Cambridge University Press.*

[6] Embrechts, P., Klüppelberg, C., and Mikosch, T. (1999). Modelling extremal events. *British Actuarial Journal, 5(2), 465-465.*

[7] Fan, J. and Gijbels, I. (1996). Local Polynomial Modelling and its Applications. *Chapman and Hall, London.*

[8] Hall, P., and Tajvidi, N. (2000). Nonparametric analysis of temporal trend when fitting parametric models to extreme-value data. *Statistical Science, 15(2), 153-167.*

[9] Heras, A., Moreno, I., and Vilar-Zanón, J. L. (2018). An application of two-stage quantile regression to insurance ratemaking. *Scandinavian Actuarial Journal, 2018(9), 753-769.*

[10] Hull, J. C. (2018), Risk Management and Financial Institutions. *Hoboken: Wiley.*

[11] Jin, Z., Ying, Z., and Wei, L. J. (2001). A simple resampling method by perturbing the minimand. *Biometrika, 88(2), 381-390.*

[12] Kang, S., Peng, L., and Golub, A. (2021). Two-step risk analysis in insurance ratemaking. *Scandinavian Actuarial Journal, 2021:6, 532-542.*

[13] Kang, S., Peng, L., and Xiao, H. (2020). Risk analysis with categorical explanatory variables. *Insurance: Mathematics and Economics, 91, 238-243.*

[14] Kelly, B., and Jiang, H. (2014). Tail risk and asset prices. *The Review of Financial Studies, 27(10), 2841-2871.*

[15] Koenker, R. (2005). Quantile Regression (Econometric Society Monographs). *Cambridge University Press.*

[16] Kudryavtsev, A. A. (2009). Using quantile regression for rate-making. *Insurance: Mathematics and Economics, 45(2), 296-304.*

[17] Massacci, D. (2017). Tail risk dynamics in stock returns: Links to the macroeconomy and global markets connectedness. *Management Science, 63(9), 3072-3089.*

[18] Resnick, S. I. (2013). Extreme Values, Regular Variation and Point Processes. *Springer.*

[19] Van der Vaart, A. W. (2000). Asymptotic Statistics (Vol. 3). *Cambridge University Press.*

[20] Wolny-Dominiak, A., and Trzesiok, M. (2014). InsuranceData: A Collection of Insurance Datasets Useful in Risk Classification in Nonlife Insurance. *R package version, 1.*

[21] Zhu, K. (2016). Bootstrapping the portmanteau tests in weak auto-regressive moving average models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 78, 463-485.*

Optimal pooling and distributed inference for the tail index and extreme quantiles

Gilles Stupfler [1] & Abdelaati Daouia [2] & Simone A. Padoan [3]

[1] *ENSAI & CREST, Campus de Ker Lann, 51 Rue Blaise Pascal, 35172 Bruz Cedex, France,* gilles.stupfler@ensai.fr

[2] *Université Toulouse 1 Capitole, TSE, 21 allée de Brienne, Manufacture des Tabacs, 31000 Toulouse, France,* abdelaati.daouia@tse-fr.eu

[3] *Department of Decision Sciences, Bocconi University of Milan, via Roentgen 1, 20136 Milano, Italy,* simone.padoan@unibocconi.it

Résumé. Ce travail, motivé par des problématiques d'estimation de risque provenant de plusieurs sources distinctes, se concentre sur des méthodes de pooling (ou groupement) pour l'estimation de paramètres extrêmes d'une loi à queue lourde. On construit et étudie une classe d'estimateurs de Hill groupés de l'indice de valeurs extrêmes, et une classe d'estimateurs de Weissman groupés calculés par moyenne géométrique. On montre la convergence de ces estimateurs lorsque les distributions au sein des échantillons peuvent présenter de l'hétérogénéité, et en présence de dépendance entre les échantillons. On obtient des estimateurs optimaux du point de vue de la variance ou de l'erreur quadratique, dont on étudie les propriétés asymptotiques en comparaison à celles de l'estimateur de Hill sur l'échantillon fusionné dans un cadre d'inférence distribuée. On considère diverses extensions, comme le cas où le nombre d'échantillons augmente, ou la présence de covariables et/ou de dépendance temporelle. Deux applications sur des données de pluie et d'assurance illustrent le comportement des estimateurs en pratique.

Mots-clés. Valeurs extrêmes, queue lourde, inférence distribuée, estimation groupée.

Abstract. Motivated by applications to the estimation of risk arising from several sources, we investigate pooling strategies for tail index and extreme quantile estimation from heavy-tailed data. To fully exploit the information contained in several samples, we present general weighted pooled Hill estimators of the tail index and weighted pooled Weissman estimators of extreme quantiles calculated through a nonstandard geometric averaging scheme. We develop their large-sample asymptotic theory across a fixed number of samples, covering the general framework of heterogeneous sample sizes with different and asymptotically dependent distributions. Our results include optimal choices of pooling weights based on asymptotic variance and MSE minimization and we assess the asymptotic properties of the corresponding estimators in comparison with the unfeasible Hill estimator on pooled data in the distributed inference setup. We consider additional scenarios where the number of subsamples grows with the total sample size and we extend our methodology to handle serial dependence and the presence of covariates. Two applications to real weather and insurance data are presented.

Keywords. Extreme values, heavy tails, distributed inference, pooling.

1 Introduction and model

The contemporary problem of efficient analysis of massive data has led to the development of the divide-and-conquer approach, or distributed inference in statistics, allowing to alleviate the computational challenges and constraints in storage imposed by the availability of extremely large datasets, or to handle data privacy issues as in banking and insurance. In this context, tail index estimation has been recently explored in Chen *et al.* (2022) by taking a simple average of subsample tail index estimators as the final distributed estimator. Here we go further by addressing several important questions about and beyond distributed inference for the tail index, but also for extreme quantiles of heavy-tailed data. Our approach is based on the use of a general terminology and theory of weighted pooling that encompasses distributed estimation from different samples whose distributions have a common target parameter, given in our setup by the tail index or an extreme quantile.

The model is the following. Let $\boldsymbol{X} = (X_1, \ldots, X_m)^\top$ denote an $m-$dimensional random vector, and $\boldsymbol{X}_i = (X_{i,1}, \ldots, X_{i,m})^\top$ $(i \geq 1)$ denote independent copies of $\boldsymbol{X}$. We assume that the available data consists of the $X_{i,j}$, for $1 \leq j \leq m$ and $1 \leq i \leq n_j = n_j(n)$, with $n = \sum_{j=1}^m n_j$ being the total number of univariate data points available across all samples. We suppose that $n_j \to \infty$ as $n \to \infty$. We focus on the general framework where the components X_j of the random vector $\boldsymbol{X}$ have continuous, right heavy-tailed distribution functions F_j, with associated survival functions $\overline{F}_j = 1 - F_j$ and tail quantile functions $U_j : t \mapsto \inf\{x \in \mathbb{R} \,|\, 1/\overline{F}_j(x) \geq t\}$ that satisfy

$\mathcal{C}_2(\boldsymbol{\gamma}, \boldsymbol{\rho}, \boldsymbol{A})$ For any $j \in \{1, \ldots, m\}$, the function U_j satisfies the second-order condition:

$\mathcal{C}_2(\gamma_j, \rho_j, A_j)$ U_j is second-order regularly varying in a neighborhood of $+\infty$ with index $\gamma_j > 0$, second-order parameter $\rho_j \leq 0$ and an auxiliary function A_j having constant sign and converging to 0 at infinity, that is,

$$\forall x > 0, \ \lim_{t\to\infty} \frac{1}{A_j(t)} \left[\frac{U_j(tx)}{U_j(t)} - x^{\gamma_j} \right] = x^{\gamma_j} \frac{x^{\rho_j} - 1}{\rho_j}$$

where the right-hand side should be read as $x^{\gamma_j} \log x$ when $\rho_j = 0$.

To incorporate the dependence between samples into the inference procedure, we assume an appropriate pairwise tail dependence structure based on the functions $\overline{C}_{j,\ell}(u, v) = \mathbb{P}(\overline{F}_j(X_j) \leq u, \ \overline{F}_\ell(X_\ell) \leq v)$ $(u, v \in [0, 1])$ that are essentially the bivariate survival copulae of $\boldsymbol{X}$, namely:

$\mathcal{J}(\boldsymbol{R})$ For any (j, ℓ) with $j \neq \ell$, there is a function $R_{j,\ell}$ on $[0, \infty]^2 \setminus \{(\infty, \infty)\}$ such that

$$\forall (x_j, x_\ell) \in [0, \infty]^2 \setminus \{(\infty, \infty)\}, \ \lim_{s\to\infty} s\, \overline{C}_{j,\ell}(x_j/s, x_\ell/s) = R_{j,\ell}(x_j, x_\ell).$$

This condition imposes the existence of a limiting dependence structure in the joint right tail of X_j and X_ℓ, given by the *tail copula* $R_{j,\ell}$ (see Schmidt and Stadtmüller, 2006). It can be viewed as a minimal assumption when it comes to assessing the dependence structure between extreme value estimators.

2 Pooling extreme value estimators

After ordering the data in the jth sample as $X_{1:n_j,j} \le X_{2:n_j,j} \le \cdots \le X_{n_j:n_j,j}$, we introduce the marginal Hill estimators $\widehat{\gamma}_j(k_j) = k_j^{-1}\sum_{i=1}^{k_j}\log(X_{n_j-i+1:n_j,j}/X_{n_j-k_j:n_j,j})$ for $k_j = k_j(n) \ge 1$. The integer k_j is the *effective sample size* in sample j, and we set $k = \sum_{j=1}^m k_j$ to be the *total effective sample size* in the vector of estimators $\widehat{\boldsymbol{\gamma}}_n = \widehat{\boldsymbol{\gamma}}_n(k_1,\ldots,k_m) = (\widehat{\gamma}_1(k_1),\ldots,\widehat{\gamma}_m(k_m))^\top$. Our ultimate interest is in the case $\boldsymbol{\gamma} = \gamma\mathbf{1}$ where the γ_j are equal to a common γ estimated by the weighted pooling estimator $\widehat{\gamma}_n(\boldsymbol{\omega}) = \widehat{\gamma}_n(\omega_1,\ldots,\omega_m) = \sum_{j=1}^m \omega_j\widehat{\gamma}_j(k_j) = \boldsymbol{\omega}^\top\widehat{\boldsymbol{\gamma}}_n$, with $\boldsymbol{\omega}^\top\mathbf{1} = 1$. Our first contribution is to obtain the asymptotic distribution of any such estimator along with the joint asymptotic normality of $\widehat{\gamma}_j(k_j)$, for $j \in \{1,\ldots,m\}$, when m is fixed. We obtain formulae for the asymptotic variance-optimal and asymptotic MSE-optimal choices of $\boldsymbol{\omega}$, as well as estimators of these optimal weights and the asymptotic properties of the corresponding composite plug-in estimators.

We then turn to extreme quantile estimation for a very small exceedance probability $p = p(n) \to 0$ as $n \to \infty$. In each sample, one may estimate $q_j(1-p)$ by the extrapolated Weissman estimator of Weissman (1978), that is, $\widehat{q}_j^\star(1-p|k_j) = (k_j/(n_jp))^{\widehat{\gamma}_j(k_j)}X_{n_j-k_j:n_j,j}$, where k_j satisfies $k_j/(n_jp) \to \infty$. When the samples are believed or known to have the same tail index γ, it is natural to harness the power of pooling by substituting the weighted estimator $\widehat{\gamma}_n(\boldsymbol{\omega})$ in place of the individual estimator $\widehat{\gamma}_j(k_j)$, to get $\widehat{q}_j^\star(1-p|k_j,\boldsymbol{\omega}) = (k_j/(n_jp))^{\widehat{\gamma}_n(\boldsymbol{\omega})}X_{n_j-k_j:n_j,j}$. Going one step further, the marginal quantile estimators themselves can be pooled to gain more insight when the datasets have the same extreme quantiles, or equivalently, if one assumes that

($\mathcal{H}$) For any $j,\ell \in \{1,\ldots,m\}$ with $j \ne \ell$, we have $U_j(t)/U_\ell(t) \to 1$ as $t \to \infty$.

The validity of this assumption can be tested, as we show by constructing a likelihood ratio-type test statistic based on the joint asymptotic normality of Weissman estimators. Under assumption ($\mathcal{H}$), the quantiles $q_j(1-p)$ are all asymptotically equivalent and hence the estimators $\widehat{q}_j^\star(1-p|k_j)$ all estimate the same quantity. A straightforward way of combining these individual extreme quantile estimators would be to take again a weighted sum, as $\overline{q}_n^\star(1-p|\boldsymbol{\omega}) = \sum_{j=1}^m \omega_j\widehat{q}_j^\star(1-p|k_j)$. This is not the best possible solution when it comes to pooling the Weissman estimators, because the use of geometric weighted sums better suits the multiplicative and power structure of these extrapolated estimators. We then consider the alternative version

$$\widehat{q}_n^\star(1-p|\boldsymbol{\omega}) = \prod_{j=1}^m [\widehat{q}_j^\star(1-p|k_j)]^{\omega_j} = \prod_{j=1}^m \left[\left(\frac{k_j}{n_jp}\right)^{\widehat{\gamma}_j(k_j)} X_{n_j-k_j:n_j,j}\right]^{\omega_j}.$$

Our next contribution is to derive the asymptotic normality of both $\widehat{q}_j^\star(1-p|k_j,\boldsymbol{\omega})$ and $\widehat{q}_n^\star(1-p|\boldsymbol{\omega})$, when $\boldsymbol{\omega}$ is a fixed weight vector, or an asymptotic variance- or MSE-optimal choice.

3 The framework of distributed inference

3.1 Fixed parameter m

Our general pooling theory naturally applies to the context of distributed inference. A restriction in this framework is that the n individual data points cannot be processed by a standalone machine and very restricted or no communication is allowed between the m machines, making it impossible to estimate the tail dependence structure between the different subsamples. We thus make the assumption that the data within and across machines are independent, that is, the $X_{i,j}$ are i.i.d. for $1 \le i \le n_j$ and $1 \le j \le m$, with a common distribution satisfying the second-order condition $\mathcal{C}_2(\gamma, \rho, A)$.

A benchmark for the distributed estimator $\widehat{\gamma}_n(\boldsymbol{\omega})$ is then the Hill estimator based on the unfeasible combination of subsamples $\{X_i, 1 \le i \le n\} = \{X_{i,j}, 1 \le j \le m, 1 \le i \le n_j\}$ with effective sample size $k = \sum_{j=1}^m k_j$, that is, $\widehat{\gamma}_n^{(\mathrm{Hill})}(k) = k^{-1} \sum_{i=1}^k \log(X_{n-i+1:n}/X_{n-k:n})$ where $X_{1:n} \le X_{2:n} \le \cdots \le X_{n:n}$ are the order statistics of the random variables X_i. We derive the asymptotic normality of $\widehat{\gamma}_n(\boldsymbol{\omega})$ in this setup, and we compare its asymptotic distribution with that of $\widehat{\gamma}_n^{(\mathrm{Hill})}(k)$. In particular, we prove that adjusting the effective sample sizes k_j so that k_j/n_j is constant across machines produces a variance-optimal distributed estimator that is asymptotically equivalent to $\widehat{\gamma}_n^{(\mathrm{Hill})}(k)$. This is much stronger than only sharing the same asymptotic distribution. We also prove the remarkable property that the AMSE-optimal distributed estimator can actually have a smaller AMSE than the benchmark Hill estimator itself under the necessary and sufficient condition that the bias is sufficiently large. Similar properties hold for the corresponding geometrically weighted pooled extreme quantile estimators.

3.2 Growing parameter m

When all the $k_j = k_j(n)$ are bounded in n, consistency of the distributed estimators necessarily requires a growing number of machines with n, namely $m = m(n) \to \infty$. Otherwise, when m is fixed, the pooled estimators will contain only a bounded number of summands. In this context, we require the following fundamental assumption:

($\mathcal{A}$) $m = m(n) \to \infty$ and the $n_j = n_j(n)$ satisfy $\inf_{1 \le j \le m} n_j / \log m \to \infty$ as $n \to \infty$.

While the condition $m \to \infty$ ensures consistency, the condition $\inf_{1 \le j \le m} n_j / \log m \to \infty$ is required to establish a precise control of the statistical errors arising in each machine. The proof is in this regard fundamentally different from the proofs of our theorems in the case of a fixed m. Another important difference is that the weight vector $\boldsymbol{\omega} \in \mathbb{R}^m$ now implicitly varies with n, so restrictions are needed to define a class of admissible weights. For example, the pooled estimator corresponding to the weight $(1, 0, 0, \ldots)$ is simply $\widehat{\gamma}_1(k_1)$, which is not consistent when k_1 is bounded. We thus introduce a *balanced allocation condition* on $\boldsymbol{\omega} = \boldsymbol{\omega}(n)$:

($\mathcal{W}$) The weight vector $\boldsymbol{\omega} = \boldsymbol{\omega}(n) \in \mathbb{R}^m$ satisfies $\sum_{j=1}^m \omega_j = 1$ as well as

$$\limsup_{n\to\infty} k \sum_{j=1}^m \frac{\omega_j^2}{k_j} < \infty \text{ and } \exists \delta > 0,\ \frac{\sum_{j=1}^m k_j^{-\delta/2}(\omega_j^2/k_j)^{1+\delta/2}}{(\sum_{j=1}^m \omega_j^2/k_j)^{1+\delta/2}} \to 0 \text{ as } n \to \infty.$$

Any weight vector satisfying $m(n)\sup_{1\le j\le m}|\omega_j(n)| \le \omega_0 < \infty$, for all n, will automatically fulfill ($\mathcal{W}$). This encompasses the naive pooled estimator of γ, studied in Chen *et al.* (2022).

We obtain the asymptotic properties of $\widehat{\gamma}_n(\boldsymbol{\omega})$ in a unified fashion in the sense that the k_j can all be bounded, can all diverge to infinity, or anything in-between. We also correct an unfortunate mistake in an asymptotic variance term obtained by Chen *et al.* (2022) and find that the distributed Hill estimator with $\omega_j = 1/m$ typically does not achieve the so-called oracle property, although the estimator with asymptotic variance-optimal weights does. We then turn to the asymptotic behavior of the geometrically weighted extreme quantile estimator $\widehat{q}_n^{\star}(1-p|\boldsymbol{\omega})$ and show that, perhaps surprisingly, this distributed estimator is in fact generally not consistent when the k_j are bounded, because the individual scale parameter estimators $X_{n_j-k_j:n_j,j}$ are then fundamentally biased estimators of $q(1-k_j/n_j)$. With growing k_j, however, the distributed extreme quantile estimator is, as expected, consistent and asymptotically normal.

4 Filtering to handle dependence and covariates

In practice, the data are often recorded with relevant covariates, or are stationary but weakly dependent in a way that can be modeled by a standard time series. Besides, when the X_j share the same tail index, they typically have also asymptotically proportional extreme quantiles. This suggests that $(X_1, \ldots, X_m)$ can be modeled in many situations by a general location-scale model $X_j = g_j(\boldsymbol{Z}_j) + \sigma_j(\boldsymbol{Z}_j)\varepsilon_j,\ 1 \le j \le m$ where the unobserved noise vector $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_m)$ has marginal tail quantile functions U_j satisfying the conditions $\mathcal{C}_2(\gamma\mathbf{1}, \boldsymbol{\rho}, \boldsymbol{A})$ and ($\mathcal{H}$), and its bivariate survival copulae $\overline{C}_{j,\ell}$ satisfy $\mathcal{J}(\boldsymbol{R})$. The functions g_j and $\sigma_j > 0$ are unknown measurable functions of $\boldsymbol{Z}_j \in \mathbb{R}^{l_j}$, for some $l_j \ge 1$. The covariates $\boldsymbol{Z}_j$ can be fully observed (in traditional regression settings) or partially or not at all observed (in a time series model which includes past unobserved innovations or volatility terms). The noise variable ε_j is assumed to be independent of $\boldsymbol{Z}_j$.

A reasonable idea to eliminate the heteroskedasticity and dependence in data $X_{i,j} = g_j(\boldsymbol{Z}_{i,j}) + \sigma_j(\boldsymbol{Z}_{i,j})\varepsilon_{i,j}$ (for $1 \le j \le m$ and $1 \le i \le n_j$) is to first estimate the location and scale components g_j and σ_j of the model and then filter the data to obtain residuals $\widehat{\varepsilon}_{i,j}^{(n_j)}$ close to the unobserved errors $\varepsilon_{i,j}$. This results in j residual-based Hill estimators which can be combined in a pooled version, whose asymptotic normality can be proved under a high-level condition on the discrepancy between $\widehat{\varepsilon}_{i,j}^{(n_j)}$ and $\varepsilon_{i,j}$. This condition is typically satisfied as soon as the location and scale components g_j and σ_j are estimated at

a faster rate than $\sqrt{k_j}$, and can easily be checked theoretically in a variety of regression models, see Girard *et al.* (2021) for examples. The ultimate interest in conditional extreme value modeling is of course to estimate extreme $(1-p)$th quantiles of X_j given $\boldsymbol{Z}_j = \boldsymbol{z}_j$, defined as $q_{X_j|\boldsymbol{Z}_j=\boldsymbol{z}_j}(1-p) = g_j(\boldsymbol{z}_j)+\sigma_j(\boldsymbol{z}_j)q_j(1-p)$ by location equivariance and positive homogeneity of quantiles. When $\boldsymbol{\varepsilon}$ satisfies the condition $(\mathcal{H})$ described in Section 2, all associated quantiles $q_j(1-p)$ become asymptotically equivalent and can then be estimated by a geometrically pooled estimator. Our final results establish the asymptotic normality of these pooled filtered tail index and extreme quantile estimators.

5 Data analysis

We discuss two concrete applications of our methodology to insurance and rainfall data. The first dataset comprises total claim amounts for car insurance companies in five US states between January and February 2011. The choice of this dataset follows the same setup as in Chen *et al.* (2022). Then, we explore regional inference on tail index and extreme quantiles of monthly rainfall in the state of Florida, using data on rainfall measurements collected daily by the Florida Automated Weather Network at 49 gauge stations, over different periods between December 1997 and May 2021, see `https://fawn.ifas.ufl.edu/data/fawnpub`.

Acknowledgements

This research was supported by the French National Research Agency under the grants ANR-19-CE40-0013 and ANR-17-EURE-0010. S. A. Padoan is supported by the Bocconi Institute for Data Science and Analytics. G. Stupfler acknowledges support from an AXA Research Fund Award on "Mitigating risk in the wake of the COVID-19 pandemic".

References

Chen, L., Li, D. and Zhou, C. (2022). Distributed inference for the extreme value index, *Biometrika*, 109, pp. 257-264.

Girard, S., Stupfler, G. and Usseglio-Carleve, A. (2021). Extreme conditional expectile estimation in heavy-tailed heteroscedastic regression models, *Annals of Statistics*, 49, pp. 3358-3382.

Schmidt, R. and Stadtmüller, U. (2006). Non-parametric estimation of tail dependence, *Scandinavian Journal of Statistics*, 33, pp. 307-335.

Weissman, I. (1978). Estimation of parameters and large quantiles based on the k largest observations, *Journal of the American Statistical Association*, 73, pp. 812-815.

Clustering parcimonieux pour extrêmes multivariés

Nicolas MEYER [1] & **Olivier WINTENBERGER** [2]

[1] *Université de Montpellier, IMAG, France*
nicolas.meyer@umontpellier.fr
[2] *Sorbonne Université, LPSM, France*
olivier.wintenberger@sorbonne-universite.fr

Résumé. Identifier les directions dans lesquelles des événements exceptionnels apparaissent est un des problèmes majeurs de la théorie multivariée des valeurs extrêmes. D'un point de vue théorique, l'information concernant de tels événements est contenue dans la mesure spectrale, qui apparaît comme la limite de la composante angulaire de vecteurs aléatoires à variation régulière. Estimer cette mesure s'avère être un point délicat, notamment en grande dimension. Dans cette présentation, nous introduisons une méthode de réduction de la dimension fondée sur la projection euclidienne sur le simplexe. Cette projection a été étudiée dans le cadre des valeurs extrêmes par Meyer & Wintenberger (2021) qui ont établi plusieurs résultats théoriques. La présentation s'attachera à exposer une approche statistique fondée sur de la sélection de modèle qui permet d'identifier les groupes de coordonnées susceptibles d'être extrêmes simultanément. Cette approche donne lieu à un algorithme appelé *MUSCLE* pour *MUltivariate Sparse Clustering for Extremes*.

Mots-clés. Extrêmes multivariés, measure spectrale, projection sur le simplexe.

Abstract. Identifying directions where exceptional events occur is one of the major problems of multivariate extreme value theory. From a theoretical point of view most of the information concerning such events is contained in the spectral measure which appears as the limit of the angular component of regularly varying random vectors. Estimating this measure is a delicate point especially in large dimensions. In this presentation we introduce a dimension reduction method based on the Euclidean projection onto the simplex. This projection has been studied in the context of extreme values by Meyer & Wintenberger (2021) who established several theoretical results. The presentation will focus on a statistical approach that uses model selection to identify groups of coordinates that are likely to be extreme simultaneously. This approach gives rise to an algorithme called *MUSCLE* for *MUltivariate Sparse Clustering for Extremes*.

Keywords. Multivariate extremes, projection onto the simplex, spectral measure.

1 Valeurs extrêmes et variation régulière

1.1 Variation régulière

Étudier les valeurs extrêmes générées par un vecteur aléatoire $\mathbf{X} \in \mathbb{R}_+^d$, $d \geq 2$, revient à étudier le comportement de la queue de distribution de $\mathbf{X}$. Dans ce contexte, il est courant de supposer que le vecteur $\mathbf{X}$ est à variation régulière : il existe un vecteur aléatoire $\mathbf{\Theta}$ sur la sphère unité telle que

$$\mathbb{P}\big((|\mathbf{X}|/t, \mathbf{X}/|\mathbf{X}|) \in \cdot \mid |\mathbf{X}| > t\big) \xrightarrow{d} \mathbb{P}((Y, \mathbf{\Theta}) \in \cdot)\,, \quad t \to \infty\,, \tag{1}$$

cf Resnick (2007). Dans ce cas, $|\cdot|$ désigne n'importe quelle norme sur $\mathbb{R}^d$. Le vecteur limite $\mathbf{\Theta}$ est alors appelé vecteur spectral tandis que sa loi est appelée mesure spectrale. La convergence (**??**) permet de séparer l'étude de la composante radiale des extrêmes $|\mathbf{X}|/t$ de celle de la composante angulaire $\mathbf{X}/|\mathbf{X}|$. Cette dernière concentre l'information sur la localisation et la dépendance des valeurs extrêmes. L'étude de la mesure spectrale est donc un point central de la théorie des extrêmes multivariés.

Il est intéressant (voir par exemple Goix et al. (2017)) d'étudier le comportement du vecteur spectral sur les sous-ensembles C_β de la sphère définis par

$$C_\beta = \{\mathbf{x} \in \mathbb{R}_+^d : |\mathbf{x}| = 1,\, x_j > 0 \text{ pour } j \in \beta,\, x_j = 0 \text{ pour } j \notin \beta\}\,,$$

pour $\beta \subset \{1, \ldots, d\}$. Des groupes de directions β sont appelés *clusters*. En effet, la mesure spectrale met de la masse sur un tel ensemble si des événements extrêmes apparaissent conjointement dans la direction β. On est ainsi ramené à l'estimation des probabilités $\mathbb{P}(\mathbf{\Theta} \in C_\beta)$. Cependant, l'estimation de ces quantités se révèle délicate pour essentiellement deux raisons. Tout d'abord, le nombre de probabilités à estimer croît exponentiellement en la dimension. Par ailleurs, si $\beta \neq \{1, \ldots, d\}$ vérifie $\mathbb{P}(\mathbf{\Theta} \in C_\beta) > 0$, alors la mesure spectrale charge la frontière de C_β (qui est le sous-ensemble C_β lui-même) et donc la convergence (**??**) ne s'applique pas.

L'idée proposée par Meyer et Wintenberger (2021) pour contourner ce problème est de remplacer le vecteur unitaire $\mathbf{X}/|\mathbf{X}|$ de (**??**) par un autre projeté qui permet de mieux tenir compte de la masse mise par la mesure spectrale sur les sous-ensembles C_β. Cette modification de la convergence (**??**) donne alors naissance à la notion de variation régulière parcimonieuse.

1.2 Variation régulière parcimonieuse

Popularisée par Duchi et al. (2008), la projection euclidienne sur le simplexe a connu un usage divers et varié, notamment en théorie de l'apprentissage.

Dans la suite, $|\cdot|$ désigne la norme ℓ^1 et $\mathbb{S}_+^{d-1}$ désigne le simplexe de $\mathbb{R}^d$. Si $\mathbf{v} \in \mathbb{R}_+^d$, alors le vecteur projeté $\pi(\mathbf{v})$ est l'unique vecteur $\mathbf{w}$ de $\mathbb{S}_+^{d-1}$ qui minimise la quantité

$|\mathbf{w}-\mathbf{v}|_2$, où $|\cdot|_2$ désigne la norme ℓ^2. On note alors π la projection euclidienne sur le simplexe $\mathbb{S}_+^{d-1}$. Cette manière de projeter permet de rendre les vecteurs parcimonieux, c'est-à-dire avec plusieurs coordonnées nulles. Elle permet ainsi de mieux rendre compte du comportement des extrêmes sur les ensembles C_β.

Définition 1 (Variation régulière parcimonieuse)**.** *Un vecteur* $\mathbf{X}$ *à valeurs dans* $\mathbb{R}_+^d$ *est dit à variation régulière parcimonieuse s'il existe un vecteur aléatoire* $\mathbf{Z}$ *défini sur le simplexe et une variable aléatoire positive* Y *tels que*

$$\mathbb{P}\big((|\mathbf{X}|/t, \pi(\mathbf{X}/t)) \in \cdot \mid |\mathbf{X}| > t\big) \xrightarrow{d} \mathbb{P}((Y, \mathbf{Z}) \in \cdot)\,, \quad t \to \infty\,. \tag{2}$$

Le vecteur limite $\mathbf{Z}$ doit être vu comme la limite angulaire obtenue après avoir remplacé $\mathbf{X}/|\mathbf{X}|$ par $\pi(\mathbf{X}/t)$ dans l'Equation (**??**). Par continuité de la projection, la notion de variation régulière standard (Equation (**??**)) implique celle de variation régulière parcimonieuse. Meyer et Wintenberger (2021) ont prouvé que sous des hypothèses assez faibles, les deux notions sont en fait équivalentes.

L'intérêt principal de la Définition **??** est de pouvoir approcher le comportement des extrêmes de $\mathbf{X}$ sur les ensembles C_β. En effet, en reprenant les notations précédentes, on a la convergence suivante pour tout $\beta \subset \{1, \ldots, d\}$:

$$\mathbb{P}(\pi(\mathbf{X}/t) \in C_\beta \mid |\mathbf{X}| > t) \to \mathbb{P}(\mathbf{Z} \in C_\beta)\,, \quad t \to \infty\,. \tag{3}$$

L'objectif est donc d'estimer le support de la distribution de $\mathbf{Z}$ via l'estimation des probabilités $\mathbb{P}(\mathbf{Z} \in C_\beta)$, pour $\beta \subset \{1, \ldots, d\}$, le but étant de détecter lesquelles de ces probabilités sont positives. Autrement dit, il s'agit d'identifier l'ensemble

$$\mathcal{S}^*(\mathbf{Z}) := \{\beta \subset \{1, \ldots, d\} : \mathbb{P}(\mathbf{Z} \in C_\beta) > 0\}\,.$$

Cet ensemble $\mathcal{S}^*(\mathbf{Z})$ rassemble tous les clusters de directions β sur lesquelles le vecteur angulaire $\mathbf{Z}$ met de la masse. On note s^* son cardinal. L'objectif est alors de proposer une approche statistique pour décider quels clusters β appartiennent à $\mathcal{S}^*(\mathbf{Z})$.

2 Estimation

On considère désormais une suite de vecteurs aléatoires indépendants et identiquement distribués à variation régulière $\mathbf{X}, \mathbf{X}_1, \mathbf{X}_2, \ldots$ de vecteur spectral $\mathbf{\Theta}$. On considère également une variable aléatoire Y de loi de Pareto de paramètre $\alpha > 0$, indépendante de $\mathbf{\Theta}$. Enfin, on pose $\mathbf{Z} = \pi(Y\mathbf{\Theta})$.

Le cadre classique en statistique des valeurs extrêmes est de considérer une suite positive $(u_n)_{n\in\mathbb{N}}$ telle que $u_n \to \infty$. Cette suite joue le rôle du seuil t dans les Equations (**??**) et (**??**). Cela signifie que pour $n \in \mathbb{N}$, la quantité u_n doit être vue comme le seuil au-dessus duquel les données $\mathbf{X}_1, \cdots, \mathbf{X}_n$ sont considérées comme des valeurs extrêmes.

Il est également usuel de définir un niveau $k = k_n = n\mathbb{P}(|\mathbf{X}| > u_n)$ et de supposer que $k_n \to \infty$ quand $n \to \infty$. Il est à noter que l'hypothèse $u_n \to \infty$ implique que $k_n/n = \mathbb{P}(|\mathbf{X}| > u_n) \to 0$. Ainsi, k_n tend vers l'infini à une vitesse plus lente que n. En pratique k_n est choisi tel que $k_n = \sum_{j=1}^n \mathbb{1}_{|\mathbf{X}_j|>u_n}$ qui correspond au nombre de dépassements au-dessus du seuil u_n, c'est-à-dire au nombre de valeurs extrêmes. Dans ce cas u_n peut être restreint à prendre des valeurs parmi les statistiques d'ordre $|\mathbf{X}|_{(1)} \geq \cdots \geq |\mathbf{X}|_{(n)}$. Dans toute la suite nous considérons donc $u_n = |\mathbf{X}|_{(k_n)}$.

Notre objectif est d'estimer les probabilités $p^*(\beta) := \mathbb{P}(\mathbf{Z} \in C_\beta)$ pour $\beta \subset \{1, \ldots, d\}$. Ces probabilités apparaissent comme les limites des probabilités pré-asymptotiques $p_n(\beta) := \mathbb{P}(\pi(\mathbf{X}/u_n) \in C_\beta \mid |\mathbf{X}| > u_n)$ (voir Equation (**??**)). Le problème principal est alors de décider si $p(\beta)$ est positif ou nul. On définit pour cela l'estimateur

$$T_n(\beta) := \sum_{j=1}^n \mathbb{1}\{\pi(\mathbf{X}_j/|\mathbf{X}|_{(k_n)}) \in C_\beta,\, |\mathbf{X}_j| > |\mathbf{X}|_{(k_n)}\}\,,$$

pour $\beta \subset \{1, \ldots, d\}$. L'idée est de sélectionner parmi les valeurs extrêmes celles qui sont projetées dans l'ensemble C_β. Ces estimateurs vérifient les résultats asymptotiques suivants.

Théorème 1. *On reprend les notations précédentes.*

1. *(Consistance). Le vecteur $k_n^{-1}(T_n(\beta))_{\beta\subset\{1,\ldots,d\}}$ converge en probabilité vers $\mathbf{p}^* := (p(\beta))_{\beta\subset\{1,\ldots,d\}}$.*
2. *(Normalité asymptotique). On a la convergence en loi suivante :*

$$\sqrt{k_n}\,\mathrm{Diag}\left(\mathbf{p}_{\mathcal{S}^*(\mathbf{Z})}\right)^{-1/2}\Big(\frac{\mathbf{T}_{n,\mathcal{S}^*(\mathbf{Z})}}{k_n} - \mathbf{p}_{n,\mathcal{S}^*(\mathbf{Z})}\Big) \xrightarrow{d} \mathcal{N}(0, Id_{s^*} - \sqrt{\mathbf{p}^*_{\mathcal{S}^*(\mathbf{Z})}}\sqrt{\mathbf{p}^*_{\mathcal{S}^*(\mathbf{Z})}}^T)\,, \quad n \to \infty\,.$$

où $\mathbf{T}_{n,\mathcal{S}^*(\mathbf{Z})}$ (resp. $\mathbf{p}_{n,\mathcal{S}^*(\mathbf{Z})}$ et $\mathbf{p}^*_{\mathcal{S}^*(\mathbf{Z})}$) correspond au vecteur de $\mathbb{R}^{s^*}$ dont les composantes sont les $T_n(\beta)$ (resp. $p_n(\beta)$ et $p^*(\beta)$) pour $\beta \in \mathcal{S}^*(\mathbf{Z})$ (on se restreint aux probabilités positives).

3 Sélection de modèle

La répartition des k données extrêmes sur les $2^d - 1$ sous-ensembles $(C_\beta)_{\beta\subset\{1,\ldots,d\}}$ suggère d'utiliser le modèle multinomial $\mathbf{M}_k$ sur R^{2^d-1} de vecteur de probabilités $\mathbf{p}$ défini par

$$\mathbf{p} = (\overbrace{p_1, \ldots, p_s, \underbrace{p, \ldots, p}_{r-s}, 0, \ldots, 0}^{2^d-1 \text{ composantes}})\,,$$

avec $p_1 \geq \ldots \geq p_s, p \in (0,1)$ satisfaisant la contrainte :

$$p_1 + \ldots + p_s + (r-s)\tilde{p} = 1\,.$$

L'idée est de séparer les clusters β en trois catégories. La première correspond aux clusters sur lesquels les extrêmes apparaissent, leur probabilité d'apparition étant alors p_j. La deuxième catégorie correspond aux clusters qui concentrent peu de données extrêmes. Ce phénomène résulte du biais entre l'aspect non-asymptotique de l'étude et le modèle théorique des $p^*(\beta)$. Ce biais est modélisé par une probabilité d'occurence p considérée comme proche de 0. Enfin, la dernière catégorie concerne les clusters β sur lesquels aucune donnée n'est apparue ; on estime alors que ces clusters ne concentrent pas de valeurs extrêmes (d'où une probabilité d'occurrence nulle). L'objectif est alors d'ajuster au mieux le nombre s de faces significatives. Cet ajustement s'effectue par le calcul de la divergence de Kullback-Leibler entre les données et le modèle théorique $\mathbf{M}_k$. L'estimation de cette vraisemblance implique que le modèle qui correspond le mieux aux données est celui qui minimise une log-vraisemblance pénalisée.

La sélection des clusters contenant les extrêmes de $\mathbf{X}$ s'est faite jusqu'à présent pour le choix d'un nombre de statistiques d'ordre k_n arbitraire. L'idée est alors d'inclure le choix de ce nombre dans la sélection de modèle. L'approche utilisée consiste à considérer des modèles avec un nombre de valeurs extrêmes différent et de déterminer le modèle le plus approprié. Cette analyse est réalisée en partitionnant les données en un groupe d'extrêmes et un groupe de non-extrêmes, le but de la sélection de modèle étant d'identifier la partition qui correspond le mieux aux données.

Notre approche donne lieu à un algorithme appelé MUSCLE pour *MUltivariate Sparse Clustering for Extremes* qui donne les clusters extrêmes à partir de données $\mathbf{X}_1, \ldots, \mathbf{X}_n$. Cet algorithme ne nécessite aucun hyper-paramètre, contrairement aux méthodes existantes dans la littérature. Ce travail est par ailleurs le premier qui combine l'étude de la dépendance des extrêmes et le choix du seuil.

La fin de notre présentation est consacrée à l'étude de divers exemples sur des données simulées qui illustrent la pertinence de notre approche. Enfin, nous mettons en évidence les clusters extrêmes sur des données financières et environnementales.

Bibliographie

Duchi, J. Shalev-Shwartz, S. Singer, Y. Chandra, T. (2008), *Efficient Projections onto the ℓ_1-Ball for Learning in High Dimensions*, ICML.
Goix, N. Sabourin, A. Clémençon, S. (2017), *Sparsity in Multivariate Extremes with Applications to Anomaly Detection*, Journal of Multivariate Analysis.
Meyer, N. et Wintenberger, O. (2021), Sparse regular variation, à paraître dans *Advances in Applied Probability*, arXiv : 1907.00686.

Resnick, S.I. (2007), *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling*, Springer.

Applications

Votes par évaluation avec des fonctions de profondeur

Jean-Baptiste Aubin[1], Irène Gannaz[1], Samuela Leoni[1], Antoine Rolland[2]

[1] *Univ Lyon, INSA de Lyon, CNRS UMR 5208, Institut Camille Jordan, F-69621 Villeurbanne, France*
[2] *Université Lyon 2, Laboratoire ERIC*

Résumé. Cet article propose un cadre unifié pour les procédures de vote basées sur une évaluation des candidats. L'idée est de représenter les notes de chaque électeur sur d candidats comme un point dans $\mathbb{R}^d$ et de définir le vainqueur du vote en utilisant le point le plus profond du nuage de points ainsi formé. Le point le plus profond est obtenu par la maximisation d'une fonction de profondeur. Il est démontré que les principales procédures par évaluation (jugement majoritaire, vote par approbation...) rentrent dans ce cadre. De plus, les propriétés des procédures obtenues étudiées sont l'universalité, l'unanimité, la neutralité, la monotonie et l'indépendance aux alternatives non pertinentes (IIA).

Mots-clés. Procédure de vote ; vote par évaluation ; fonctions de profondeur

Abstract. This article aims to present a unified framework for grading-based voting processes. The idea is to represent the grades of each voter on d candidates as a point in $\mathbb{R}^d$ and to define the winner of the vote using the deepest point of the scatter plot. The deepest point is obtained by the maximization of a depth function. The studied properties of the resulting processes are universality, unanimity, neutrality, monotonicity and IIA properties.

Keywords. Voting process ; Grade modeling ; Depth functions

Balinski et Laraki [1] ont montré que les procédures de vote basées sur des évaluations plutôt que sur des classements, satisfont des propriétés intéressantes. Ils proposent une règle de vote, le jugement majoritaire, basée sur les médianes des évaluations, et prouvent qu'elle n'entre pas dans le cadre du théorème d'impossibilité d'Arrow [1]. Cette propriété ressort des informations contenues dans chaque vote qui fournissent des informations plus nuancées qu'un vote basé sur un simple classement des candidats par les votants (comme dans le cas des scrutins majoritaires, de Borda, etc.). Le modèle par évaluation a suscité beaucoup d'intérêt au cours des dernières décennies, et les alternatives au jugement majoritaire sont, par exemple, le vote par approbation [2] et le *range voting* [3].

Cet article présente un cadre unifié pour les processus de vote basés sur les évaluations, reposant sur les notions statistiques des fonctions de profondeur et des points les plus profonds associés. Considérons dans la suite que nous avons n électeurs et d candidats. Chaque électeur donne une note à chaque candidat. Chaque électeur peut alors être

assimilé à un point de $\mathbb{R}^d$, dont les coordonnées sont les notes données à chacun des candidats. L'idée clé est de considérer l'électeur *le plus central* du nuage de points obtenu. Le gagnant du vote est alors le candidat qui a la note maximale de cet électeur central.

La définition du point le plus central d'un nuage de points est un sujet de recherche bien connu en statistique. L'idée est d'introduire une fonction de profondeur, qui associe à chaque point une valeur, de sorte que la fonction de profondeur soit maximale au point le plus central du nuage de points et minimale pour les valeurs aberrantes. De nombreuses notions de profondeur de données ont été introduites [4, 5]. Les fonctions de profondeur peuvent apporter un point de vue intéressant sur les procédures de vote.

1 La procédure de vote

Nous considérons ici le cadre des votes par évaluation défini par [1]. Considérons le cas de n électeurs et d candidats. Chaque électeur v_j, $j = 1, \ldots, n$ donne une note au candidat c_i, $i = 1, \ldots, d$. Notons $\Phi(i, j)$ cette note. Soit Λ l'ensemble des valeurs possibles des notes $\Phi(i, j)$. Comme le font remarquer Balinski et Laraki [1], l'ensemble des notes Λ doit être strictement ordonné mais il peut être fini ou être un intervalle réel. Supposons sans perte de généralité que $\Lambda \subseteq [0, 1]$. Soit $\Phi = \{\Phi(i, j), i = 1, \ldots, d, j = 1, \ldots, n\} \in \Lambda^{d \times n}$ la matrice des notes. Une collection de d notes s'appelle un *profil*. Chaque colonne $\Phi(., j)$ est un profil.

Une procédure de vote est une fonction G qui à une matrice Φ associe un unique profil. La fonction G est ainsi définie sur $\Lambda^{d \times n}$ et à valeurs dans $[0, 1]^d$. Balinski et Laraki [1] proposent des conditions, ou axiomes, qu'une telle procédure doit vérifier :

Neutralité : G est invariante par permutation sur les lignes de Φ (*i.e.* si on permute les candidats).

Universalité : G est invariante par permutation sur les colonnes de Φ (*i.e.* si on permute les électeurs).

Unanimité : si tous les électeurs donnent une même note α à un candidat, alors G lui assigne la valeur α.

Monotonie : si $\Phi = \widetilde{\Phi}$ à l'exception d'un ou de plusieurs électeurs qui donnent une note plus élevée au candidat c_i dans Φ que dans $\widetilde{\Phi}$, alors $G(\Phi)(c_i) \geq G(\widetilde{\Phi})(c_i)$.

IIA (Indépendance aux alternatives non pertinentes) : si les notes obtenues par un candidat c_i dans deux votes Φ et $\widetilde{\Phi}$ sont identiques, alors $G(\Phi)(c_i) = G(\widetilde{\Phi})(c_i)$.

2 Les fonctions de profondeur

Étant donnée une distribution F sur $\mathbb{R}^d$, une fonction de profondeur mesure la centralité des points de $\mathbb{R}^d$ relativement à F. Une valeur élevée de la fonction de profondeur correspond à un point *central* tandis qu'une valeur faible correspond à une valeur *en périphérie*. Appliquées aux votes, ces fonctions permettent ainsi de déterminer quel serait l'électeur *central*, déterminant l'issue du vote.

Notons $\mathcal{F}$ l'ensemble des lois de probabilités sur les boréliens de $\mathbb{R}^d$ et notons F_X la loi d'un vecteur aléatoire X.

Definition 1. *Soit $D : \mathbb{R}^d \times \mathcal{F} \to \mathbb{R}$ une fonction bornée, non négative, telle que:*

(P1) Soit $X = (X_1, \dots, X_d)$ un vecteur aléatoire dans $\mathbb{R}^d$, $x \in \mathbb{R}^d$, et σ une permutation de $\{1, \dots, d\}$. Notons $X_\sigma = (X_{\sigma(1)}, \dots, X_{\sigma(d)})$ et $x_\sigma = (x_{\sigma(1)}, \dots, x_{\sigma(d)})$. Alors $D(x_\sigma, F_{X_\sigma}) = D(x; F_X)$.

(P2) Pour tout $a \in \mathbb{R}$, $b \in \mathbb{R}^d$, pour tout vecteur aléatoire $X \in \mathbb{R}^d$, $\operatorname{argmax}_{x \in \mathbb{R}^d} D(a\,x + b; F_{a\,X+b}) = \operatorname{argmax}_{x \in \mathbb{R}^d} D(x; F_X)$.

*(P3) Pour toute distribution $F \in \mathcal{F}$ ayant un unique "centre" θ (*e.g.* un point de "symétrie"), $D(\theta; F) = \sup_{x \in \mathbb{R}^d} D(x; F)$.*

(P4) Pour tout $F \in \mathcal{F}$, $D(\cdot; F)$ est quasi-concave. C'est-à-dire, si $\theta \in \operatorname{argmax}_{x \in \mathbb{R}^d} D(x; F)$, alors $D(x; F) \leq D(\theta + \lambda(x - \theta); F)$ pour tout $0 \leq \lambda \leq 1$.

(P5) $D(x; F) \to 0$ quand $\|x\| \to \infty$ pour chaque $F \in \mathcal{F}$.

(P6) Soit $F \in \mathcal{F}$ ayant pour distributions marginales $F_1, \dots, F_d$. Supposons que pour $i \in \{1, \dots, d\}$, le support de F_i contient un unique point $\{\alpha\}$. Alors pour tout $x^* \in \operatorname{argsup}_{x \in \mathbb{R}^d} D(x; F)$, la $i^{ème}$ coordonnée de x^* est $x^*_i = \alpha$.

Alors $D(\cdot; \cdot)$ est appelée **fonction de profondeur**.

La définition d'une fonction de profondeur varie selon les auteurs. Cette définition est moins contraignante que certaines, dans la mesure où nous avons restreint les conditions à ce qui est nécessaire dans le cadre des votes.

Il existe dans la littérature de nombreuses fonctions de profondeur, par exemple *halfspace*, par projection, d'Oja, les profondeurs par moyennes pondérées, etc. Nous renvoyons à [4] et [5] pour une revue détaillée de celles-ci. Une visualisation des fonctions de profondeur sur un exemple est donné en Figure 1.

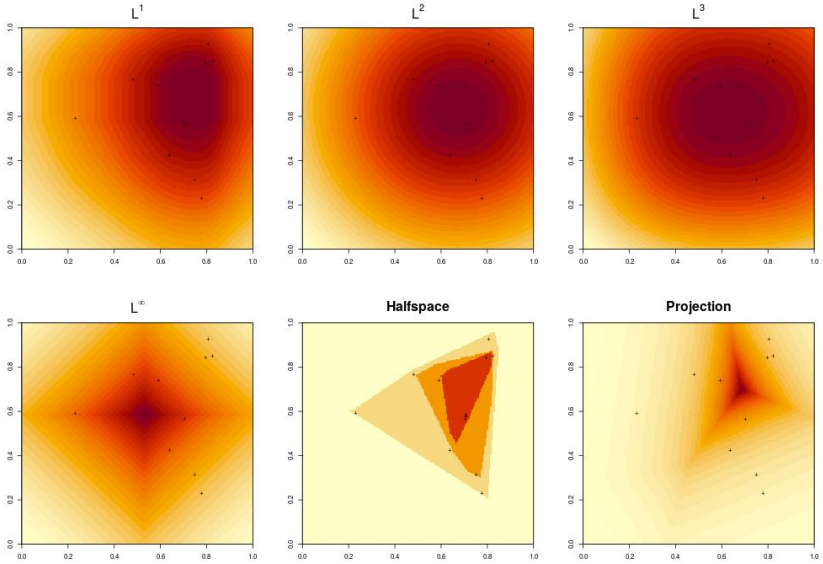


Figure 1: Exemples de fonctions de profondeur sur une même matrice de notes avec 2 candidats et 5 électeurs. L'axe horizontal correspond aux notes données pour le candidat c_1 et l'axe vertical aux notes pour le candidat c_2. Chaque croix représente un électeur.

3 Les votes par profondeur

Definition 2 (Vote par profondeur)**.** *Considérons une matrice de notes* Φ*, et une fonction de profondeur* D*. Soit*

$$\mathcal{X}_D^* := \{x \in \mathbb{R}^d : D(x, \Phi) = \sup(D(., \Phi))\}$$

l'ensemble des points les plus profonds Φ *en fonction de* D*. La fonction* $G_D : \Phi \to \mathcal{X}_D^*$ *est une procédure de vote associée à la profondeur* D*.*

Soit

$$i_D := \underset{1 \leq i \leq d}{\operatorname{argmax}} \{x_{D,i}^*,\ x_D^* = (x_{D,1}^*, \ldots, x_{D,d}^*) \in \mathcal{X}_D^*\}.$$

Si i_D *est unique, le vainqueur de l'élection associé à la procédure* G_D *est le candidat* c_{i_D}*. Si* i_D *n'est pas unique, il n'y a pas de vainqueur.*

Une visualisation des procédures de vote associées à diverses fonctions est représentée en Figure 2.

Dans [6], il a été montré que si D est une fonction de profondeur vérifiant la définition 1, alors la procédure de vote G_D associée à D vérifie les propriétés de *Neutralité*, d'*Universalité* et d'*Unanimité*. La propriété de *Monotonie* a été établie pour certaines classes de fonctions, mais il apparaît que des fonctions de profondeur usuelles telles que la fonction *halfspace* ou la fonction de Oja ne vérifient pas la *Monotonie*. Enfin, la condition *IIA*

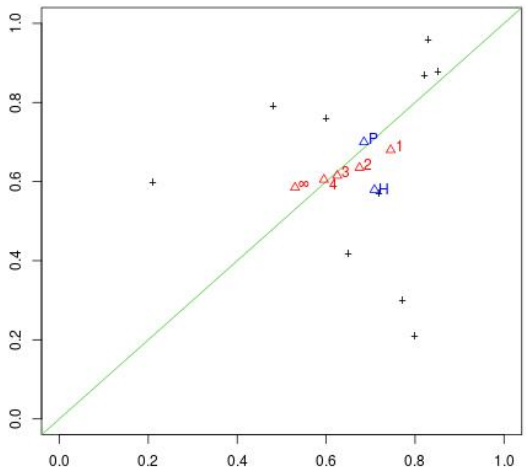


Figure 2: Résultats de procédures de vote basées sur des fonctions de profondeur dans l'exemple de la Figure 1. L'axe horizontal correspond aux notes données pour le candidat c_1 et l'axe vertical aux notes pour le candidat c_2. Chaque croix représente un électeur et chaque triangle un résultat de vote. Les résultats pour les profondeurs L^p, $p \in \{1, 2, 3, 4, \infty\}$, sont donnés par le label p. Les résultats pour les profondeurs *halfspace* et par projection sont données respectivement avec les labels H et P.

correspond à un ensemble de points le plus profond obtenu composante par composante, ce qui signifie que les scores finaux sont calculés indépendamment pour chaque candidat. Ces résultats sont synthétisés dans la Table 1.

Fonction de profondeur	Unicité	*Neutralité*	*Universalité*	*Unanimité*	*Monotonie*	*IIA*
L^1	N	✓	✓	✓	✓	✓
L^p avec $1 < p \leq \infty$	✓	✓	✓	✓	✓	✓
Halfspace	N	✓	✓	✓	N	N
Oja	N	✓	✓	✓	N	N
Projection	✓	✓	✓	✓	N	N
Moyennes pondérées	✓	✓	✓	✓	✓	✓

Table 1: Cette table résume les propriétés de procédures de votes pour diverses fonctions de profondeur. Elle précise si la fonction G_D associe un unique point (le plus profond pour D), et si les conditions définies dans la section 1 sont vérifiées (avec $d > 1$ candidats). ✓ correspond à des propriétés valides et N à des non-valides. Remarquons que les moyennes pondérées et la L^2 donnent la même procédure de vote.

Comme précisé en Table 1, pour $p \geq 1$, la procédure de vote associée à la fonction de profondeur L^p vérifie la *Neutralité*, l'*Universalité*, l'*Unanimité*, la *Monotonie* et l'*IIA*. Cette classe des profondeur L^p est intéressante dans la mesure où elle inclut trois procédures de vote classiques qui sont le jugement majoritaire [1], le vote par approbation [2] et le *range voting* [3]. Ces procédures correspondent respectivement aux fonctions L^2, L^1 et L^1 avec un support des notes réduit à $\Lambda = \{0; 1\}$. Le choix de p dans la profondeur L^p est critique. Schématiquement, plus p est élevé, plus on accorde d'importance aux notes extrêmes. Nous renvoyons à [6] pour une analyse plus approfondie de cette famille de vote par profondeur.

Conclusion

Le vote basé sur des évaluations offre une information plus riche que le modèle basé sur des préférences. Dans cet article, nous introduisons un nouveau point de vue sur les procédures de vote, en faisant appel à la notion statistique de profondeur. Les fonctions de profondeur offrent une manière unifiée de définir les processus de vote, avec un support pour les notes Λ discret ou continu. Des procédures de vote classiques telles que le jugement majoritaire, le vote par approbation et le *range voting* correspondent à des votes par profondeur liés aux profondeurs L^p. Ce formalisme devrait pouvoir également permettre de proposer de nouvelles procédures de vote en choisissant une fonction de profondeur adaptée à la situation considérée.

Les procédures de vote par profondeur vérifient des propriétés usuellement demandées, à savoir la neutralité, l'universalité et l'unanimité. La monotonie des processus de vote est établie pour plusieurs fonctions, telles que les profondeurs L^p, mais certaines fonctions de profondeur usuelles, telles que la fonction *halfspace*, par ou la profondeur d'Oja, ne satisfont pas la monotonie. La non-vérification de cette propriété semble un inconvénient majeur pour ces processus de vote. Nous établissons finalement que l'indépendance aux alternatives non pertinentes est liée à une caractérisation composante par composante des points de profondeur maximale. Nous renvoyons à [6] pour une étude plus approfondie de ces propriétés.

Bibliographie

[1] Michel Balinski and Rida Laraki. A theory of measuring, electing, and ranking. *Proceedings of the National Academy of Sciences*, 104(21):8720–8725, 2007.

[2] Steven Brams and Peter C. Fishburn. *Approval voting*. Springer, 2007.

[3] Warren D. Smith. Range voting. http://rangevoting.org/RangeVoting.html, 2000. Accessed: 2014-10-12.

[4] Yijun Zuo and Robert Serfling. General notions of statistical depth function. *Annals of statistics*, pages 461–482, 2000.

[5] Karl Mosler. Depth statistics. In *Robustness and complex data structures*, pages 17–34. Springer, 2013.

[6] Jean-Baptiste Aubin, Irène Gannaz, Samuela Leoni, and Antoine Rolland. Deepest voting: A new way of electing. *Mathematical Social Sciences*, 116:1–16, 2022.

Impact of the energy transition on long-term factor productivity

Guillaume Flament[1,2]

[1] *Ensai, 51 Rue Blaise Pascal, 35172 BRUZ, guillaume.flament@ensai.fr*
[2] *Square Research Center, 163 avenue Achille Peretti, 92200 Neuilly-sur-Seine, guillaume.flament@adway-conseil.fr*

Résumé. Dans cet article, nous construisons un modèle hiérarchique Bayésien, pour endogénéiser la trajectoire de la productivité des facteurs puis, nous illustrons cette modélisation à travers le modèle DICE proposé par Nordhaus (2018). Les données montrent que la variation de l'*exergy* primaire est un bon prédicteur de la variation de la productivité des facteurs. Grâce à cette modélisation de la productivité des facteurs, il est possible de générer de nouveaux scénarios économiques respectant une trajectoire d'émissions de gaz à effet de serre. Ces scénarios sont particulièrement intéressants pour l'élaboration de politiques publiques. De plus, notre modèle est capable de conciler des projections très différentes, d'une part, si l'*exery* augmente, alors nous obtenons des scénarios similaires à ceux de Nordhaus (2018), à l'inverse, si elle diminue, alors nous obtenons des scénarios similaires à Meadows et al. (2013).

Mots-clés. Modèle intégré, risque climatique et de transition, modèle bayésien hiérarchique, prédictions à long-terme

Abstract. In this article, we propose to build a statistical Bayesian model for factor productivity and illustrate our predictions on a common toy model : DICE (Nordhaus, 2018). Data show that the variation of primary *exergy*, or simply called *exergy* is a good predictor of the Total Factor Productivity variation. That remark allows us to design new economic scenarios that respect a given global warming threshold. These scenarios are of particular interest for policy makers. Moreover, our model is capable to reconcile different projections by the introduction of environmental constraints. On one hand, if *exergy* increases, we have similar projections with standard economic models. On the other hand, if *exergy* decreases, we obtain projections closer to other Integrated Assessment Models, where economic output declines such as World 3.

Keywords. Integrated assessment model, Climate risk, transition risk, long-run predictions, bayesian hierarchical model

1 Introduction

Depuis 1988, le Groupe d'experts intergouvernemental sur l'évolution du climat (GIEC) tente d'évaluer et comprendre l'impact des activités humaines sur le climat. Ce groupe a très récemment publié un dernier rapport inquiétant : IPCC (2021). Au moment de la création du GIEC, Nordhaus (1991) a été le premier à proposer un modèle étudiant les liens entre le monde économique et le changement climatique. Dans ce modèle dénommé DICE (Dynamic Integrated model of Climate and Economy) basé sur une analyse coût bénéfice, il discute pour la première fois de l'impact d'une réduction des émissions de CO_2 sur des trajectoires économiques de long-terme. Son travail, a semble-t-il eu un grand succès puisqu'il a obtenu le prix de la banque de Suède en sciences économiques en mémoire d'Alfred Nobel en 2018 pour l'introduction du risque climatique dans les modèles macroéconomiques. Ces travaux ont donc influencé la recherche économique et les scénarios utilisés par les instances de régulation en sont largement inspirés. Dans le modèle DICE, la croissance du PIB provient principalement de l'augmentation de la productivité des facteurs et sa trajectoire est simplement fixée par le modélisateur, on dit qu'elle est exogène. Des économistes ont émis des critiques à l'encontre de cette hypothèse, en particulier Stern (2013) qui précise que la croissance économique ne devrait pas être considérée comme acquise et des modélisations doivent être développées pour mieux comprendre son origine, en somme, cela revient à endogénéiser cette variable.

Nous souhaitons donc modifier le modèle DICE en endogénéisant la productivité des facteurs par un modèle Bayésien. Nous présenterons d'abord rapidement le modèle DICE avant de nous intéresser plus précisément à notre modélisation pour conclure en comparant nos projections avec celles utilisées par la banque centrale européenne ou celles obtenues par Nordhaus and Sztorc (2013).

1.1 Présentation du modèle DICE

Les travaux de Nordhaus s'appuient sur les travaux de Solow qui modélise le PIB à partir de deux facteurs de production, l'un représente la quantité de travail ($L(t)$) et l'autre l'investissement passé en outils de production ($K(t)$) (machines, usines, infrastructures, ...). En introduisant seulement ces facteurs, Solow n'arrive pas à expliquer la trajectoire observée du PIB. Il introduit alors un terme d'erreur pour corriger ses prédictions. Ce terme est la productivité des facteurs qui sera ensuite interprété comme l'amélioration de l'efficacité de l'économie (avancée technologique, meilleure organisation, ...). Seulement, cette interprétation ne nous semble pas suffisante puisqu'elle interdit toutes chutes importantes du PIB. En effet, il est peu satisfaisant de penser que les avancées technologiques ou organisationnelles d'une année soient perdues l'année suivante. Ainsi, il est difficile de modéliser des variations brusques à la baisse de la TFP. Reprenons le modèle classique utilisé par Nordhaus (2018), le PIB sera modélisé comme suit :

$$Y(t) = \Omega(t)A(t)K(t)^{\beta}L(t)^{1-\beta}.$$

Comme dans le modèle DICE, nous fixerons : $\beta = 0.3$. L'évolution du stock de capital évoluera selon l'équation :

$$K(t) = (1 - \delta_k)K(t-1) + I(t)$$

La fonction Ω rend compte des effets néfastes sur l'économie du changement climatique. Malheureusement, calibrer ces fonctions est très compliqué puisqu'aucun réchauffement de cet ampleur n'a été observé, nous testerons les projections avec deux fonctions populaires dans la littérature (Nordhaus and Sztorc, 2013):

$$\Omega_N(t) = \{1 + 0.026 \cdot T(t)^2\}^{-1}. \tag{1}$$

Et, une autre, qui quantifie mieux les dommages extrêmes Weitzman (2010) :

$$\Omega_W(t) = \{1 + 0.00239 \cdot T(t)^2 + 0.000156 \cdot T(t)^7\}^{-1}. \tag{2}$$

Enfin, en fixant la quantité de travail à partir de projections de données démographiques et en considérant I comme un taux d'épargne, il ne reste qu'à calibrer la productivité des facteurs pour être en mesure d'obtenir des projections économiques.

Symbole	Signification	Symbole	Signification
$Y_c(t)$	Produit Intérieur Brut (PIB)	$T(t)$	Variation de température entre l'ère pré-industriellee et l'année t
$A_c(t)$	Productivité des facteurs (TFP)		
$K_c(t)$	Capital	$E_c(t)$	*Exergy*
$L_c(t)$	Travail	$\mathcal{C}$	Liste des pays
$I_c(t)$	Investissement	GES	Gaz à effet de serre

Table 1: Liste des notations. Lorsque cela n'est pas nécessaire, nous enlèverons l'indice représentant le pays. Aussi, une lettre minuscule correspond au logarithme d'une lettre majuscule. Par exemple, $\log(A(t)) = a(t)$, l'opérateur différence sera noté Δ. le symbole t correspond à une année, ainsi, $Y_c(t)$ correspond au PIB du pays c durant l'année t.

1.2 Les données

Pour construire notre modèle, nous nous appuierons sur des données d'énergie primaire de l'IEA pour les pays du G7 et les données de TFP issues du travail de Feenstra et al. (2015). Nos données remontent à 1971 et s'arrêtent en 2019. À partir de ces données d'énergie primaire, nous calculerons l'*exergy*, notre variable explicative, grâce à la méthode présentée par Brockway et al. (2014).

1.3 Modélisation Bayésienne

Le travail de Nordhaus and Sztorc (2013) suppose que la TFP évolue de l'ordre de 1.5% par an. Nous voulons changer cette hypothèse par une modélisation de la variation de la TFP en fonction de la variation de l'*exergy*. Cette quantité physique représente la quantité d'énergie et l'efficacité de sa transformation. Certains auteurs considèrent que l'*exergy* permet de modéliser la productivité des facteurs (Bercegol and Benisty, 2022; Santos et al., 2021), une première relation pourrait être :

$$\Delta a_c(t) = \alpha_{0,c} + \alpha_{1,c}\Delta a(t-1) + \beta_{0,c}\Delta e_c(t) + \epsilon_{c,t}. \tag{3}$$

Où Δ est l'opérateur différence et $\epsilon_c(t)$ un bruit. Une problématique importante est qu'il est nécessaire de compter les importations, il est peu pertinent de mener ce genre d'étude sur un unique pays qui pourrait avoir délocalisé l'ensemble de ses industries polluantes, pour ne garder que des services tout en important massivement des produits manufacturés d'autres pays. Malheureusement, nous n'avons pas de données détaillées sur le contenu énergétique des importations, la première approche négligera ces aspects. Dans un second temps, nous chercherons à corriger les problèmes d'importations et exportations en considérant la variation de l'*exergy* mondiale comme seconde variable explicative :

$$\Delta a_c(t) = \alpha_{0,c} + \alpha_{1,c}\Delta a(t-1) + \beta_{0,c}\Delta e_c(t) + \beta_{1,c}\Delta \sum_{c\in\mathcal{C}} e_c(t) + \epsilon_{c,t}. \tag{4}$$

Ces modèles peuvent être schématisés par :

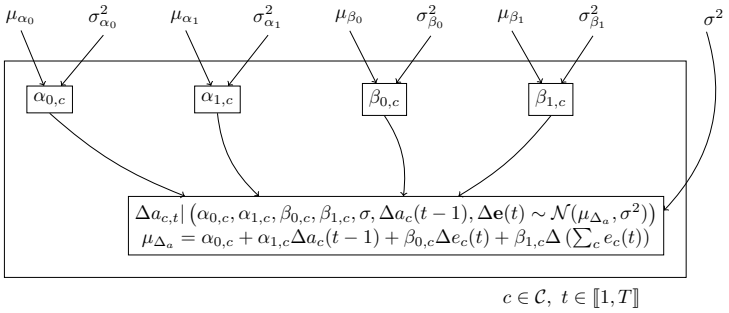


Figure 1: Illustration du modèle Bayésien hiérarchique proposé. Ici, par exemple, $\alpha_{0,c}$ aura pour prior une loi normale de moyenne μ_{α_0} et de variance $\sigma^2_{\alpha_0}$. Différentes lois et paramètres ont été testées, l'ensemble des simulations sont disponibles sur ce dépôt Github

L'idée de cette approche est de modéliser la variation de la production des facteurs par une composante spécifique au pays et une composante mondiale. Cette deuxième composante est une première approche pour traduire l'effet des importations sur l'économie d'un pays. Deuxièmement, ajouter des a priori, nous permet de modéliser le fait que les économies de ces pays dépendent de manière similaire à l'*exergy*, c'est aussi une solution contre la faible profondeur de notre historique et améliorer la précision de nos estimateurs.

Pays	Variation émissions GES (%)	Variation *Exergy* estimée (%)	Prédiction Modèle (3) (%)	Prédiction Modèle (4) (%)	Variation PIB observée (%)
États-Unis	-10.9	-10.9	-2.0 $(-2.6, -1.4)$	$-2.1(-2.9, -1.3)$	-2.3
France	-14.3	-14.3	-1.8 $(-2.7, -0.8)$	$-1.9(-2.9, -0.1)$	-4.2
Canada	-9.6	-9.6	-2.1 $(-2.8, -1.5)$	$-2.9(-3.9, -1.9)$	-5.6
Italie	-10.7	-10.7	-2.7 $(-3.6, -2)$	$-5.8(-6.7, -4.8)$	-5.9
Allemagne	-9.7	-9.7	-1 $(-1.6, -0.4)$	$-2.5(-3.8, -1.3)$	-1.4
Japon	-6.5	-6.5	-0.9 $(-1.3, -0.5)$	$-2(-3.2, -0.09)$	-1.7
Royaume-uni	-11.2	-11.2	-1.7 $(-2.4, -1.1)$	$-3(-4.2, -1.8)$	-4

Table 2: Prédictions des modèles (3) et (4) durant la crise du Covid-19 (non observé dans les données), en comparaison des variations de PIB observées. La moyenne a posteriori est donnée et les quantiles à 0.1 et 0.9 sont entre parenthèses

A notre connaissance cette approche est la seule capable de modéliser et d'obtenir des résultats cohérents avec les observations lors de crises, que ce soit la crise financière de 2008 ou la crise du Covid-19 (voir Table 2).

1.4 Résultats et conclusions

Grâce à notre modèle Bayésien, et des projections sur un mix énergétique respectant un scénario d'émissions de gaz à effet de serre, nous sommes capable de générer des scénarios macroéconomiques à long-terme (voir Figure 2). Pour faciliter la comparaison avec les projections proposées par Nordhaus and Sztorc (2013), nous avons choisi d'utiliser les paramètres qu'il définit en 2010 $(K, I, L, \dots)$, en intégrant seulement notre modélisation de la TFP. Nous avons ensuite, projeté à un horizon 30 ans, qui correspond à l'horizon fixé par les organismes de régulation pour les tests réglementaires.

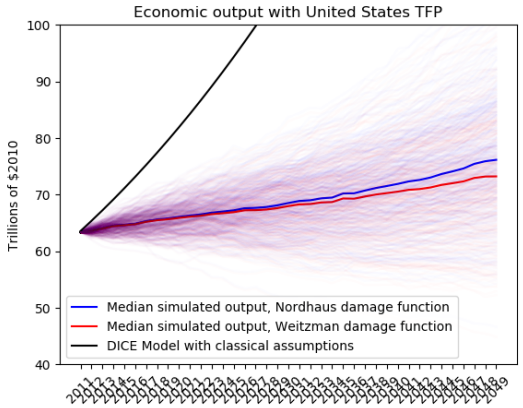


Figure 2: Projections à long-terme avec le modèle (4) pour les deux fonctions de dommages Ω_N (1) et Ω_W (2) et les hypothèses classiques de Nordhaus and Sztorc (2013).

Nous obtenons des trajectoires à horizon 2050 où la production est environ 3 fois plus faible

que celle prédite par le modèle DICE, quant aux scénarios proposés par le régulateur, qui permettent de tester la résistance des institutions bancaires dans des scénarios pessimistes, ils correspondent à nos scénarios les plus optimistes.

Bibliographie

Bercegol, H. and Benisty, H. (2022). An energy-based macroeconomic model validated by global historical series since 1820. *Ecological Economics*, 192:107253.

Brockway, P. E., Barrett, J. R., Foxon, T. J., and Steinberger, J. K. (2014). Divergence of trends in us and uk aggregate exergy efficiencies 1960–2010. *Environmental science & technology*, 48(16):9874–9881.

Feenstra, R. C., Inklaar, R., and Timmer, M. P. (2015). The next generation of the penn world table. *American economic review*, 105(10):3150–82.

IPCC (2021). The sixth assessment report - the numbers behind the science.

Meadows, D. H., Randers, J., and Meadows, D. L. (2013). *The Limits to Growth (1972)*. Yale University Press.

Nordhaus, W. (2018). Evolution of modeling of the economics of global warming: changes in the dice model, 1992–2017. *Climatic change*, 148(4):623–640.

Nordhaus, W. and Sztorc, P. (2013). User's manual for dice-2013r.

Nordhaus, W. D. (1991). To slow or not to slow: the economics of the greenhouse effect. *The economic journal*, 101(407):920–937.

Santos, J., Borges, A. S., and Domingos, T. (2021). Exploring the links between total factor productivity and energy efficiency: Portugal, 1960–2014. *Energy Economics*, 101:105407.

Stern, N. (2013). The structure of economic modeling of the potential impacts of climate change: grafting gross underestimation of risk onto already narrow science models. *Journal of Economic Literature*, 51(3):838–59.

Weitzman, M. L. (2010). What is the" damages function" for global warming—and what difference might it make? *Climate Change Economics*, 1(01):57–69.

Pl@ntNet-300K: une base de données d'images de plantes avec une ambiguïté entre classes importante et une distribution à longue traîne.

Camille Garcin [1,2] & Maximilien Servajean [3,4] & Alexis Joly [2] & Joseph Salmon [1,5]

[1] *IMAG, Université de Montpellier, CNRS, Montpellier, France*
[2] *Inria, LIRMM, Univ Montpellier, CNRS, Montpellier, France*
[3] *LIRMM, Univ Montpellier, CNRS, Montpellier, France*
[4] *AMIS, Paul Valery University, Montpellier, France*
[5] *Institut Universitaire de France (IUF)*

Résumé. Cet article présente un nouveau jeu de données d'images à forte ambiguïté intrinsèque et à distribution à longue traîne, construit à partir de la base de données de l'observatoire citoyen Pl@ntNet. Il est composé de 306 146 images de plantes couvrant 1 081 espèces. Nous soulignons deux caractéristiques particulières du jeu de données, inhérentes à la manière dont les images sont acquises et à la diversité intrinsèque de la morphologie des plantes : (i) le jeu de données présente un fort *déséquilibre entre classes*, c'est-à-dire que quelques espèces seulement représentent la plupart des images, et, (ii) de nombreuses espèces sont *visuellement similaires*, rendant l'identification difficile même pour un œil expert.

Mots-clés. Apprentissage profond, jeu de données, plantes, distribution à longue traîne, ambiguïté

Abstract. This paper presents a novel image dataset with high intrinsic ambiguity and a long-tailed distribution built from the database of Pl@ntNet citizen observatory. It consists of 306 146 plant images covering 1 081 species. We highlight two particular features of the dataset, inherent to the way the images are acquired and to the intrinsic diversity of plants morphology: (i) the dataset has a strong *class imbalance*, *i.e.,* a few species account for most of the images, and, (ii) many species are *visually similar*, rendering identification difficult even for the expert eye.

Keywords. Deep Learning, dataset, plants, long-tailed distribution, class ambiguity

1 Introduction

Lors de la classification d'images, nous sommes confrontés à deux grands types d'incertitudes [Der Kiureghian and Ditlevsen, 2009] : (i) l'incertitude des données qui découle du caractère aléatoire intrinsèque du processus sous-jacent et qui est considérée comme irréductible, et (ii) l'incertitude du modèle qui est causée par un manque de connaissances

et qui est considérée comme réductible avec des données d'entraînement supplémentaires. Dans certaines applications modernes, ces deux types d'incertitudes sont particulièrement difficiles à gérer. Le nombre toujours croissant de classes à distinguer tend à augmenter le recouvrement des classes (et donc l'incertitude des données) et, d'autre part, la distribution à longue traîne rend difficile l'apprentissage des classes moins peuplées (et donc augmente l'incertitude du modèle). Bien qu'il existe plusieurs jeux de données dans la littérature qui présentent des classes visuellement similaires : [Nilsback and Zisserman, 2008, Maji et al., 2013, Yang et al., 2015, Russakovsky et al., 2015], la plupart d'entre eux ne visent pas à conserver à la fois l'incertitude du modèle et l'incertitude des données.

Dans cet article, nous proposons un jeu de données conçu pour rester représentatif de l'ambiguïté rencontrée dans des problèmes réels. Ce jeu de données est extrait d'images collectées avec l'application mobile Pl@ntNet [Affouard et al., 2017], qui permet aux utilisateurs d'envoyer une photo d'une plante qu'ils rencontrent et d'obtenir en retour une liste des espèces les plus probables pour cette photo. L'application est utilisée par plus de 10 millions d'utilisateurs dans environ 170 pays et couvre plus de 35 000 espèces illustrées par près de 12 millions d'images validées. Une caractéristique essentielle de Pl@ntNet est que l'ensemble d'apprentissage utilisé pour entraîner le classifieur est enrichi et révisé de manière collaborative.

La base de données comprenant toutes les images collectées par Pl@ntNet est beaucoup trop volumineuse pour permettre une utilisation généralisée par la communauté de l'apprentissage automatique. L'extraction d'un sous-échantillon de la base de données doit être effectuée avec précaution car nous voulons préserver l'incertitude naturellement présente dans l'ensemble de la base de données. Le jeu de données présenté dans cet article est construit en ne retenant qu'un sous-ensemble des genres de l'ensemble de la base de données Pl@ntNet (échantillonné uniformément au hasard). Toutes les espèces appartenant aux genres sélectionnés sont ensuite retenues. Cela permet de maintenir l'ambiguïté initiale, car les espèces d'un même genre sont susceptibles de partager des caractéristiques visuelles communes.

Dans la Section 2, nous décrivons la procédure de construction du jeu de données Pl@ntNet-300K. Enfin, nous fournissons le lien vers le jeu de données dans la Section 3 avant de conclure.

Le document présent est un condensé de l'article Garcin et al. [2021]. L'article cité comprend notamment une introduction à classification d'ensembles et l'intérêt de Pl@ntNet-300K pour l'évaluation de telles méthodes, avec la description des métriques associées. De plus, de nombreuses architectures de réseaux de neuronnes sont utilisées et leurs performances comparées, illustrant notamment la difficulté de Pl@ntNet-300K du fait de l'ambiguïté et de la distribution à longue traîne.

2 Construction du jeu de données

2.1 Validation des étiquettes et nettoyage des données

La validation des étiquettes est basée sur un algorithme de vote majoritaire pondéré prenant en entrée les étiquettes proposées par les utilisateurs de Pl@ntNet avec un principe de pondération adaptatif selon l'expertise et l'engagement de l'utilisateur. Ainsi, un seul annotateur de confiance peut suffire à valider l'étiquette d'une image. En revanche, les images dont les étiquettes sont proposées par plusieurs utilisateurs novices peuvent ne pas être validées car elles n'ont pas un poids suffisant. Au moment de la construction de Pl@ntNet-300K, le nombre total d'annotateurs dans la base de données Pl@ntNet était égal à 2.079.003. Le nombre moyen d'annotateurs par image est égal à 2,03. En plus de la procédure de validation des étiquettes, le pipeline de Pl@ntNet comprend d'autres procédures de nettoyage des données : (i) filtrage automatique des contenus inappropriés ou non pertinents (visages, humains, animaux, bâtiments, etc.) à l'aide d'un CNN, et (ii) filtrage sur la qualité des images (évaluée par les utilisateurs).

2.2 Construction de Pl@ntNet-300K

En taxonomie, les espèces sont organisées en genres, chaque genre contenant une ou plusieurs espèces, et les différents genres ne se chevauchent pas, comme illustré dans la Figure 1. Au lieu de retenir des espèces ou des images choisies au hasard dans l'ensemble

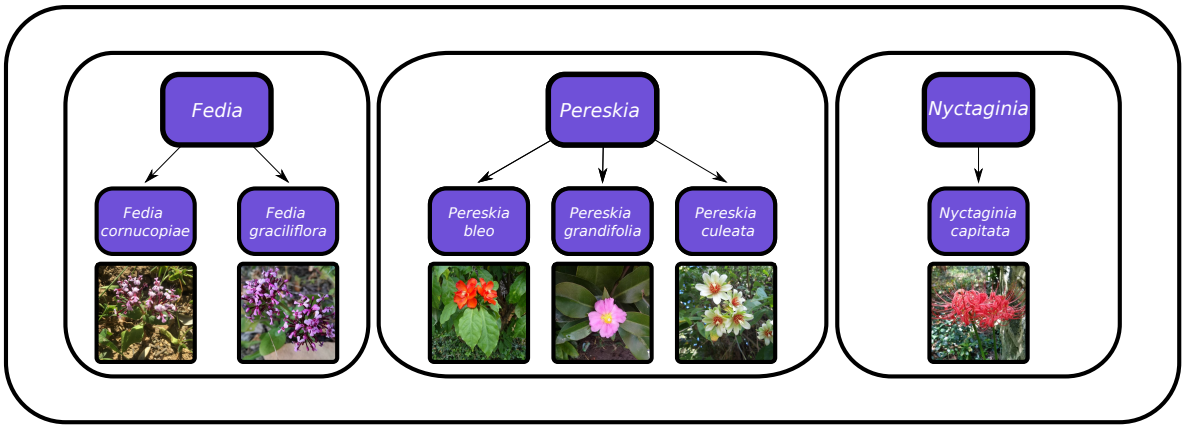


Figure 1: Taxonomie des genres : nous présentons trois genres présents dans l'ensemble de données proposé—*Fedia*, *Pereskia* et *Nyctaginia*— qui contiennent respectivement deux, trois et une espèces.

de la base de données Pl@ntNet, nous choisissons de retenir des genres choisis au hasard et de conserver toutes les espèces appartenant à ces genres. Ce choix vise à préserver la grande quantité d'ambiguïté présente dans la base de données originale, car les espèces

appartenant au même genre ont tendance à partager des caractéristiques visuelles. Le jeu de données présenté dans cet article est construit en échantillonnant uniformément au hasard seulement 10% des genres de l'ensemble de la base de données Pl@ntNet.

Nous ne retenons ensuite que les espèces ayant plus de 4 images, ce qui donne un total de 303 genres et $L = 1\,081$ espèces. Les images sont divisées en un ensemble d'entraînement, un ensemble de validation et un ensemble de test.[1] Pour chaque espèce, 80% des images sont placées dans l'ensemble d'apprentissage ($n_{train} = 243\,916$), 10% dans l'ensemble de validation ($n_{val} = 31\,118$), et 10% dans l'ensemble de test ($n_{test} = 31\,112$), avec au moins une image de chaque espèce dans chaque ensemble. Cela représente un total de $n_{tot} = n_{train} + n_{val} + n_{test} = 306\,146$ images en couleur. La taille moyenne des images est de 570×570, allant de 180×180 à 900×900. La construction de l'ensemble de données préserve le déséquilibre des classes. Pour le montrer, nous traçons les courbes de Lorenz [Gastwirth, 1971] de l'ensemble du jeu de données Pl@ntNet et celle du jeu de données Pl@ntNet-300K dans la Figure 2.

2.3 Incertitude du modèle

L'incertitude du modèle se réfère principalement au manque de données nécessaires pour estimer correctement les probabilités conditionnelles. Dans Pl@ntNet, les espèces les plus communes sont facilement observées par les utilisateurs dans la nature et représentent donc une grande fraction des images, tandis que les espèces les plus rares sont plus difficiles à trouver et donc moins fréquentes dans la base de données. Dans la Figure 2, nous constatons que 80% des espèces (celles dont le nombre d'images est le plus faible) ne représentent que 11% du nombre total d'images. Par conséquent, l'entraînement des modèles d'apprentissage automatique est difficile pour un tel ensemble de données, car pour de nombreuses classes, le modèle ne dispose que d'une poignée d'images sur lesquelles s'entraîner, ce qui rend l'identification difficile pour ces espèces.

2.4 Incertitude des données

Dans notre cas, la source d'incertitude des données réside principalement dans les informations limitées qui nous sont données pour prendre une décision (attribuer une étiquette à une plante). Certaines espèces, notamment celles appartenant au même genre, peuvent être visuellement très similaires. Par exemple, considérons le cas où deux espèces produisent les mêmes fleurs mais des feuilles différentes, généralement parce qu'elles ont évolué différemment à partir de la même espèce mère. Si une personne photographie uniquement la fleur d'un spécimen de l'une des deux espèces, il sera impossible, même pour un expert, de savoir à quelle espèce la fleur appartient. L'information discriminante n'est pas présente dans l'image. Cette ambiguïté irréductible se traduit par des

[1] La division est effectuée au niveau de l'espèce en raison de la distribution à longue traîne.

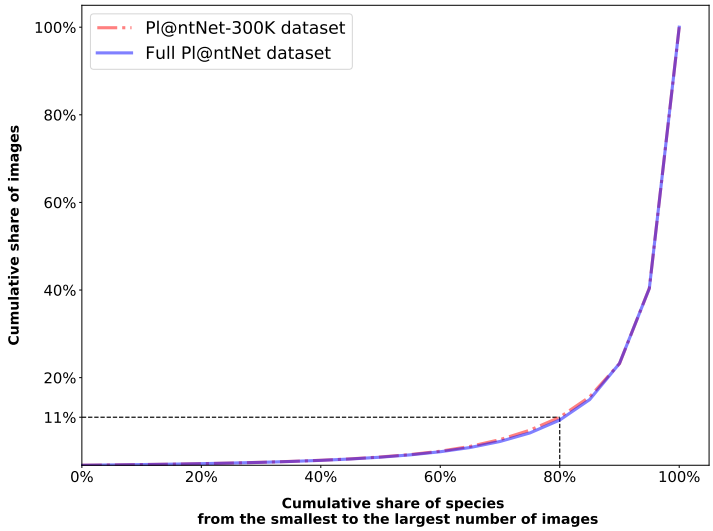


Figure 2: Courbes de Lorenz de la base de données Pl@ntNet et de l'ensemble de données proposé. Notez que, pour une comparaison équitable, nous écartons les espèces ayant moins de 4 images dans la base de données Pl@ntNet.

paires d'images appartenant à des espèces différentes mais difficiles voire impossibles à distinguer, voir Figure 3.

Figure 3: Exemples d'images similaires appartenant à deux classes différentes.

3 Lien vers le dataset

Le jeu de données Pl@ntNet-300K est disponible au lien suivant : https://doi.org/10.5281/zenodo.5645731. Il est organisé en trois dossiers nommés "`train`", "`val`" et "`test`". Chacun de ces dossiers contient $L = 1\,081$ sous-dossiers. Nous fournissons la correspondance entre les noms des sous-dossiers et les noms des classes dans le fichier "`plantnet300K_species_id_2_name.json`". Les noms des classes sont de la forme *Genus_species*, *e.g., Cymbalaria_aequitriloba*. Nous fournissons également un fichier de métadonnées nommé "`plantnet300K_metadata.json`" contenant pour chaque image les informations suivantes : l'identifiant de l'espèce (classe), l'organe de la plante (fleur, feuille, ...), le nom de l'auteur, la licence et le split (*i.e.,* train, validation ou test set).

Bibliographie

Antoine Affouard, Hervé Goëau, Pierre Bonnet, Jean-Christophe Lombardo, and Alexis Joly. Pl@ntnet app in the era of deep learning. In *ICLR - Workshop Track*, 2017.

Armen Der Kiureghian and Ove Ditlevsen. Aleatory or epistemic? does it matter? *Structural safety*, 31(2):105–112, 2009.

Camille Garcin, Alexis Joly, Pierre Bonnet, Jean-Christophe Lombardo, Antoine Affouard, Mathias Chouet, Maximilien Servajean, Joseph Salmon, and Titouan Lorieul. Pl@ntNet-300K: a new plant image dataset for the evaluation of set-valued classifiers. In *NeurIPS Datasets and Benchmarks*, 2021.

Joseph L Gastwirth. A general definition of the Lorenz curve. *Econometrica: Journal of the Econometric Society*, pages 1037–1039, 1971.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729. IEEE, 2008.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vision*, 115(3):211–252, 2015.

Linjie Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. A large-scale car dataset for fine-grained categorization and verification. In *CVPR*, pages 3973–3981, 2015.

Variance estimators for weighted and stratified linear dose-response function estimators using generalized propensity score

Valérie Garès [1] & Guillaume Chauvet [2] & David Hajage [3]

[1] *Univ Rennes, INSA, CNRS, IRMAR - UMR 6625, F-35000 Rennes, France, valerie.gares@insa-rennes.fr*

[2] *Univ Rennes, ENSAI, CNRS, IRMAR - UMR 6625, F-35000 Rennes, France, chauvet@ensai.fr*

[3] *Sorbonne Université, INSERM, Institut Pierre Louis d'Epidémiologie et de Santé Publique, AP-HP, Hôpital Pitié-Salpêtrière, Département de Santé Publique, Centre de Pharmacoépidémiologie, Paris, France*

Résumé. Les méthodes utilisant le score de propension sont largement utilisées dans les études observationnelles pour évaluer les effets marginaux du traitement. Le score de propension généralisée (GPS) est une extension du score de propension, historiquement développé dans le cas des expositions binaires, développé pour des expositions quantitatives ou continues. Nous proposons des estimateurs de variance pour les estimateurs de l'effet du traitement sur des variables d'intéret continues. Les fonctions dose-réponse (DRF) ont été estimées en pondérant par l'inverse du GPS ou en utilisant la stratification. Les estimateurs de variance ont été évalués à l'aide de simulations Monte Carlo. Malgré l'utilisation de poids stabilisés, la variabilité de l'estimateur pondéré du DRF était particulièrement élevée, et aucun des estimateurs de la variance n'a pu capturer adéquatement cette variabilité, ce qui a donné des taux de couvertures inférieures à la valeur nominale, en particulier lorsque la proportion de la variabilité de l'exposition quantitative expliquée par les covariables était importante. L'estimateur stratifié était plus stable et les estimateurs de la variance plus efficaces pour saisir la variabilité empirique des paramètres du DRF. L'estimateur de la variance linéarisé groupé et l'estimateur groupé basé sur le modèle avaient tendance à surestimer la variance, tandis que l'estimateur bootstrap, qui prend intrinsèquement en compte l'étape d'estimation du GPS, a donné des estimations de la variance et des taux de couverture corrects. Ces méthodes ont été appliquées à un ensemble de données réelles dans le but d'évaluer l'effet de l'indice de masse corporelle de la mère sur le poids du nouveau-né à la naissance.

Mots-clés. Score de propension généralisée, étude observationnelle, exposition quantitative, estimateur de variance

Abstract. Propensity score methods are widely used in observational studies for evaluating marginal treatment effects. The generalized propensity score (GPS) is an extension of the propensity score framework, historically developed in the case of binary exposures, for use with quantitative or continuous exposures. We proposed variance estimators for treatment effect estimators on continuous outcomes. Dose-response functions (DRF) were

estimated through weighting on the inverse of the GPS or using stratification. Variance estimators were evaluated using Monte Carlo simulations. Despite the use of stabilized weights, the variability of the weighted estimator of the DRF was particularly high, and none of the variance estimators were able to adequately capture this variability, resulting in coverages below the nominal value, particularly when the proportion of the variation in the quantitative exposure explained by the covariates was large. The stratified estimator was more stable, and variance estimators more efficient at capturing the empirical variability of the parameters of the DRF. The pooled linearized estimator and the pooled model-based estimator tended to overestimate the variance, whereas the bootstrap-based estimator, which intrinsically takes into account the estimation step of the GPS, resulted in correct variance estimations and coverage rates. These methods were applied to a real data set with the aim of assessing the effect of maternal body mass index on newborn birth weight.

Keywords. Generalized propensity score, observational study, quantitative exposure, variance estimator

1 Introduction

In observational cohort studies, confounding may occur when the distribution of baseline covariates differs between treated and control subjects. The propensity score is one of the methods that helps in reducing or minimizing this confounding to get valid inferences on treatment effects. It was first developed for binary or categorical exposures. In many studies, the exposure of interest is continuous rather than binary. For example, we may not only know whether an individual is a smoker or not, but also the pack-years of cigarettes smoked, or the duration of smoking. Considering this type of exposure variable, one may be interested in estimating the dose-response function (DRF). If this term may evoke the dose of a medication, we will use it regardless of the nature of the exposure as long as it is quantitative. The propensity score has been generalized into a propensity function for quantitative exposures which is known as the generalized propensity score (GPS). Different propensity score methods have been proposed to estimate the treatment effects on outcomes using the GPS: covariate adjustment, stratification (Zhang et al (2016)) and inverse probability of treatment weighting (IPTW) (Austin (2018)).

In the case of binary exposure, several authors have proposed valid closed-form variance estimators adapted to each treatment effect estimators. Note that all these estimators take into account the fact that the theoretical propensity score value of an individual is unknown, and is estimated from the data in the first-stage of analysis. To our knowledge, variance estimation for treatment effect estimated using the GPS framework has received little attention. In this work, we develop and evaluate closed-form variance estimators for

stratified and weighted treatment effect estimators using the influence function linearization technique. These variance estimators are also compared to bootstrap-based variance estimators.

2 Treatment effect estimator using the generalized propensity score

Notations and assumptions. Let T denote the level of a quantitative exposure which is a continuous variable, and Z a set of K baseline measured covariates. Let $Y(t),\ t \in \Psi$, denote a set of potential outcomes which is assumed to exist under Rubin's framework for causal inference. More precisely, we assume that T is a continuous exposure (i.e., Ψ is a subset of $\mathbb{R}$) and that $Y_i(t)$ is the outcome that would be observed for subject i if he/she received (maybe contrary to the reality) the level of exposure $T = t$. In practice, we only observe one level of exposure for each subject i and the corresponding outcome. The observed data consists of (Z_i, T_i, Y_i) for subjects $i = 1, \ldots, n$. We are interested in estimating the dose-response function

$$\mu(t) \quad = \quad \mathbb{E}[Y_i(t)], \tag{1}$$

which corresponds to the average response if all subjects were exposed to the level $T = t$. In this work, we only assume that $Y(t)$ is independent of T given Z, which is known as the weak unconfoundedness assumption and denoted as $Y(t) \perp\!\!\!\perp T|Z, \quad \forall t \in \Psi$. This assumption means that any association between the actual exposure and the potential outcomes is explained by a set of baseline covariates Z. Let us denote by

$$r(t \mid z) \quad \equiv \quad f_{T|Z}(t|z), \tag{2}$$

the conditional density of exposure variable T given the covariates, which is called the generalized propensity score (GPS). We make the positivity assumption, namely $r(t \mid z) > 0$ for any $t \in \Psi$ and for any z. We focus on two approaches for the estimation of the dose-response function: inverse probability of treatment weighting, and stratification.

Weighted treatment effect estimator. The first estimator is obtained by fitting a generalized linear regression model between the dose-response function and the exposure, used as the sole dependent variable. We considered a linear model for the dose response function, namely

$$\mu(T) \quad = \quad \beta_0 + \beta_1\ T, \tag{3}$$

where β_0 is the average response observed in case of null exposure ($T = 0$), and β_1 is the average response change if the level of exposure is increased by one unit. The parameter

$\beta = (\beta_0, \beta_1)^\top$ is estimated by weighted least squares. This leads to the first estimator $\hat{\mu}_w(t) = \hat{\beta}_{w0} + \hat{\beta}_{w1}\ t$ for the dose-response function. The weights used in the weighted least squares are the estimations of the theoretical Generalized Propensity Score (GPS) weights, defined as

$$w_i(\gamma) \quad = \quad \frac{W(T_i|\gamma)}{r(T_i|Z_i,\gamma)}, \tag{4}$$

where $r(t|z,\gamma)$ is the conditional density of the exposure variable defined in (2), and where $W(\cdot|\gamma)$ is a stabilization factor. As is currently done in the literature, we use $W(t|\gamma) \equiv f_T(t|\gamma)$ the marginal density of the exposure variable. Note that the weights depend on some unknown vector of parameters γ, which needs to be estimated.
We suppose that T_i follows a normal distribution, both conditionally on Z_i and non conditionally. We have $\gamma = (\mu_T, \sigma_T^2, \alpha^\top, \sigma^2)^\top$ with μ_T and σ_T^2 the parameters associated with the marginal distribution of T, and α and σ^2 the parameters associated with the conditional distribution $T|Z$. The parameters α and σ^2 are estimated by fitting a linear regression model between the exposure variable and the covariates, namely

$$T_i \quad = \quad \tilde{Z_i}^\top \alpha + \eta_i, \tag{5}$$

This leads to the estimator $\hat{\gamma} = (\hat{\mu}_T, \hat{\sigma}_T^2, \hat{\alpha}^\top, \hat{\sigma}^2)^\top$. By plugging this estimator in (4), we obtain the estimated weights $\hat{w}_i \equiv w_i(\hat{\gamma})$. The model (5) is called the propensity model.

Stratified treatment effect estimator. An alternative approach consists in partitioning the sample into L strata, in such a way that the units inside a given stratum are somewhat similar with respect to the covariates. Inside any stratum $l = 1, \ldots, L$, we fit the regression model $Y(T) = \beta_{l0} + \beta_{l1}\ T + \epsilon_l$ and by estimating the parameter $\beta_l = (\beta_{l0}, \beta_{l1})^\top$ by ordinary least squares , we obtain

$$\hat{\beta}_l \quad \equiv \quad (\hat{\beta}_{l0}, \hat{\beta}_{l1})^\top = \left(\sum_{i \in S_l} \tilde{T}_i \tilde{T}_i^\top \right)^{-1} \left(\sum_{i \in S_l} \tilde{T}_i Y_i \right), \tag{6}$$

with S_l the subset of sampled units which belong to the stratum l. The stratified estimator of the parameter β in (3) is obtained by pooling these L estimators, which leads to $\widehat{\beta}_{st} \equiv (\hat{\beta}_{st0}, \hat{\beta}_{st1})^\top = \sum_{l=1}^{L} \frac{n_l}{n} \hat{\beta}_l$ with n_l the number of sampled units in the stratum S_l. This leads to the second estimator of the dose-response function $\hat{\mu}_{st}(t) = \hat{\beta}_{st0} + \hat{\beta}_{st1}\ t$ for the dose-response function.

Our objective is to develop closed-form variance estimators for the estimators of the dose-response function. Without loss of generality, we focus on variance estimation for the estimated coefficients of regression $\hat{\beta}_w$ and $\widehat{\beta}_{st}$. We follow the influence function linearization technique developped by Deville (1999). For an estimator $\hat{\beta}$, this technique

consists in finding a so-called estimated linearized variable $\hat{I}_i$, summarizing the variability in the estimation of the parameter. The derivation of the linearized variance estimators is detailed in Garès et al. (2021).

3 Simulations

Data-generating process. First, we randomly generate $K+1$ normally distributed covariates $Z_1, \ldots, Z_k \ldots, Z_K, Z_U$ from the following multivariate normal distribution. Thus, $Z_1, \ldots, Z_K$ are mutually independent following a standard normal distribution, but are each correlated to a standard normal variable Z_U through covariance parameters σ_k, $k = 1, \ldots, K$. The treatment allocation T is drawn from a linear model where $T = \alpha_0 + \sum_{k=1}^{K} \alpha_k Z_k + \eta$ with $\eta \sim \mathcal{N}(0, \sigma^2)$. The parameter σ^2 is linked to the coefficient of determination R^2 which measures the proportion of the variance (of the exposure) explained by the regression model. The continuous outcome is then generated from Z_U as $Y = \beta_0 + \beta_1 T + \sigma_Y Z_U$ and therefore $Y \sim \mathcal{N}(\mu, \sigma_Y^2)$ where $\mu = \beta_0 + \beta_1 T$.

Estimation of the parameters of the dose-response function and their variance. A total of $B = 1000$ datasets were generated. All statistical methods estimating the dose-response function described in Section 2 were applied to each simulated dataset, and compared to the naive (unweighted) maximum likelihood estimator. Three different variance estimators were associated with the weighted estimator of the dose-response function: the sandwich variance estimator (Zhang et al (2016)), the proposed linearized variance estimator and a bootstrap variance estimator. The stratified estimator of the dose-response function was used with strata defined according to the deciles of the linear predictors of the propensity model. We also considered three variance estimators: the pooled linearized variance estimator, the pooled variance estimator using model-based variance estimator from the maximum likelihood estimator in each stratum, and the boostrap-based variance estimator. The bootstrap variance estimators were based on $N_{\text{boot}} = 200$ bootstrap samples and the propensity model and the parameters of the dose-response function were reestimated within each bootstrap samples.

Results. Coverage rates for dose-response function estimates are reported on Figure 1 for different values of parameters R^2 and σ_Y^2. Despite the use of stabilized weights, the variability of the weighted estimator of the DRF was particularly high, and none of the variance estimators were able to adequately capture this variability, resulting in coverages below the nominal value, particularly when the proportion of the variation in the quantitative exposure explained by the covariates was large. The stratified estimator was more stable, and variance estimators more efficient at capturing the empirical variability of the parameters of the DRF. The pooled linearized estimator and the pooled model-based estimator tended to overestimate the variance, whereas the bootstrap-based estimator,

which intrinsically takes into account the estimation step of the GPS, resulted in correct variance estimations and coverage rates.

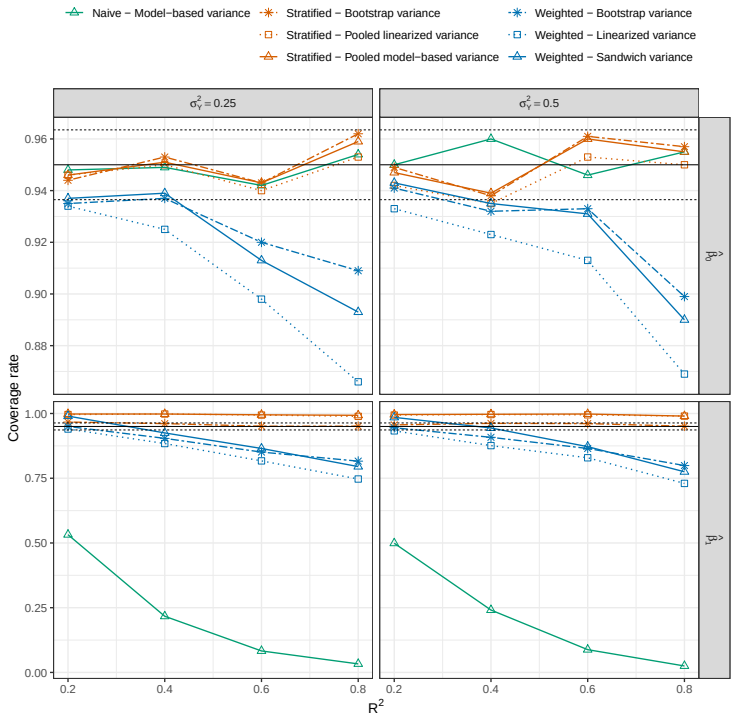


Figure 1: Coverage rate for different values of $R^2 \in \{0.2, 0.4, 0.6, 0.8\}$ and $\sigma^2_Y \in \{0.25, 0.5\}$, $\beta_0 = 0$, $\beta_1 = 1$ and $n = 1000$.

Bibliography

Austin, P. C. (2018). Assessing the performance of the generalized propensity score for estimating the effect of quantitative or continuous exposures on binary outcomes, *Statistics in medicine*, 37(11):1874–1894.

Deville, J. C. (1999). Variance Estimation for Complex Statistics and Estimators: Linearization and Residual Techniques, *Survey methodology*, 25(2):193–204.

Garès, V., Chauvet, G., Hajage, D. (2021). Variance estimators for weighted and stratified linear dose–response function estimators using generalized propensity score, *Biomedical Journal*.

Zhang Z., Zhou J., Cao W., and Zhang J.. (2016). Causal inference with a quantitative exposure, *Statistical methods in medical research*, 25(1):315–335.

Topological period counting method for reparametrized functions

Wojciech Reise [1] & Thomas Bonis [2] & Frédéric Chazal [1] & Bertrand Michel [3]

[1] *LMO, 307 rue Michel Magat, 91400 Orsay, wojciech.reise@inria.fr, frederic.chazal@inria.fr*
[2] *LAMA, 5 boulevard Descartes, 77420 Champs-sur-Marne, thomas.bonis@univ-eiffel.fr*
[3] *École Centrale Nantes, 1 Rue de la Noë, 44300 Nantes, bertrand.michel@ec-nantes.fr*

Résumé. On considère un signal constitué de plusieurs périodes d'une fonction périodique. On observe une reparamétrisation bruitée de ce signal et on cherche à retrouver le nombre de périodes observées. C'est une version plus facile du problème d'estimation de phase, traitée souvent quand la fonction est simple ou connue. On s'intéresse ici à des estimateurs basés sur la quantification de la forme du signal et au cas où la fonction reparamétrisée est inconnue. On étudie l'homologie persistante du signal qui synthétise l'évolution de ses sous–niveaux et on propose des estimateurs du nombre de périodes. Ceux-ci ne nécessitent pas la connaissance de cette fonction et sont adaptés à une classe de signaux génériques.

Mots-clés. Analyse topologique des données, Traitement du signal, Estimation de phase.

Abstract. We consider a signal composed of several periods of a function. We observe its reparametrisation corrupted by noise and we are interested in recovering the number of observed periods. It is a version of the phase estimation problem, often stated for functions which are simple or known. Interested in the case when the underlying function is unknown, we concentrate on estimators which quantify the shape of the function. We study the persistent homology of the signal, which synthesizes the evolution of its sublevel sets and we introduce estimators of the number of observed periods. Those estimators do not assume prior knowledge of the reparametrized function and can be used on a class of generic signals.

Keywords. Topological data analysis, Signal processings, Phase estimation.

1 Model

Let $f : \mathbb{R} \to \mathbb{R}$ be a 1–periodic, continuous function and $\gamma : [0, 1] \to [0, N]$ a continuous and increasing bijection for some $N \in \mathbb{N}$. Consider a noisy reparametrisation of N periods of f, namely

$$S_t = (f \circ \gamma)(t) + W_t, \tag{1}$$

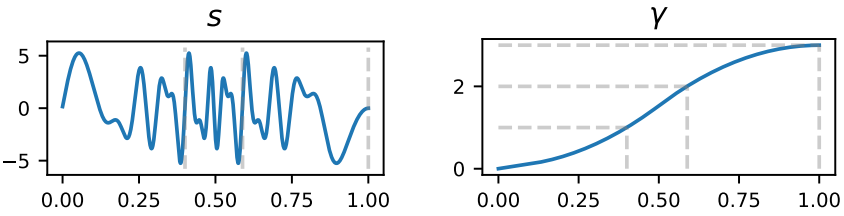


Figure 1: On the right, $f \circ \gamma$ for an unknown function f. The dashed vertical lines mark $\gamma^{-1}(\mathbb{N})$ and help recognise the reparametrised periods of f. On the left, γ, with $\gamma(1) = 3$.

where W_t is a continuous (stochastic) process. Figure 1 illustrates the model, with a function f featuring 10 local extrema per period. We are interested in inference on γ, without knowing f. In the scope of this communication, we focus on the estimation of N.

When f is a (linear combination of) phase–shifted sine function(s), inference on γ is known as the phase estimation problem in the signal processing community (Boashash, 2015). The Hilbert Transform of the signal yields a classic estimator of γ. Alternatively, a frequency decomposition of the signal allows one to extract the instantaneous frequency in S, a quantity related to γ' (Colominas et al., 2020). In Daubechies et al. (2009), the authors establish a connection between the wavelet transform, a technique related to spectral decompositions, with the Empirical Mode Decomposition, which extracts interpretable components from the signal.

The zero–crossings method extracts phase information from the shape of S in a simple way. It consists of counting the number of times that the graph of S crosses a certain level. Whenever f has a simple structure and does not have local extrema close to the chosen level, this method can be used to construct an interpretable estimator of γ.

Topological methods have been used to quantify the recurrent structure in dynamical systems or functions (Perea, 2019). Various methods have been experimentally shown to successfully detect recurrence in the signal and even construct a parametrization of the phase space (de Silva et al., 2012), providing effective dimensionality reduction and visualization techniques. However, we are not aware of their use for estimation of γ. An algorithm proposed in Khasawneh and Munch (2018) and based on topological descriptors counts the number of jumps in a piece-wise constant signal. It outperforms spectral approaches on the task of counting revolutions performed by a machine.

In the scope of this communication, we introduce an estimator of N, which relies on a topological descriptor of the input signal. We explain how it is related to the zero–crossings method and we show conditions under which the estimator is correct. Finally, we show experimental results, comparing it to a benchmark constructed from the zero-crossings method.

2 Topological estimator of N

Consider the case $W = 0$ in (1) and let k be the number of times that f crosses level α in one period. If k is known, then the crossings method gives an estimator of N

$$N_\alpha(S) = \frac{|S^{-1}(\alpha)|}{2k}. \tag{2}$$

We notice that the number of crossings $|S^{-1}(\alpha)|$ has a topological interpretation. It is the number of connected components in the sublevel set $S^{-1}(]-\infty, \alpha])$, plus or minus 1.

As α varies from $-\infty$ to ∞, the number of connected components in $S^{-1}(]-\infty, \alpha])$ changes. The persistence diagram $D(S)$ associated to the sublevel sets of S is a multi–set of points in $\mathbb{R}^2$ which summarizes these changes (Boissonnat et al., 2018). Each point $(b, d) \in D(S)$ encodes the birth of a connected component at level $\alpha = b$ and, at $\alpha = d$, the merging of that connected component into another component born no later than b. An example is shown in Figure 2. The diagram is seen as a sum of point masses centered at the points in the multi–set.

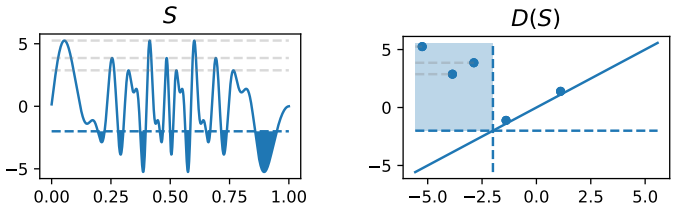


Figure 2: On the left, the graph of S with points (x, y) colored, where $S(x) \leq y \leq \alpha$. The persistence diagram of S from Fig 1. Points (b, d) that fall in the shaded region correspond to connected components born before α ($b \leq \alpha$) and die after α ($d \geq \alpha$).

For our purposes, we slightly modify the construction of the persistence diagram, by considering the signal as defined on a circle, instead of the interval $[0, 1]$. This allows us to account for the boundary effects and yields the homogeneity property, which motivates the introduction of an estimator of N based on the persistence diagram.

Proposition 1 *For $N \in \mathbb{N}$, we consider f_N which consists of N periods of f*

$$\begin{array}{rccc} f_N : & [0,1] & \to & \mathbb{R} \\ & t & \mapsto & f(Nt). \end{array}$$

Then,

$$D(f_N) = ND(f_1). \tag{3}$$

Let $D(f)(A)$ be the number of points in a measurable set A. We define $N(f)$ as the greatest common divisor of the multiplicities of points in the diagram $D(f)$

$$\hat{N}(f) = \gcd\{D(f)(\{p\}) \mid p \in D(f)\}. \tag{4}$$

Proposition 1 implies that $\hat{N}(f_N) = N \cdot \hat{N}(f_1)$. If we assume that $\hat{N}(f_1) = 1$, $\hat{N}(f_N)$ recovers N. This is a reasonable assumption in terms of the persistence diagram. Indeed, it is the weakest hypothesis that allows us to ensure the problem of estimation of N is well–posed.

We notice that (3) is similar to the homogeneity of the crossings: $|S^{-1}(\alpha)| = Nk$. Interestingly, the number crossings of the level α can be read in $D(S)$ as the sum of multiplicities of points (b, d) such that $b < \alpha$ and $\alpha < d$,

$$|S^{-1}(\alpha)| = D(S)(]-\infty, \alpha] \times]\alpha, \infty[).$$

Let us now consider the case when S is noisy, $W \neq 0$. When the noise is a stochastic process with density, $D(S)$ almost-surely does not contain points with multiplicity greater than one. Also, $D(S)$ might contain new points which correspond to local extrema generated by the noise. However, by stability of the persistence diagram, the points remain in $\|W\|_\infty$-balls around their counterparts in $D(f_N)$ and the diagonal $\{(x, x) \mid x \in \mathbb{R}\}$. We therefore introduce a new estimator, which discards points near the diagonal and which counts points in certain neighborhoods, instead of their multiplicities. Let $\tau > 0$ be a neighborhood size parameter and denote by $\Delta_\tau = \{(b, d) \in \mathbb{R}^2 \mid 0 \leq d - b \leq 2\tau\}$ the set of points which are above the diagonal, but not further from it than τ. We define $\hat{N}(S, \tau)$ as the greatest common divisor of cardinalities of τ-neighborhoods in $D(S) \setminus \Delta_\tau$,

$$\hat{N}(S, \tau) = \gcd\{D(S)(B(p, \tau) \setminus \Delta_\tau) \mid p \in D(S) \setminus \Delta_\tau\}. \tag{5}$$

As long as the noise W remains small, compared to the topological signature of f, $\hat{N}(S, \tau) = N \cdot \hat{N}(f_1)$.

Theorem 1 *Let δ_f be the minimum between pairwise distances of points in $D(f)$ and their distances to the diagonal $\{(x, x) \mid x \in \mathbb{R}\}$. Suppose that $(W_t)_{t\in[0,1]}$ is a Gaussian stochastic process with covariance function $\Gamma(t) = \sigma^2 \exp\left(-\frac{t^2}{2l^2}\right)$. Then, for any $\tau \in]0, \delta_f/3[$, we have*

$$\hat{N}(S, \tau) = N \cdot \hat{N}(f_1, \tau) = N \cdot \hat{N}(f_1),$$

with probability at least $1 - \left(\frac{1}{l^2\pi} \exp(\frac{-\kappa^2}{2}) + 2\phi(-\kappa)\right)$, where $\kappa = \frac{\tau}{2\sigma}$ and ϕ is the cumulative distribution function of a normal random variable.

Theorem 1 quantifies the probability of estimating N correctly, in terms of τ. As we would expect, the lower bound increases in τ, but only until $\tau = \delta_f/3$. The persistence diagram captures the periodic structures and τ defines the scale at which different extrema are identified as realizations of the same extremum. Hence, increasing τ past $\delta_f/3$ confoundes the topological signature of f and will yield an incorrect estimate of N.

3 Practical adaptations and numerical results

We assume no prior knowledge about f and, in practice, we do not know δ_f either. However, we know that $\delta_f \leq A_S$, with $A_S \coloneqq \frac{3}{5}(\max S - \min S)$. We introduce $\hat{N}^B$, which is the mode of $\hat{N}(S, \tau)$, for several values of τ sampled in $]0, A_S/3]$.

Experimentally, we notice points in perturbed diagrams form non-overlapping clusters. Their diameters are too big to be counted correctly with $\hat{N}(\cdot, \tau)$, for any choice of τ, but can be often captured with a single–linkage hierarchical clustering algorithm, initialized at scale τ. Hence, we define $\hat{N}_C$ by counting points in the resulting clusters. We also construct the analogue estimator $\hat{N}_C^B$, which does not require the choice of a particular threshold τ.

We evaluate the proposed estimators on a synthetic dataset. Depicted in Figure 3 are five template functions for which $\hat{N}(f_1) = 1$. We generate 100 reparametrizations γ by sampling N uniformly from $[\![5, 50]\!]$. This defines a collection of 500 signals, which we perturb with independent realizations of Gaussian noise $(W_t)_{t\in[0,1]}$ with covariance as in Theorem 1 and $\sigma = \mathrm{SNR} \in [0.0001, 6]$. We compare our estimators to the *zero-crossings oracle* defined in (2), where we provide the method with the correct k. The results are shown in the graphs on the right in Figure 3.

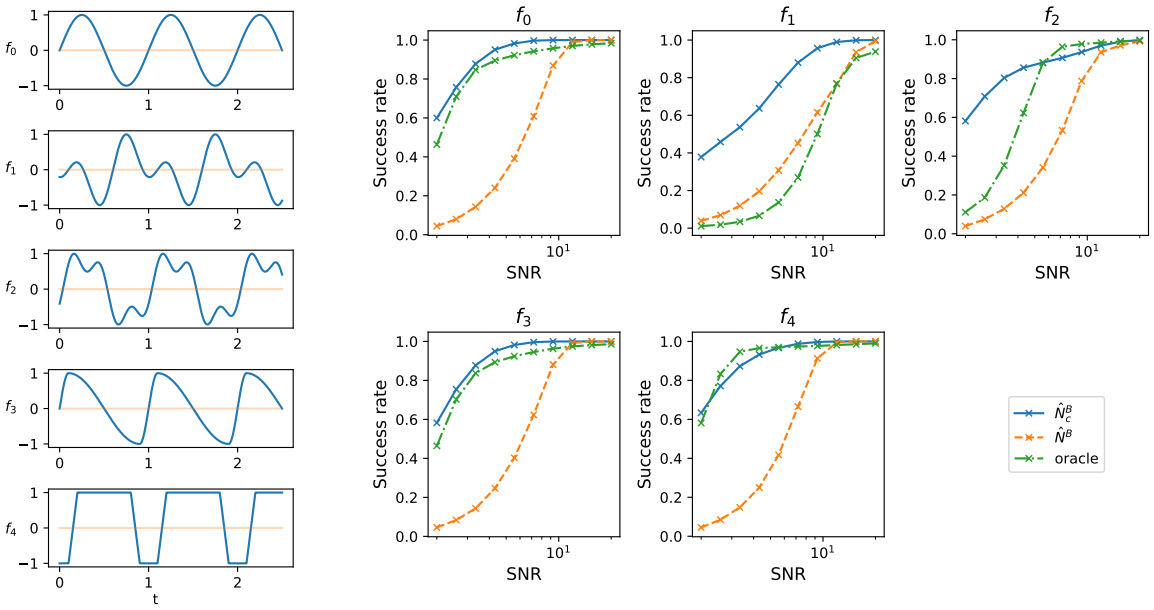


Figure 3: On the left, the graphs of the template functions $(f_i)_{i=0}^4$ on $[0, 2.5]$. On the right, the fraction of samples for which N was correctly estimated, for different template functions f_i.

The success rate of $\hat{N}_C^B$ is similar to that of the oracle, for f_0, f_3, f_4, while that of

$\hat{N}^B$ is much lower. The graph of f_1 crosses $y = 0$ four times in a period, but two of the extremal values of f_1 are near 0. Hence, the noise is likely to induce spurious zero–crossings to which $\hat{N}_C^B$ is less sensitive. On the other hand, a reparametrisation of f_4 corrupted with noise will have many local extrema of on the segments $f_4^{-1}(\{-1\} \cup \{1\})$. The zero–crossings method will be insensitive to those extrema as long as they are far from zero. Meanwhile, $\hat{N}^B$ and $\hat{N}_C^B$ are sensitive to any additional points in the diagram, especially for small τ.

In this communication, we have focused on the problem of estimating $\gamma(1) = N$. The estimators we propose are an integral part of a procedure to construct a finer estimator of γ. Still, in this part of our work, the theoretical guarantees do not reflect the true performance of the estimators. Improving those guarantees should be followed by extending them to $\hat{N}^B$ and $\hat{N}_C^B$.

References

Boashash, B. (2015). *Time-Frequency Signal Analysis and Processing : A Comprehensive Reference Ed. 2*. Elsevier Science.

Boissonnat, J.-D., Chazal, F., and Yvinec, M. (2018). *Geometric and Topological Inference*. Cambridge University Press.

Colominas, M. A., Meignen, S., and Pham, D.-H. (2020). Fully Adaptive Ridge Detection Based on STFT Phase Information. *IEEE Signal Processing Letters*, 27:620–624.

Daubechies, I., Lu, J., and Wu, H.-T. (2009). Synchrosqueezed Wavelet Transforms: a Tool for Empirical Mode Decomposition. *arXiv:0912.2437 [math]*. arXiv: 0912.2437.

de Silva, V., Skraba, P., and Vejdemo-Johansson, M. (2012). Topological Analysis of Recurrent Systems. page 5.

Khasawneh, F. A. and Munch, E. (2018). Topological data analysis for true step detection in periodic piecewise constant signals. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 474(2218):20180027. Publisher: Royal Society.

Perea, J. A. (2019). Topological Time Series Analysis. *Notices of the American Mathematical Society*, 66(05):1.

Extrêmes & Prediction

Extremal cluster inference for heavy-tailed time series

Gloria Buriticá[1], Thomas Mikosch[2] & Olivier Wintenberger[1]

[1] *LPSM, Sorbonne Université, 4 place Jussieu, 75005 Paris, France*
[2] *University of Copenhagen, Universitetsparken 5, DK-2100 Copenhagen, Denmark*

Résumé. Nous étudions des séries chronologiques à variation régulière. Dans ce contexte, les valeurs extrêmes peuvent entrainer des observations consécutives extremales. Nous nous référons aux épisodes extrémaux de norme ℓ^p large, comme des p-clusters. Pour évaluer cet effet de clustering, nous considérons des fonctionnelles agissant sur les p-clusters. Par exemple, l'indice extrémal est une statistique que nous pouvons obtenir de cette manière. L'inverse de l'indice extrémal est interprétée comme la moyenne des grandes valeurs simultanées dans un cluster. Nous examinons le cadre d'inférence de p-cluster dans Buriticá et al. (2021b), basé sur les ℓ^p–blocs extrémaux. Nous comparons deux estimateurs de l'indice extrémal, basés respectivement sur les ℓ^α, ℓ^∞-blocs extrémaux, où $\alpha > 0$ est l'indice de queue de la série. Des expériences numériques à partir du modèle auto-régressif d'ordre un montrent que l'approche avec $p = \alpha$ est plus robuste pour traiter les dépendances temporelles que la stratégie $p = \infty$.

Abstract. We study regularly varying time series. In this setting, extremes can appear as consecutive large observations. We refer to extreme episodes with large ℓ^p-norm, as p-clusters. To evaluate the clustering effect, we consider functionals acting on p-clusters. A summary statistic that we can recover in this manner is the extremal index. Its reciprocate is interpreted as the mean simultaneous large values in one cluster. We review the p-cluster inference setting in Buriticá et al. (2021b) based on extremal ℓ^p–blocks. We compare two estimators of the extremal index, based on extremal ℓ^α,ℓ^∞–blocks, respectively, where $\alpha > 0$ is the tail index of the series. Numerical experiments from the auto-regressive model of order one show the approach with $p = \alpha$ is more robust to handle time dependencies than the $p = \infty$ strategy.

Keywords: Regular variation; the extremal index; cluster processes; cluster inference.

1 Introduction

In time series analysis of extremes, statisticians commonly need to assess time dependencies to improve accuracy Embrechts et al. (2013). This is the case, for examples, with environmental data analysis including rainfall records, or wind speeds, or sea-levels. For these applications, assessing the risk of extreme events is highly important from a societal and economic point of view. Typically, the practitioner aim to predict when the next catastrophic event will happen to prevent its damage. The physical dynamics of weather

cause temporal dependencies that often traduce into a concatenation of heavy records. In this setting, it is well known that time dependencies can perturb classical statistical procedures. These examples motivates the study of extremal clustering; see Leadbetter et al. (1983).

We consider regularly varying time series $(\mathbf{X}_t)$, taking values in $(\mathbb{R}^d, |\cdot|)$, with tail index $\alpha > 0$. We aim to study the extreme behavior of the series in time. In this setting, typically recording a significant value entails a period of extreme behavior lasting for numerous time units. To model these extreme episodes, we define $p-$clusters as short periods of consecutive observations, or blocks, with large ℓ^p-norm, for $p > 0$; (see e.g. Buriticá et al., 2021b). Recall for $(\mathbf{x}_t) \in (\mathbb{R}^d)^{\mathbb{Z}}$, ℓ^p-norms are defined as $\|(\mathbf{x}_t)\|_p = (\sum_{t\in\mathbb{Z}} |\mathbf{x}_t|^p)^{1/p}$, for $p \in (0, \infty)$, and the as supremum norm for $p = \infty$. Moreover, we focus on p-cluster inference as we aim to summarize the shape of extremal ℓ^p–blocks. Our goal is to infer the asymptotic limit of certain functionals $g_p : \ell^p \to \mathbb{R}$, acting on extremal ℓ^p–blocks, as their ℓ^p-norm grows. For example, the extremal index can be recovered in this way using different g_p, p pairs. The extremal index is a classical measure of temporal dependence summarizing the mean counts of pick values triggered by one first large recording (see e.g. Buriticá et al., 2021a).

We review the p-cluster inference theory in Buriticá et al. (2021b) and compare estimation of the extremal index based on extremal ℓ^α-blocks and extremal ℓ^∞-blocks, where α is the tail index of the regularly varying series. Our numerical experiments support that inference based on extremal ℓ^α–blocks is a nice competitor to inference based on extremal ℓ^∞-blocks. Typically, cluster inference has focused solely on methods for extremal blocks with at least one exceedance, i.e., large ℓ^∞-norm (see e.g. Kulik and Soulier, 2020).

1.1 Assumptions

Consider an $\mathbb{R}^d$-valued stationary time series $(\mathbf{X}_t)$ with values in $(\mathbb{R}^d, |\cdot|)$. Assume it admits an extremal index $\theta_{|\mathbf{X}|}$, as defined in Leadbetter et al. (1983), i.e., for all $\tau \in (0, \infty)$, there exists a sequence $(a_n(\tau))$ such that $n\mathbb{P}(|\mathbf{X}_0| > a_n(\tau)) \to \tau$, as $n \to +\infty$, and

$$\mathbb{P}(\|(\mathbf{X}_t)_{t=1,\dots,n}\|_\infty \le a_n(\tau)) = e^{-\tau\theta_{|\mathbf{X}|}}. \tag{1}$$

In addition, we assume the conditions $\mathbf{RV}_\alpha, \mathbf{AC}, \mathbf{CS}_\alpha$ defined below. The first condition is regular variation and models the heaviness of tails. The two last are needed to control the impact of a heavy records on future events. They are tailored for maxima of blocks and α-sums, respectively. To define the extremal index, it is common to prevent the long range dependence of extremes. We state the aforementioned conditions next.

$\mathbf{RV}_\alpha$: Assume all finite cuts of $(\mathbf{X}_t)$ are regularly varying, i.e., for all $h \ge 0$ there exists $(\mathbf{\Theta}_t)_{|t|=0,\dots,h} \in (\mathbb{R}^d)^{2h+1}$ such that

$$\lim_{x\to+\infty} \mathbb{P}(x^{-1}(\mathbf{X}_t)_{|t|=0,\dots,h} \in \cdot \,|\, |\mathbf{X}_0| > x) = \mathbb{P}(Y(\mathbf{\Theta}_t)_{|t|=0,\dots,h} \in \cdot). \tag{2}$$

such that Y is independent of $(\mathbf{\Theta}_t)_{|t|=0,\dots,h}$, and $\mathbb{P}(Y > y) = y^{-\alpha}$, for $y > 1$.

Assume also there exist a sequence (x_n) such that $n\,\mathbb{P}(|\mathbf{X}_0| > x_n) \to 0$, and some $\kappa > 0$ such that $n/x_n^{\alpha-\kappa} \to 0$, as $n \to +\infty$, and for all $\delta > 0$,

AC : Assume the anti-clustering condition:

$$\lim_{k\to+\infty} \limsup_{n\to+\infty} \mathbb{P}\big(\|(\mathbf{X}_t)_{t=k,\dots,n}\|_\infty > \delta\, x_n \mid |\mathbf{X}_0| > \delta\, x_n\big) \;=\; 0,$$

CS$_\alpha$: Assume the vanishing small values condition:

$$\lim_{\epsilon\to 0} \limsup_{n\to\infty} \frac{\mathbb{P}(\|(\,\mathbf{X}_t \mathbb{1}(|\mathbf{X}_t| \le \epsilon\, x_n)\,)_{t=0,\dots,n}\|_\alpha > \delta\, x_n)}{n\,\mathbb{P}(|\mathbf{X}_0| > x_n)} \;=\; 0.$$

The time series $(\mathbf{\Theta}_t)$ is known as the *spectral tail process* and was introduced in Basrak and Segers (2009) as a tool for modeling multivariate regular variation of finite vectors of the time series; see Theorem 2.1 in Basrak and Segers (2009). It satisfies the time-change formula: for any $s \le 0 \le t$ and for any measurable bounded function $f : (\mathbb{R}^d)^{t-s+1} \to \mathbb{R}$,

$$\mathbb{E}[f(\mathbf{\Theta}_{s-i},\dots,\mathbf{\Theta}_{t-i})\mathbb{1}(|\mathbf{\Theta}_{-i}| \neq 0)] \;=\; \mathbb{E}[|\mathbf{\Theta}_i|^\alpha\, f(\mathbf{\Theta}_s/|\mathbf{\Theta}_i|,\dots,\mathbf{\Theta}_t/|\mathbf{\Theta}_i|)]. \tag{3}$$

1.2 Cluster processes

We recall the large deviation result in Buriticá et al. (2021b) and reformulate below Proposition 2.1. In the remaining, the sequence (x_n) is fixed to satisfy **AC**, **CS**$_\alpha$.

Theorem 1. *Let $(\mathbf{X}_t)$ be a stationary time series admitting an extremal index $\theta_{|\mathbf{X}|}$, satisfying* **RV**$_\alpha$, **AC**, **CS**$_\alpha$. *Then, $\|\mathbf{\Theta}\|_\alpha < +\infty$ a.s., $\theta_{|\mathbf{X}|} = \mathbb{E}[\|\mathbf{\Theta}\|_\alpha^\alpha / \|\mathbf{\Theta}\|_\infty^\alpha]$ and*

$$\mathbb{P}(\|(\mathbf{X}_t)_{t=0,\dots,n}\|_\alpha > x_n) \sim\; n\mathbb{P}(|\mathbf{X}_0| > x_n), \tag{4}$$

$$\sim \theta_{|\mathbf{X}|}^{-1}\mathbb{P}(\|(\mathbf{X}_t)_{t=0,\dots,n}\|_\infty > x_n), \quad n \to +\infty. \tag{5}$$

Moreover, there exists a p-cluster processes $\mathbf{Q}^{(\alpha)} \in \ell^\alpha$ distributed as $\mathbf{\Theta}/\|\mathbf{\Theta}\|_\alpha$ such that

$$\mathbb{P}(x_n^{-1}(\mathbf{X}_t)_{t=0,\dots,n} \in \cdot \mid \|(\mathbf{X}_t)_{t=0,\dots,n}\|_\alpha > x_n) \to \mathbb{P}(Y\mathbf{Q}^{(\alpha)} \in \cdot), \tag{6}$$

and $\mathbf{Q}^{(\infty)} \in \ell^\infty$, such that $\|\mathbf{Q}^{(\infty)}\|_\infty = 1$ a.s., and

$$\mathbb{P}(x_n^{-1}(\mathbf{X}_t)_{t=0,\dots,n} \in \cdot \mid \|(\mathbf{X}_t)_{t=0,\dots,n}\|_\infty > x_n) \to \mathbb{P}(Y\mathbf{Q}^{(\infty)} \in \cdot), \tag{7}$$

Convergence holds in the space of shift-invariant sequences of ℓ^α, ℓ^∞, respectively, and Y is independent of $\mathbf{Q}^{(\alpha)}$ and $\mathbf{Q}^{(\infty)}$ with $\mathbb{P}(Y > y) = y^{-\alpha}$, for $y > 1$.

Moreover, cluster processes $\mathbf{Q}^{(\alpha)}, \mathbf{Q}^{(\infty)}$ are related in the following way by means of the time shift-formula (3). We recall Proposition 3.1. in Buriticá et al. (2021b).

Proposition 1. *Let $g : \ell^\infty \to \mathbb{R}$, be a continuous and bounded shift-invariant functional. Assume also it vanishes in a neighborhood of the origin. Then, $\theta_{|\mathbf{X}|} = \mathbb{E}[\|\mathbf{Q}^{(\alpha)}\|_\infty^\alpha]$ and*

$$\int_0^\infty \mathbb{E}\big[g(y\mathbf{Q}^{(\infty)})\big]d(-y^{-\alpha}) \;=\; \theta_{|\mathbf{X}|}^{-1}\int_0^\infty \mathbb{E}\big[\,\|\mathbf{Q}^{(\alpha)}\|_\infty^\alpha\, g\big(y\tfrac{\mathbf{Q}^{(\alpha)}}{\|\mathbf{Q}^{(\alpha)}\|_\infty}\big)\big]d(-y^{-\alpha}). \tag{8}$$

Equation (8) suggests that the same quantity, let's say $\theta_{|X|}$, can be obtained by choosing adequately functionals g_α, g_∞ to act on extremal ℓ^α, ℓ^∞ blocks, respectively. We review next this inference strategy in the extremal index setting. Developing the equation above we obtain

$$\begin{aligned}
&\big(\textstyle\int_0^\infty \mathbb{E}[\sum_{t\in\mathbb{Z}} \mathbb{1}(|y\mathbf{Q}_t^{(\infty)}| > 1)]d(-y^{-\alpha})\big)^{-1}\\
&= \;\theta_{|\mathbf{X}|}\,\big(\textstyle\int_0^\infty \mathbb{E}\big[\|\mathbf{Q}^{(\alpha)}\|_\infty^\alpha \sum_{t\in\mathbb{Z}} \mathbb{1}(\big|y\tfrac{\mathbf{Q}_t^{(\alpha)}}{\|\mathbf{Q}^{(\alpha)}\|_\infty}\big| > 1)\big]d(-y^{-\alpha})\big)^{-1}\\
&= \;\theta_{|\mathbf{X}|}\,\big(\mathbb{E}\big[\|\mathbf{Q}^{(\alpha)}\|_\infty^\alpha \textstyle\sum_{t\in\mathbb{Z}} \big|\tfrac{\mathbf{Q}_t^{(\alpha)}}{\|\mathbf{Q}^{(\alpha)}\|_\infty}\big|^\alpha\big]\big)^{-1}\\
&= \;\theta_{|\mathbf{X}|}\,\big(\mathbb{E}\big[\|\mathbf{Q}^{(\alpha)}\|_\infty^\alpha \tfrac{\|\mathbf{Q}^{(\alpha)}\|_\alpha^\alpha}{\|\mathbf{Q}^{(\alpha)}\|_\infty^\alpha}\big]\big)^{-1}\\
&= \;\theta_{|\mathbf{X}|} \;=\; \mathbb{E}[\|\mathbf{Q}^{(\alpha)}\|_\infty^\alpha].
\end{aligned}$$

The previous calculation points to new cluster inference procedures of the extremal index.

2 Extremal index cluster inference

For cluster inference, we start by dividing a sample $(\mathbf{X}_t)_{t=1,\dots,n}$ into disjoint time periods of length b_n. We define disjoint blocks $(\mathcal{B}_t)_{t=1,\dots,n/b_n}$ by

$$\underbrace{\mathbf{X}_1,\dots,\mathbf{X}_{b_n}}_{:=\mathcal{B}_1},\ \underbrace{\mathbf{X}_{b_n+1},\dots,\mathbf{X}_{2b_n}}_{:=\mathcal{B}_2},\ \dots,\ \underbrace{\mathbf{X}_{n-b_n+1},\dots,\mathbf{X}_n}_{:=\mathcal{B}_{n/b_n}}.$$

In view of Theorem 1, we can infer the extremal properties of clusters using extremal ℓ^α or ℓ^∞-blocks. We aim to compare both inference procedures.

We review the inference approach from Theorem 4.1. in Buriticá et al. (2021b). For a shift-invariant function $g_p : \ell^\alpha \to \mathbb{R}$, our goal is to estimate $\mathbb{E}[g_p(Y\mathbf{Q}^{(p)})]$ by $\widehat{g}_p$ defined as

$$\widehat{g}_p \;=\; \tfrac{1}{k}\textstyle\sum_{t=1}^m g(\mathcal{B}_t/\|\mathcal{B}\|_{p,(k+1)})\mathbb{1}(\|\mathcal{B}_t\|_p > \|\mathcal{B}\|_{p,(k+1)}),$$

where $\|\mathcal{B}\|_{p,(1)} \geq \|\mathcal{B}\|_{p,(2)} \geq \cdots \geq \|\mathcal{B}\|_{p,(m_n)}$ are order statistics of p-norms and $p \in \{\alpha,\infty\}$.

2.1 Estimators

In the extremal index setting, we can consider two functionals g_α, g_∞ defined by $g_\alpha : (\mathbf{x}_t) \mapsto \|(\mathbf{x}_t)\|_\infty^\alpha/\|(\mathbf{x}_t)\|_\alpha^\alpha\,\mathbb{1}(\|(\mathbf{x}_t)\|_\alpha > 1), g_\infty : (\mathbf{x}_t) \mapsto \sum_{t\in\mathbb{Z}}\mathbb{1}(|\mathbf{x}_t| > 1)\,\mathbb{1}(\|(\mathbf{x}_t)\|_\infty > 1)$. In

view of Theorem 1 and Proposition 1, we propose to infer the extremal index $\theta_{|\mathbf{X}|}$ letting g_α, g_∞ act on extremal ℓ^α and ℓ^∞-blocks, respectively. We obtain two estimators

$$\widehat{\theta}_{|\mathbf{X}|,\alpha} := \tfrac{1}{k}\sum_{t=1}^{m} \tfrac{\|\mathcal{B}_t\|_\infty^\alpha}{\|\mathcal{B}_t\|_\alpha^\alpha} \mathbb{1}(\|\mathcal{B}_t\|_\alpha > \|\mathcal{B}\|_{\alpha,(k+1)}), \tag{9}$$

$$\widehat{\theta}^2_{|\mathbf{X}|,\infty} := \big(\tfrac{1}{k}\sum_{t=1}^{n} \mathbb{1}(|\mathbf{X}_t| > \|\mathcal{B}\|_{\infty,(k+1)})\big)^{-1}. \tag{10}$$

Both estimators depend on a tuning parameter k selecting the number of extremal ℓ^α, ℓ^∞ blocks used for inference. Moreover, if (x_n) is as in Theorem 1, choosing $k = k_n$ carefully entails $\|\mathcal{B}_{\alpha,(k)}\|_\alpha / x_{b_n} \xrightarrow{\mathbb{P}} 1$, as $n \to +\infty$. For this $k = k(\alpha)$ must satisfy

$$k(\alpha) \sim \mathbb{P}(\|\mathbf{X}_{t=0,\dots,b_n}\|_\alpha > x_{b_n}) \sim n\mathbb{P}(|\mathbf{X}| > x_{b_n}). \tag{11}$$

Choosing $k = k(\alpha)$ in this way yields, as $n \to +\infty$,

$$\begin{aligned}
\mathbb{E}[\widehat{\theta}_{|\mathbf{X}|,\alpha}] &\sim k^{-1} m_n \mathbb{E}[\,\|\mathbf{X}_{t=0,\dots,b_n}\|_\infty^\alpha / \|\mathbf{X}_{t=0,\dots,b_n}\|_\alpha^\alpha \mathbb{1}(\|\mathbf{X}_{t=0,\dots,b_n}\|_\alpha > x_{b_n})] \\
&\sim k^{-1}\mathbb{P}(\|\mathbf{X}_{t=0,\dots,b_n}\|_\alpha > x_{b_n})\mathbb{E}[\|\mathbf{Q}^{(\alpha)}\|_\infty^\alpha] \qquad (12)\\
&\sim \mathbb{E}[\|\mathbf{Q}^{(\alpha)}\|_\infty^\alpha]. \qquad (13)
\end{aligned}$$

On the other hand, for (10) we deduce by similar calculations that $k = k(\infty)$ must satisfy

$$k(\infty) \sim \mathbb{P}(\|\mathbf{X}_{t=0,\dots,b_n}\|_\infty > x_{b_n}) \sim \theta_{|\mathbf{X}|} n\mathbb{P}(|\mathbf{X}| > x_{b_n}). \tag{14}$$

To sum up, the advantage of ℓ^α-cluster inference is that the choice of k_n does not depend on the clustering effect, as can be seen from Equation (11). Instead, we see from Equation (14) that for ℓ^∞ cluster-based inference the choice of k_n is shrunken by the clustering effect. In this case, fine-tuning k can be challenging.

2.2 Simulation

We illustrate the estimators in (9) and (10) for a AR(1) model (X_t) defined by

$$X_t = \varphi X_{t-1} + Z_t, \quad t \in \mathbb{Z},$$

where $|\varphi| < 1$, (Z_t) is an iid sequence of student(α) noise, and $\theta_{|X|} = 1 - |\varphi|^\alpha$. We tune both estimators (9), (10) with k chosen as n/b_n^2, for different block lengths $b_n = 2, 4, 8, 16, 32, 64$. To implement the estimator in (9) we plug-in an estimate $\widehat{\alpha}$ of the index of regular variation. We use the bias-correction procedure in De Haan et al. (2016).

2.3 Conclusions

We mentioned $\widehat{\theta}_{|X|,\alpha}$ in Equation (9) is a robust estimator since the choice of k does not depend on the temporal memories of the series. This can be seen from Equation (11). For comparison, we implement $\widehat{\theta}_{|X|,\widehat{\alpha}}$, $\widehat{\theta}_{|X|,\infty}$ following the details in Section 2.2. We can see in Figure 1 that the estimator $\widehat{\theta}_{|X|,\widehat{\alpha}}$, based on extremal ℓ^α–blocks, is a nice competitor against $\widehat{\theta}_{|X|,\infty}$. Choosing k as in Section 2.2 we can see that our estimator $\widehat{\theta}_{|X|,\widehat{\alpha}}$ is robust to the choice of block length, for example, $b = 64$ is a nice choice in all plots.

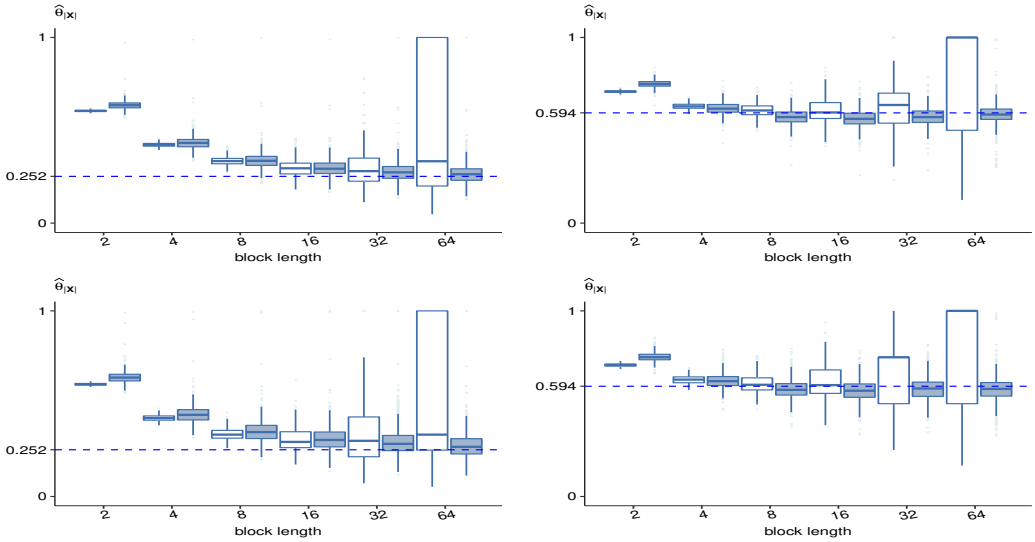


Figure 1: Boxplot of estimates $\widehat{\theta}_{|X|,\widehat{\alpha}}$ in (9), as a function of b_n (in blue) and $\widehat{\theta}_{|X|,\infty}$ in (10) (in white). We simulate 1 000 samples $(X_t)_{t=1,\dots,n}$ from a causal AR(1) model with student(α) noise, $\alpha = 1.3$ and $\varphi = 0.8$ (left column), $\varphi = 0.5$ (right column). In the first row we fix $n = 8\,000$, and in the second row we fix $n = 4\,000$.

References

Basrak, B., Segers, J., 2009. Regularly varying multivariate time series. Stochastic processes and their applications 119, 1055–1080.

Buriticá, G., Meyer, N., Mikosch, T., Wintenberger, O., 2021a. Some variations on the extremal index. arXiv:2106.12822 .

Buriticá, G., Mikosch, T., Wintenberger, O., 2021b. Threshold selection for cluster inference based on large deviation principles. arXiv:2106.12822 .

De Haan, L., Mercadier, C., Zhou, C., 2016. Adapting extreme value statistics to financial time series: dealing with bias and serial dependence. Finance and Stochastics 20, 321–354.

Embrechts, P., Klüppelberg, C., Mikosch, T., 2013. Modelling extremal events: for insurance and finance. volume 33. Springer Science & Business Media.

Kulik, R., Soulier, P., 2020. Heavy-Tailed Time Series. Springer, New York.

Leadbetter, M., Lindgren, G., Rootzén, H., 1983. Extremes and Related Properties of Random Sequences and Processes. Springer, Berlin.

Étude de la dépendance spatiale dans les extrêmes : un point de vue géostatistique.

Marine Demangeot [1] & Emilie Chautru [2] & Anne Sabourin [3]

[1] *Sorbonne Université, LPSM, 4 Place Jussieu, 75005 Paris*
marine.demangeot@sorbonne-universite.fr
[2] *MINES ParisTech - PSL university, centre de Géoscience, équipe Géostatistique*
35 Rue Saint-Honoré, 77300 Fontainebleau
emilie.chautru@mines-paristech.fr
[3] *Télécom Paris, centre IDS, équipe L2A, 9 place Marguerite Perey, 91120 Palaiseau*
anne.sabourin@telecom-paris.fr

Résumé. La fonction coefficient extrémal est une mesure bivariée de dépendance spatiale pour les processus stationnaires max-stables, *cf.* Schlather and Tawn (2003). Elle est généralement estimée à partir de données temporelles, lorsque le phénomène spatial étudié est observé plusieurs fois dans le temps (e.g. des pluies ou des températures extrêmes, de fortes concentrations de polluants dans l'air). Parfois, nous n'avons pas accès à de telles données : seulement un ou quelques enregistrements sont disponibles. C'est le cas, par exemple, des études sur l'estimation des ressources minières ou sur l'évaluation de la pollution des sols et plus généralement de toute recherche dont l'objet d'étude varie très peu au cours du temps ou pour lequel le coût d'échantillonnage est trop élevé. Ce cas de figure est très peu abordé par la communauté des extrêmes. Au contraire, c'est un cadre d'analyse auquel la Géostatistique s'intéresse particulièrement. Un des outils fondamentaux de cette discipline est le variogramme qui est aussi une mesure bivariée de dépendance spatiale. En considérant le variogramme des indicatrices, au dessus d'un certain seuil, d'un processus stationnaire max-stable, nous proposons un nouvel estimateur non paramétrique de la fonction coefficient extrémal basé sur l'estimateur de type Nadaraya-Watson de ce variogramme. Ce dernier a notamment été étudié par García-Soidán et al. (2004) and García-Soidán (2007) ; à partir de leurs travaux, nous établissons les propriétés asymptotiques de notre estimateur lorsqu'il est calculé à partir d'un unique jeu de données spatialisées. En particulier, sous certains conditions, nous montrons qu'il est consistant et asymptotiquement normal. Ces résultats sont illustrés sur des simulations numériques puis sur des données de précipitations maximales. Nous comparons également les performances de notre estimateur avec celle de l'estimateur basé sur le F-madogramme proposé par Cooley et al. (2006).

Mots-clés. Extrêmes spatiaux, processus max-stables, fonction coefficient extrémal, Géostatistique, réalisation unique, conditions de mélange.

Abstract. The extremal coefficient function is a bivariate measure of spatial dependence for stationary max-stable processes, see Schlather and Tawn (2003). It is usually

estimated from time series, when the spatial object under study is observed through time (e.g. extreme precipitations, extreme temperatures, high concentrations of pollution in the air). However, in some cases, such types of data cannot be accessed: only one or just a few records are made available. This is the case, for instance, in mining resources estimation, soil contamination evaluation or any other applications where the phenomenon of interest either varies too slowly across time to hope for a decent time series, or is too expensive to sample from. This situation is rarely addressed in the spatial extremes community, contrary to Geostatistics, which typically deals with such issues. A basic geostatistical tool is the so-called variogram, see Chilès and Delfiner (2012), which is also a bivariate measure of spatial dependence. Considering the indicator variogram, above some threshold, of a stationary max-stable random field, we propose a new nonparametric estimator of the extremal coefficient function based on the variogram's Nadaraya-Watson estimator. The latter has been studied by García-Soidán et al. (2004) and García-Soidán (2007); from their work, we derive asymptotic properties of our estimator when it is computed from a single spatial set of observations. Namely, under some assumptions, we show that it is consistent and asymptotically normal. These results are illustrated by numerical experiments and a comparison with the well-known F-madogram based estimator proposed by Cooley et al. (2006) is performed. An application on precipitation maxima dataset is also presented.

Keywords. Spatial extremes, max-stable processes, extremal coefficient function, Geostatistics, single realization, mixing conditions.

1 Précipitations bourguignonnes : un exemple introductif

Dans cet exposé, nous considérons le jeu de données illustratif suivant : pendant 30 ans, la précipitation journalière maximale a été mesurée en 146 endroits différents en Bourgogne. Afin d'étudier le comportement extrémal de ces précipitations sur une longue période, nous nous intéressons, en chaque station, au maximum des précipitations observées sur ces 30 années. Le jeu de données final est donc constitué d'une seule observation spatiale de maxima. Notons que les données ont été traitées au préalable afin que cette observation puisse être considérée comme stationnaire (au sens strict) et isotrope. La question à laquelle nous souhaitons répondre est : comment étudier la dépendance spatiale de cet unique champ de maxima ?

1.1 La théorie spatiale des valeurs extrêmes

Puisque nous nous intéressons à un champ de maxima, il peut être raisonnable de modéliser ces données par une fonction aléatoire max-stable $\big(Z(x), x \in \mathbb{R}^d\big)$, voir e.g. Ribatet (2013).

Sans perte de généralité, nous supposons également que les marginales de Z sont toutes distribuées selon la loi de Fréchet standard : $\forall x \in \mathbb{R}^d, \forall z \in \mathbb{R}, \mathbf{P}[Z(x) \leq z] = \exp\{-1/z\}$. Z est alors un processus max-stable *simple*. Cette standardisation permet de séparer l'estimation de la structure de dépendance spatiale de Z de l'évaluation de ses marginales. En accord avec les données, nous supposons aussi que Z est stationnaire (au sens strict), c'est-à-dire que ses lois finies dimensionnelles sont invariantes par translation. Nous pouvons alors considérer la fonction coefficient extrémal $\theta : h \in \mathbb{R}^d \to [1, 2]$ introduite par Schlather and Tawn (2003) et qui satisfait

$$\mathbf{P}[Z(0) \leq z,\, Z(h) \leq z] = \exp\left\{-\frac{\theta(h)}{z}\right\},$$

pour tout $h \in \mathbb{R}^d$, $z \in (0, +\infty)$. La fonction θ est une mesure bivariée de la dépendance spatiale entre des paires de points séparés par la même distance. En particulier, $\theta(h) = 1$ correspond à l'égalité (p.s.) entre $Z(0)$ et $Z(h)$ alors que $\theta(h) = 2$ correspond à leur indépendance.

Afin d'étudier la dépendance spatiale du champ de maxima bourguignons, nous cherchons à évaluer θ. Dans la littérature extrémale, cette fonction est généralement estimée à partir de données temporelles et peu de travaux s'intéressent à son évaluation lorsque Z n'est observée qu'une seule fois et de manière partielle. À notre connaissance, seuls les travaux de Bel et al. (2008), Naveau et al. (2009) et Dombry and Eyi-Minko (2012) ont abordé cette question ; les deux premiers d'un point de vue empirique et le dernier de manière théorique, lorsque Z est définie sur $\mathbb{Z}^d$. Au contraire, la problématique de l'inférence dans un contexte de réalisation unique a beaucoup été étudiée par la Géostatistique.

1.2 Un point de vue géostatistique ?

Historiquement, la Géostatistique s'intéresse à des phénomènes régionalisés, modélisés par une fonction aléatoire, qui ne sont observés que très peu de fois et bien souvent qu'une seule fois dans le temps. C'est le cas, par exemple, des études sur l'estimation des ressources minières ou sur l'évaluation de la pollution des sols et plus généralement de toute recherche dont l'objet d'étude varie très peu au cours du temps ou pour lequel le coût d'échantillonnage est trop élevé. Puisque le phénomène n'est observé qu'une seule fois, pour espérer s'en sortir d'un point de vue statistique, la Géostatistique fait l'hypothèse d'une certaine répétition spatiale de l'information. Une hypothèse fréquente est celle de stationnarité d'ordre 2. Elle stipule que l'espérance et la variance ponctuelle sont finies et identiques en tout point de l'espace. De plus, la fonction de corrélation entre deux marginales ne dépend que du vecteur les séparant : $\mu := \mathbf{E}[Z(x)] < +\infty$, $\sigma^2 := \mathbf{Var}[Z(x)] < +\infty$ et $\rho(h) := \mathbf{Corr}[Z(x), Z(x+h)]$, pour tout x, $h \in \mathbb{R}^d$. Cette hypothèse nous permet d'introduire le variogramme γ qui est également une mesure bivariée de la dépendance spatiale de Z. Il est défini, pour tout $h \in \mathbb{R}^d$, par

$$\gamma(h) := \frac{1}{2}\mathbf{E}\left[\left(Z(h) - Z(0)\right)^2\right].$$

Plusieurs travaux ont proposé des estimateurs de γ et ont étudié leurs propriétés asymptotiques lorsque ces derniers sont calculés à partir d'une unique observation partiellement observée de Z. Par la suite, nous utilisons certains de ces travaux pour proposer un nouvel estimateur de θ et étudier ses propriétés asymptotiques lorsque le processus max-stable associé est défini sur $\mathbb{R}^d$.

2 Un nouvel estimateur de θ

Nous supposons de nouveau que $\left(Z(x), x \in \mathbb{R}^d\right)$ est un processus max-stable simple stationnaire muni de sa fonction coefficient extrémal θ. Pour plus de simplicité, nous considérons également que Z est isotrope. Puisque ses marginales sont distribuées selon la loi de Fréchet standard, elles ont des moments d'ordre 1 et 2 infinis : Z n'admet donc pas de variogramme. À la place, nous nous intéressons au variogramme du champ des excès I_z, $z \in (0, +\infty)$, défini par $I_z(x) := \mathbf{1}\{Z(x) > z\}$ pour tout $x \in \mathbb{R}^d$. En effet, la fonction aléatoire I_z est stationnaire d'ordre deux et son variogramme γ_z satisfait

$$\theta(h) = -z \ln\left\{\exp\left\{-\frac{1}{z}\right\} - \gamma_z(h)\right\}.$$

Un estimateur de γ_z fournit donc un estimateur de θ. Pour limiter les hypothèses sur la structure de dépendance de Z, nous cherchons un estimateur non-paramétrique de γ_z avec de bonnes propriétés asymptotiques. Supposons que Z est observée en $n \in \mathbb{N}^*$ points $x_1, \ldots, x_n \in \mathbb{R}^d$. Parmi les différents estimateurs non paramétriques proposés et étudiés dans la littérature, nous choisissons de considérer l'estimateur suivant :

$$\hat{\gamma}_{z,\tau_n}(x) := \frac{\displaystyle\sum_{i=1}^{n}\sum_{j=1}^{n} K\left(\frac{\|x\| - \|x_i - x_j\|}{\tau_n}\right)\left[I_z(x_i) - I_z(x_j)\right]^2}{\displaystyle 2\sum_{i=1}^{n}\sum_{j=1}^{n} K\left(\frac{\|x\| - \|x_i - x_j\|}{\tau_n}\right)},$$

pour tout $x \in \mathbb{R}^d \setminus \{0\}$, avec $K > 0$ une densité de probabilité symétrique, bornée et à support compact sur $\mathbb{R}$ et $\tau_n \in (0, +\infty)$. Cet estimateur a été proposé par García-Soidán et al. (2004) et García-Soidán (2007) qui ont également démontré sa consistance et sa normalité asymptotique. À notre connaissance, ce sont les seuls travaux établissant de telles propriétés sans faire l'hypothèse que Z est m-dépendent, pour $m \in \mathbb{R}_+^*$. À l'aide de leurs résultats, nous avons établi les propriétés asymptotiques de l'estimateur correspondant $\hat{\theta}_z$ de θ.

2.1 Propriétés asymptotiques

García-Soidán et al. (2004) et García-Soidán (2007) ont montré sous certaines conditions que $\hat{\gamma}_{z,\tau_n}$ était consistant et asymptotiquement normal lorsque $n \to +\infty$. Concernant les observations, ils supposent que l'échantillonage est aléatoire et qu'il devient de plus en plus dense et de plus en plus étendu spatialement, à mesure que $n \to +\infty$. Cela correspond à un mélange entre l'asymptotique *infill* et l'asymptotique *increasing domain*. Concernant la structure de dépendance de I_z, ils font certaines hypothèses de mélange qui ne sont pas évidentes à vérifier dans le cadre des processus max-stables. À partir de leurs travaux, nous avons donc établi de nouvelles conditions de mélange, plus faciles à interpréter, qui garantissent les propriétés asymptotiques de $\hat{\gamma}_{z,\tau_n}$ et donc de $\hat{\theta}_z$. En particulier, pour $z > 1.45$, nous montrons que si

$$\lim_{\|h\|\to+\infty} \theta(h) \text{ existe et } \int_{\mathbb{R}^d} 2 - \theta(h)\, dh < +\infty \tag{2.1}$$

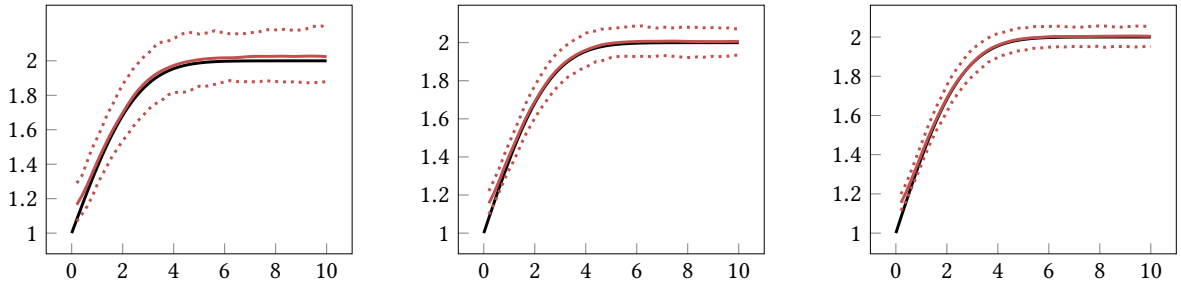
alors $\hat{\theta}_z$ est consistant. Nous montrons également que si $\lim_{\|h\|\to+\infty} \theta(h) = 2$ et si

$$\int_{\substack{\|x_1\|\leq s \quad \|x_2-x_3\|\leq s,\\ \|x_4-x_5\|\leq s \quad \|x_6-x_7\|\leq s}} \left| \mathbf{E}\left[\prod_{j=0}^{3} \left(\left[I_z(x_{2j}) - I_z(x_{2j+1})\right]^2 - \mathbf{E}\left[\left[I_z(x_{2j}) - I_z(x_{2j+1})\right]^2\right] \right) \right] \right| dx_1 \dots dx_7 \tag{2.2}$$

est finie alors $\hat{\theta}_z$ est asymptotiquement normal. Notons que cette dernière condition est déjà proposée par García-Soidán (2007) et reste difficile à vérifier pour des processus max-stables. Nous cherchons actuellement a montré qu'elle peut être remplacée par la condition (2.1). Les résultats concernant la consistance de $\hat{\theta}_z$ sont illustrés à l'aide de simulations numériques, *cf.* Sous-section 2.2. Nous comparons également les performances de notre estimateur avec celle de l'estimateur basé sur le F-madogramme proposé par Cooley et al. (2006). Enfin nous utilisons ce nouvel estimateur $\hat{\theta}_z$ pour évaluer la dépendance spatiale bivariée du champ des précipitations maximales bourguignonnes.

2.2 Quelques illustrations

Soit Z un processus *Smith*, *i.e.* un modèle spécifique de processus max-stables simples stationnaires et isotropes, *cf.* Ribatet (2013). Sa fonction coefficient extrémal θ satisfait la condition (2.1) ; $\hat{\theta}_z$ est donc consistant. Posons $z_3 = 3.48$, le 3ième quartile de la loi Fréchet standard. Pour $n \in \{300, 1000, 2000\}$, nous avons simulé 500 réalisations de $(Z(x_1), \dots, Z(x_n))$. Pour $z \approx z_3$, K un noyau gaussien tronqué et (τ_n) une suite de fenêtres déterminées par validation croisée, nous avons ensuite estimé γ_z puis θ, pour chaque simulation. Les courbes présentées en Figure 1 semblent indiquer que $\hat{\theta}_z$ est consistant, ce qui concordent avec les résultats théoriques. Concernant la normalité asymptotique, nous ne savons pas si Z satisfait la condition (2.2) mais une investigation numérique suggère que $\sqrt{n^\alpha}(\hat{\theta}_z(h) - \theta(h))$ converge vers une loi normale, pour $h \in \mathbb{R}^d \setminus \{0\}$ et $\alpha = 8/9$.



(a) $n = 300$ (b) $n = 1000$ (c) $n = 2000$

Figure 1: Courbe moyenne (en rouge) et bande de confiance à 90% (pointillés rouges) de $\hat{\theta}_z$ évaluées en différentes distances $\|h\|$ (en abscisse) à partir des 500 simulations de $(Z(x_1), \ldots, Z(x_n))$. Pour comparaison, θ est également représenté (courbe noire).

Bibliographie

L. Bel, J. N. Bacro, and C. Lantuéjoul. Assessing extremal dependence of environmental spatial fields. *Environmetrics*, 19(2):163–182, 2008.

J.-P. Chilès and P. Delfiner. *Geostatistics: Modeling Spatial Uncertainty*. Wiley series in probability and statistics. John Wiley & Sons Inc, second edition, 2012.

D. Cooley, P. Naveau, and P. Poncet. *Variograms for spatial max-stable random fields*, pages 373–390. Springer New York, New York, NY, 2006.

C. Dombry and F. Eyi-Minko. Strong mixing properties of max-infinitely divisible random fields. *Stochastic Processes and their Applications*, 122(11):3790 – 3811, 2012.

P. García-Soidán. Asymptotic normality of the nadaraya–watson semivariogram estimators. *TEST*, 16(3):479–503, Dec 2007.

P. H. García-Soidán, M. Febrero-Bande, and W. González-Manteiga. Nonparametric kernel estimation of an isotropic variogram. *JSPI*, 121(1):65 – 92, 2004.

P. Naveau, A. Guillou, D. Cooley, and J. Diebolt. Modelling pairwise dependence of maxima in space. *Biometrika*, 96(1):1–17, 03 2009.

M. Ribatet. Spatial extremes: Max-stable processes at work. *Journal de la Société Française de Statistique*, 154(2):156–177, 2013.

M. Schlather and J. Tawn. A dependence measure for multivariate and spatial extreme values: Properties and inference. *Biometrika*, 90(1):139–156, 2003.

Maximum Mean Discrepancy invariant par translation et convergence en loi

Thibault Modeste [1], Clément Dombry [2] & Anne-Laure Fougères [1]

[1] *Univ. Lyon, Université Claude Bernard Lyon 1, CNRS UMR 5208, Institut Camille Jordan, F-69622 Villeurbanne, France E-mail :* modeste@math.univ-lyon1.fr, fougeres@math.univ-lyon1.fr.

[2] *Université Bourgogne Franche-Comté, Laboratoire de Mathématiques de Besançon, CNRS UMR 6623, F-25000 Besançon, France. E-mail :* clement.dombry@univ-fcomte.fr

Résumé. Dans ce travail nous nous intéresserons aux relations entre les RKHS et les mesures de probabilités sur $\mathbf{R}^d$. Nous nous focaliserons plus particulièrement sur les noyaux ayant un Maximum Mean Discrepancy (MMD) invariant par translation, ce qui permet d'inclure des noyaux k non bornés. Nous étudierons en premier les liens entre ces MMD et la convergence en loi. Puis nous utiliserons le fait que les noyaux k ne soient pas forcément bornés pour considérer une distance plus forte, la distance de Wasserstein.

Mots-clés. RKHS, MMD, Score à noyaux, Convergence en loi, distance de Wasserstein.

Abstract. In this talk, we will focus on the relations between RKHS and probability measures on $\mathbf{R}^d$. We will be interested in kernels with a translation invariant Maximum Mean Discrepancy (MMD) which allows unbounded kernels. First, we will look at the links between these MMD and weak convergence. Then we will use the fact that the kernels k are not necessarily bounded to consider a stronger distance, the W_1 Wasserstein distance.

Keywords. RKHS, MMD, Kernel Score, Weak convergence, Wasserstein distance.

1 Introduction

Les « espaces de Hilbert à noyaux reproduisants » (RKHS) sont des structures mathématiques aux applications nombreuses, notamment en apprentissage machine (ML). Ils permettent en particulier de résoudre des problèmes de régression et de classification tels que l'analyse en composantes principales (ACP) ou les machines à vecteurs de support (SVM) voir par exemple Muandet et al. (2017) pour plus de détails entre les liens étroits entre RKHS et ML. Dans ce travail nous nous intéresserons aux relations entre ces RKHS et les mesures de probabilités sur un espace topologique $\mathcal{X}$. Notons k le noyau associé qui permet de caractériser le produit scalaire sur le RKHS H. À partir du noyau k, on peut construire une pseudo métrique γ_k, appelée Maximum Mean Discrepancy (MMD), sur l'espace des mesures signées finies. Cette pseudo métrique vérifie toutes les propriétés d'une distance, excepté l'axiome de séparation. Notre problème se décompose en deux questions. La première concerne la capacité de l'espace de Hilbert H à différencier les mesures signées, c'est à dire à voir si γ_k est une vraie distance. L'article Steinwart and Ziegel (2021) caractérise les noyaux k rendant γ_k séparées dans le cas où $\mathcal{X}$ est un groupe abélien compact. Dans Sriperumbudur et al. (2010), la question est résolue dans le cas où $\mathcal{X} = \mathbf{R}^d$ et le noyau k est continu borné et invariant par translation, c'est à dire qu'il existe $\psi \in \mathcal{C}_b$ tel que

$$k(x, y) = \psi(x - y).$$

Dans le cas où γ_k est une distance, on peut regarder ces liens avec les différentes distances sur l'espace des mesures. On trouve dans la littérature essentiellement des réponses lorsque le noyau k est borné et pour la convergence en loi. Dans le cas où l'espace de Hilbert H est inclus dans l'espace des fonctions continues tendant vers 0 en dehors de tout compact, noté $\mathcal{C}_0$, l'article Simon-Gabriel et al. (2021) montre que γ_k métrise la convergence en loi si et seulement si le noyau k est continu et γ_k est une distance. Sriperumbudur et al. (2010) caractérise lui aussi à quel moment, les noyaux continus bornés k invariants par translation métrisent avec γ_k la convergence en loi. Le résultat n'est pas inclus dans l'article précédent car les auteurs ne supposent pas $H \subset \mathcal{C}_0$.

Le but de cet exposé est de traiter le cas où l'on suppose l'invariance par translation non pas de k mais du MMD γ_k c'est à dire pour μ, ν deux mesures et $x \in \mathbf{R}^d$,

$$\gamma_k(\mu, \nu) = \gamma_k(\mu * \delta_x, \nu * \delta_x), \tag{1}$$

où δ_x est la mesure Dirac en x et $*$ le produit de convolution de mesures. On commencera d'abord à donner quelques définitions et propriétés de base sur les RKHS, en rappelant le lien avec les Scores à noyaux. On généralisera ensuite les résultats de Sriperumbudur et al. (2010) à ces noyaux et pour une distance plus forte que la convergence en loi. Cette généralisation est résumée dans le Théorème 1.

2 RKHS

Pour éviter des problèmes de mesurabilité de certaines fonctions, on présente la construction des RKHS que sur $\mathbf{R}^d$, muni de sa tribu borélienne, et avec un noyau k continu. Se référer à (Steinwart and Christmann, 2008, Section 4.3) pour plus de détails.

2.1 Définition

Définition 1. *Soit H un espace de Hilbert inclus dans $\mathcal{F}(\mathbf{R}^d, \mathbf{R})$ l'ensemble des fonctions à valeurs réelles. L'espace H est appelé RKHS si pour tout $x \in \mathbf{R}^d$, l'évaluation $ev_x \colon f \mapsto f(x)$ est continue.*

Par le théorème de Riesz, on peut alors trouver $K(x) \in H$ tel que

$$\forall f \in H,\ \ f(x) = \langle f, K(x) \rangle.$$

On définit alors le noyau $k(x, y) = \langle K(x), K(y) \rangle$. On supposera dans la suite que k est continu. Dans ce cas, les fonctions de H sont mesurables. Comme k n'est pas nécessairement borné, on va devoir se restreindre à un sous ensemble de $\mathcal{M}(\mathbf{R}^d)$, l'espace des mesures boréliennes finies signées

$$\mathcal{M}^k = \{\mu \in \mathcal{M}(\mathbf{R}^d) \mid \int \sqrt{k(x,x)}\ \mathrm{d}|\mu|(x) < +\infty\}.$$

Pour $\mu \in \mathcal{M}^k$, on peut montrer par une inégalité de Cauchy-Schwarz que le RKHS H est inclus dans $\mathcal{L}^1(\mathbf{R}^d, \mu)$, l'ensemble des fonctions mesurables pour μ. En utilisant le Théorème du Graphe Fermé pour les espaces de Banach, on obtient la continuité de l'identité de $(H, \|\cdot\|_H)$ dans $(H, \|\cdot\|_{\mu,1})$. Avec cette continuité, on obtient le résultat suivant que l'on utilisera pour construire la pseudo métrique γ_k.

Proposition 1. *Il existe un unique $\Phi(\mu) \in H$ tel que, $\forall f \in H$, $\langle \Phi(\mu), f \rangle = \int f(x)\ \mathrm{d}\mu(x)$.*

L'application Φ définie sur $\mathcal{M}^k$ permet de comparer les mesures à travers H. Elle est naturellement linéaire. Pour μ, ν deux mesures, on pose

$$\gamma_k(\mu, \nu) = \|\Phi(\mu) - \Phi(\nu)\|_H.$$

On voit facilement que γ_k est une distance si et seulement Φ est injective. Par un petit calcul, on peut donner une autre forme à la pseudo distance

$$\gamma_k^2(\mu, \nu) = \int k\ \mathrm{d}(\mu - \nu) \otimes (\mu - \nu).$$

2.2 Lien avec les Scores à noyau

Pour celles et ceux qui sont déjà familiers avec le langage des scores, on peut voir la divergence d'un score à noyau S_k comme le carré d'un MMD pour un autre noyau $\tilde{k}$. Nous évoquerons ce lien sans définir les scores (voir Gneiting and Raftery (2007)). À la différence des RKHS, le noyau k n'est pas positif mais conditionnellement négatif. En effet, la preuve qu'un score à noyau est propre se repose sur l'inégalité de Hoeffding (Chapitre 7 Théorème 2.1 Berg et al. (1984)) qui fonctionne pour les noyaux conditionnellement négatifs. Dans ce même livre, le lemme 2.1 du Chapitre 3 montre qu'à tout noyau conditionnellement négatif k, on peut lui associer un noyau positif $\tilde{k}$. Puis par un calcul direct, on trouve que

$$\overline{S_k}(\mu,\nu) - \overline{S_k}(\nu,\nu) = \gamma_{\tilde{k}}^2(\mu,\nu).$$

3 MMD invariant par translation

Soit H un RKHS sur $\mathbf{R}^d$ et soit k le noyau associé. On suppose que k est continu et que le MMD γ_k vérifie (1). On peut montrer qu'alors il existe un noyau $\tilde{k}$ définissant le même MMD tel que $\tilde{k}$ soit la fonction de covariance d'un processus $(X_t)_{t\in\mathbf{R}^d}$ à accroissements stationnaires. Ainsi par l'équation (2.63) de Cohen and Istas (2013), on peut écrire

$$\tilde{k}(x,y) = \int \left(\mathrm{e}^{ix\cdot t} - 1\right)\left(\mathrm{e}^{-iy\cdot t} - 1\right)\, \mathrm{d}\Lambda(t) + y^T\Sigma x,$$

où Σ est une matrice symétrique positive et Λ une mesure sigma-finie positive vérifiant

$$\int 1 \wedge \|t\|_2^2 \, \mathrm{d}\Lambda(t) < +\infty.$$

De cette écriture, nous montrerons que

Théorème 1. *1. γ_k est une distance si et seulement si* $\mathrm{supp}(\Lambda) = \mathbf{R}^d$ *;*

2. γ_k métrise la convergence en loi si et seulement si $\mathrm{supp}(\Lambda) = \mathbf{R}^d$*, Λ est une mesure finie et $\Sigma = 0$;*

3. Si $\mathrm{supp}(\Lambda) = \mathbf{R}^d$*, Λ est une mesure finie et Σ strictement positive alors γ_k métrise la distance de Wasserstein W_1.*

Références

Berg, C., Christensen, J. P. R., and Ressel, P. (1984). *Harmonic Analysis on Semigroups*. Springer-Verlag.

Cohen, S. and Istas, J. (2013). *Fractional Fields and Applications*. Mathématiques et Applications. Springer. volume 76.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477) :359–378.

Muandet, K., Fukumizu, K., Sriperumbudur, B., and Schölkopf, B. (2017). Kernel mean embedding of distributions : A review and beyond. *Foundations and Trends in Machine Learning*, 10(1-2) :1–141.

Simon-Gabriel, C.-J., Barp, A., Schölkopf, B., and Mackey, L. (2021). Metrizing weak convergence with maximum mean discrepancies.

Sriperumbudur, B. K., Gretton, A., Fukumizu, K., Schölkopf, B., and Lanckriet, G. R. (2010). Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11(50) :1517–1561.

Steinwart, I. and Christmann, A. (2008). *Support Vector Machines*. Springer Publishing Company, Incorporated, 1st edition.

Steinwart, I. and Ziegel, J. F. (2021). Strictly proper kernel scores and characteristic kernels on compact spaces. *Applied and Computational Harmonic Analysis*, 51 :510–542.

Vitesse de Convergence Minimax pour la Régression Distributionnelle basée sur le Continuous Ranked Probability Score

Romain Pic [1], Clément Dombry [1], Philippe Naveau [2] & Maxime Taillardat [3]

[1] *Université Bourgogne Franche-Comté, Laboratoire de Mathématiques de Besançon, UMR CNRS 6623, 25000 Besançon, France.*
E-mails : romain.pic@univ-fcomte.fr, clement.dombry@univ-fcomte.fr
[2] *Université Paris-Saclay, Laboratoire des Sciences du Climat et de l'Environnement, UMR 8212, 91191 Gif-sur-Yvette, France. E-mail : philippe.naveau@lsce.ipsl.fr*
[3] *Météo France, Centre National de Recherches Météorologique, UMR 3589, 31057 Toulouse, France. E-mail : maxime.taillardat@meteo.fr*

Résumé. La régression distributionnelle répond à un besoin fondamental de l'analyse statistique : permettre de faire des prévisions tout en quantifiant leur incertitude. Cette approche surmonte les limites de la régression classique qui estime uniquement l'espérance conditionnellement aux covariables en fournissant un estimateur de l'intégralité de la loi conditionnelle. Cette méthodologie, dite de prédiction probabiliste, est largement adoptée dans de nombreux domaines tels que la météorologie et la production d'énergie, mais ses aspects théoriques restent peu développés. Par analogie avec la théorie classique de l'apprentissage statistique, nous définissons un cadre où le prédicteur est une loi de probabilité, dite loi prédictive, et où la fonction de perte est donnée par un score strictement propre au sens de Gneiting et Raftery (2007). Le prédicteur de Bayes coïncide alors avec la loi conditionnelle. Dans le cas du CRPS, nous étudions ensuite la vitesse minimax de convergence et montrons en particulier que l'algorithme des k plus proches voisins atteint le taux minimax optimal en dimension supérieure ou égale à 2 et que les méthodes à noyau réalise ce taux optimal en dimension quelconque.

Mots-clés. régression distributionnelle, apprentissage statistique, prédicteur de Bayes, taux minimax, méthode des plus proches voisins, méthodes à noyau

Abstract. The distributional regression fulfills a fundamental need of statistical analysis : being able to make forecasts and quantify their uncertainty. This approach overcomes the limits of classical regression which estimates only the conditional mean by estimation the whole distribution law. This methodology, called probabilistic forecast, is widely used in numerous fields such as meteorology and energy production, but its theoretical aspects have not been studied. By analogy with the classical theory of statistical learning, we define a framework where the predictor is a law of probability, called prediction law, and where the loss function is given by a strictly proper scoring rule in the sense of Gneiting and Raftery (2007). Bayes predictor is then the conditional law. In the case of the Continuous Ranked Probability Score, we study then the minimax rate of convergence and show that the k nearest neighbor algorithm reaches the optimal rate of convergence

in dimension higher or equal to 2 and that the kernel methods reach this optimal rate in any dimension.

Keywords. distributional regression, statistical learning, Bayes predictor, minimax rate, nearest neighbors method, kernel methods

1 Introduction

De nombreux domaines nécessitent la capacité de prédire la valeur de variables d'intérêt en utilisant l'information de variables auxiliaires mais aussi de quantifier l'incertitude de la prédiction. Cela a amené à un changement transdisciplinaire de paradigme de la régression ponctuelle vers la régression distributionnelle (Gneiting, 2008).

La régression distributionnelle est largement utilisée en post-traitement météorologique avec une approche visant à minimiser le risque empirique d'un algorithme à partir du Continuous Ranked Probability Score (CRPS). Pour plus de détails sur les méthodes de post-traitement, Taillardat et al. (2019) et Schulz and Lerch (2021) présentent l'état de l'art et comparent les performances de nombreux modèles.

Dans les travaux présentés ci-après, nous commençons par introduire les différentes notions liées à la régression distributionnelle et à l'apprentissage statistique. Puis, nous présentons des résultats sur les vitesses de convergence minimax de la méthode des k plus proches voisins et des méthodes à noyau dans le cadre de la régression distributionnelle.

2 Régression Distributionnelle et Apprentissage Statistique

Nous considérons $Y \in \mathbb{R}$ la variable aléatoire cible (e.g. les précipitations) et $X \in \mathbb{R}^d$ la variable aléatoire représentant les covariables (e.g. température, humidité, pression, etc...). L'objectif de la régression distributionnelle est d'estimer la distribution conditionnelle de Y sachant $X = x$, notée $\mathbb{P}_{Y|X=x}(\mathrm{d}y) = F_x^*(\mathrm{d}y)$.

Définition 1. *Un prédicteur est une application mesurable* $\hat{F} \colon \mathbb{R}^d \to \mathcal{P}$*, avec* $\mathcal{P} \subseteq \mathcal{M}(\mathbb{R})$ *et* $\mathcal{M}(\mathbb{R})$ *l'ensemble des distributions sur* $\mathbb{R}$*.*

L'apprentissage statistique fait intervenir une fonction de perte qui compare la prédiction avec des observations. Dans le cadre de la prédiction probabiliste, les fonctions de perte sont appelées *scores* et confrontent la distribution prédictive $\hat{F}$ avec une observation (x, y).

Définition 2. *Un score est une fonction à valeurs réelles étendues* $S : \mathcal{P} \times \mathbb{R} \to [-\infty, +\infty]$*, telle que* $S(F, .)$ *est* $\mathcal{P}$*-quasi-intégrable pour tout* $F \in \mathcal{P}$*.*

La fonction de perte est utilisée pour évaluer la qualité d'un prédicteur grâce au *risque théorique*, correspondant à l'espérance du score. Le risque dépend de P, la distribution de (X, Y).

Définition 3. *Le risque d'un prédicteur $\hat{F}$ est défini par $R_P(\hat{F}) = \mathbb{E}[S(\hat{F}_X, Y)]$.*

On note $\bar{S}(F, G) = \mathbb{E}_{Y\sim G}[S(F, Y)]$ le score moyen d'une distribution prédictive donnée F lorsque les observations suivent une loi G.

Définition 4. *Soit S un score et $\mathcal{P} \subset \mathcal{M}(\mathbb{R})$. S est dit propre sur $\mathcal{P}$ si :*

$$\bar{S}(G, G) \leq \bar{S}(F, G) \quad \textit{pour tout } F, G \in \mathcal{P}. \tag{1}$$

On dit que S est strictement propre si l'égalité dans (1) tient si et seulement si $F = G$.

Définition 5. *Le **risque de Bayes** est défini par :*

$$R_P^* = \inf_{\hat{F}} \mathbb{E}_{(X,Y)\sim P}[S(\hat{F}_X, Y)]$$

*avec $\hat{F}_X$ la fonction de répartition prédite de Y sachant X. De plus, si l'infimum est atteint, le prédicteur associé F^* est appelé **prédicteur de Bayes**.*

Nous considérons le Continuous Ranked Probability Score (CRPS) qui est un score strictement propre largement utilisé pour les prévisions météorologiques. Le CRPS a été défini dans Matheson and Winkler (1976) comme suit :

$$S(F, y) = \int_{\mathbb{R}} (F(z) - \mathbb{1}_{y\leq z})^2 \mathrm{d}z$$

Proposition 1. *Considérons la régression distributionnelle avec le CRPS comme score et supposons Y intégrable. Alors, le risque de Bayes est fini et vaut*

$$R_P^* = \int F_x^*(z)(1 - F_x^*(z)) P_X(\mathrm{d}x)\mathrm{d}z. \tag{2}$$

Tout prédicteur F vérifie $R_P(F) \geq R_P(F^*)$ avec égalité si et seulement si $F_x = F_x^*$ P_X-p.p.

Dans notre étude, nous nous sommes placés dans le cadre de l'apprentissage statistique qui correspond au cas pratique où l'on dispose d'un nombre n d'observations de la variable cible et des covariables et où l'on cherche à prédire la distribution de Y sachant les covariables X. On note l'échantillon de n observation

$$D_n = \{(X_i, Y_i), i \in [\![1; n]\!]\}.$$

Définition 6. *Un algorithme de prédiction est une application mesurable*

$$\hat{F} : (\mathbb{R}^d \times \mathbb{R})^n \to \mathcal{F}(\mathbb{R}^d, \mathcal{P}), \ \textit{avec } \mathcal{P} \subseteq \mathcal{M}(\mathbb{R}).$$
$$d_n \mapsto \hat{F}_n$$

$\hat{F}$ associe un prédicteur $\hat{F}_n$ à un échantillon $d_n = \{(x_i, y_i), i \in [\![1; n]\!]\}$. On note $\hat{F}_{n,X} = \hat{F}_n(X)$ la distribution prédite de Y sachant X à partir des observations d_n.

3 Taux de Convergence Optimal

3.1 Résultat principal

Dans ce contexte théorique, nous nous intéressons la vitesse de convergence du risque théorique d'un algorithme de prédiction vers le risque de Bayes. Nous étudions donc la grandeur suivante :

$$\mathbb{E}[S(\hat{F}_{n,X}, Y) - S(F_X^*, Y)] = \mathbb{E}\left[\int |\hat{F}_{n,X}(z) - F_X^*(z)|^2 \mathrm{d}z\right].$$

Pour étudier le taux de convergence des algorithmes, nous nous intéressons aux notions de *vitesse de convergence minimax minimale* et de *vitesse de convergence minimax optimale.*

Définition 7. *La suite de nombres positifs (a_n) est appelée vitesse de convergence minimax optimale pour la classe $\mathcal{D}$ si*

$$\liminf_{n\to\infty} \inf_{\hat{F}_n} \sup_{(X,Y)\in\mathcal{D}} \frac{\mathbb{E}[S(\hat{F}_{n,X}, Y)] - \mathbb{E}[S(F_X^*, Y)]}{a_n} = C_1 > 0$$

et

$$\limsup_{n\to\infty} \inf_{\hat{F}_n} \sup_{(X,Y)\in\mathcal{D}} \frac{\mathbb{E}[S(\hat{F}_{n,X}(Y), Y)] - \mathbb{E}[S(F_X^*(Y), Y)]}{a_n} = C_0 < \infty$$

où l'infimum est pris sur tout les modèles de régression distributionnelle $\hat{F}_n$. Si (a_n) satisfait uniquement la borne inférieure (7), elle est appelée vitesse de convergence minimax minimale.

La notion de vitesse de convergence minimax minimale est liée au fait de vouloir minimiser l'erreur maximale au sein d'une classe de distribution.

Notre résultat principal est l'obtention de la vitesse de convergence minimax optimale pour la régression distributionnelle avec le CRPS comme score. Sachant que si $(X, Y) \sim P$, alors P_X est la distribution marginale de X et F_x^* est la distribution conditionnelle de Y sachant X. Nous considérons les classes de distributions suivantes.

Définition 8. *Pour $H \in (0, 1]$, $C > 0$ et $M > 0$, soit $\mathcal{D}^{(H,C,M)}$ une classe de distributions P telle que :*

i) $X \in [0, 1]^d$ P_X*-p.s. ;*

ii) Pour tout $x \in [0, 1]^d$, $\int_{\mathbb{R}} F_x^*(z)(1 - F_x^*(z))\mathrm{d}z < M$ *; et*

iii) $\|F_{x'}^* - F_x^*\|_{L^2} \leq C\|x' - x\|^H$ *pour tout* $x, x' \in [0, 1]^d$.

Théorème 2. *La suite* $a_n = n^{-\frac{2H}{2H+d}}$ *est la vitesse de convergence minimax minimale pour la classe* $\mathcal{D}^{(H,C,M)}$.

La preuve du Théorème 2 est divisée en trois étapes :

1) Nous commençons par montrer que $a_n = n^{-\frac{2H}{2H+d}}$ est une vitesse de convergence minimax minimale ; l'argument principal est de montrer qu'il suffit de considérer le cas d'une prédiction binaire lorsque Y et F_X^* prennent des valeurs dans $\{0, 1\}$. Cela permet de se ramener à un résultat de régression ponctuelle étudiée dans Györfi et al. (2002).
2) Nous montrons ensuite que la méthode des k plus proches voisins (k-NN) permet d'atteindre cette vitesse de convergence minimax minimale pour un choix particulier de $k = k_n$; cela ne fonctionne que pour $d \geq 2$.
3) Pour compléter la preuve dans le cas $d = 1$, nous montrons que, pour tout $d \geq 1$, les méthodes à noyau atteignent la vitesse de convergence minimax minimale de l'étape 1 pour un choix adapté de largeur de bande $h = h_n$.

3.2 k plus proches voisins

Nous introduisons l'algorithme des k plus proches voisins (k-NN) adapté à la régression distributionnelle. La méthode des k-NN estime la distribution conditionnelle de Y sachant $X = x$ en moyennant des distributions de Dirac centrées sur la valeur de la variable cible prise par les k plus proches voisins de x parmi l'échantillon utilisé pour l'apprentissage d_n. Il peut être formulé de la manière suivante :

$$\hat{F}_{k,x}(z) = \frac{1}{k} \sum_{i=1}^{k} \mathbb{1}_{y_{(i)} \leq z} \tag{3}$$

où l'indexation par (i) correspond au i-ème plus proche voisin de x parmi les observations utilisées. Par analogie avec des résultats en apprentissage statistique classique (Biau and Devroye (2015) ; Györfi et al. (2002)), nous avons obtenus des résultats concernant la vitesse de convergence minimax optimale pour la classe de distribution $\mathcal{D}^{(H,C,M)}$ lorsque la dimension des covariables d est supérieure à 2.

Proposition 3. *Supposons* $P \in \mathcal{D}^{(H,C,M)}$ *et soit* $\hat{F}_n$ *le modèle des k plus proches voisins défini par l'Equation* (3). *Alors,*

$$\mathbb{E}[R_P(\hat{F}_n)] - R_P^* \leq \begin{cases} 8^H C^2 \left(\frac{k_n}{n}\right)^H + \frac{M}{k_n} & \text{if } d = 1, \\ {c_d}^H C^2 \left(\frac{k_n}{n}\right)^{2H/d} + \frac{M}{k_n} & \text{if } d \geq 2, \end{cases}$$

où $c_d = \frac{2^{3+\frac{2}{d}}(1+\sqrt{d})^2}{V_d^{2/d}}$ *et* V_d *est le volume de la boule unitaire de* $\mathbb{R}^d$. *Le choix optimal de* k_n *permet d'obtenir la vitesse de convergence minimax optimale lorsque* $d \geq 2$.

3.3 Méthodes à noyau

Les méthodes à noyau peuvent être naturellement étendues à la régression distributionnelle et nous considérons le noyau naïf suivant :

$$\hat{F}_{n,x}(z) = \frac{\sum_{i=1}^{n} \mathbb{1}_{Y_i \leq z} \mathbb{1}_{\|X_i - x\| < h_n}}{n P_n(S_{x,h_n})}, \tag{4}$$

si le dénominateur est non nul, et $\mathbb{1}_{0 \leq z}$ sinon.

Proposition 4. *Supposons $P \in \mathcal{D}^{(H,C,M)}$ et soit $\hat{F}_n$ le modèle à noyau défini par l'Equation* (4)*. Alors,*

$$\mathbb{E}[R_P(\hat{F}_n)] - R_P^* \leq \hat{c} \frac{M + \sup_{x \in [0,1]^d} \int_{\mathbb{R}} \left(\mathbb{1}_{z \geq 0} - F_x^*(z)\right)^2 \mathrm{d}z}{n {h_n}^d} + C^2 {h_n}^{2H}$$

où $\hat{c}$ dépend uniquement de d. Le choix optimal de h_n permet d'obtenir la vitesse de convergence minimax optimale pour tout $d \geq 1$.

Références

[1] Gérard Biau and Luc Devroye. *Lectures on the Nearest Neighbor Method*. Springer Series in the Data Sciences. Springer, 2015.

[2] Tilmann Gneiting. Editorial : Probabilistic forecasting. *Journal of the Royal Statistical Society : Series A (Statistics in Society)*, 171, apr 2008. doi : 10.1111/j.1467-985x.2007.00522.x.

[3] Lászlò Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer Series in Statistics. Springer, 2002.

[4] James E. Matheson and Robert L. Winkler. Scoring rules for continuous probability distributions. *Management Science*, 22, jun 1976. doi : 10.2307/2629907.

[5] Benedikt Schulz and Sebastian Lerch. Machine learning methods for postprocessing ensemble forecasts of wind gusts : A systematic comparison, 2021. arXiv:2106.09512.

[6] Maxime Taillardat, Anne-Laure Fougères, Philippe Naveau, and Olivier Mestre. Forest-based and semiparametric methods for the postprocessing of rainfall ensemble forecasting. *Weather and Forecasting*, 34, jun 2019. doi : 10.1175/WAF-D-18-0149.1.

Non-Asymptotic Analysis of Stochastic Approximation Algorithms for Streaming Data

Antoine Godichon-Baggioni, **Nicklas Werge*** & Olivier Wintenberger

LPSM, Sorbonne Université, 4 place Jussieu, 75005 Paris, France

Abstract. Motivated by the high-frequency data streams continuously generated, real-time learning is becoming increasingly important. These data streams should be processed sequentially with the property that the data stream may change over time. In this streaming setting, we propose techniques for minimizing convex objective functions through unbiased estimates of their gradients, commonly referred to as stochastic approximation problems. Our methods rely on stochastic gradient algorithms because of their applicability and computational advantages. The reasoning includes iterate averaging that guarantees optimal statistical efficiency under classical conditions. Our non-asymptotic analysis shows accelerated convergence by selecting the learning rate according to the expected data streams. We show that the average estimate converges optimally and robustly for any data stream rate. In addition, noise reduction can be achieved by processing the data in a specific pattern, which is advantageous for large-scale machine learning problems. These theoretical results are illustrated for various data streams, showing the effectiveness of the proposed algorithms.

Keywords. machine learning, large-scale, stochastic approximation, stochastic optimization, streaming data

1 Introduction

Machine learning and artificial intelligence have become an integral part of modern society. This massive utilization of intelligent systems generates an endless sequence of data, many of which come as *streaming data* such as internet traffic data, financial investments, self-driving cars, or sensor data. It requires robust and time-efficient algorithms to analyze and process such data without compromising accuracy. This problem has attracted a lot of attention in the machine learning community (Bottou and LeCun, 2004; Zhang, 2004; Xiao, 2010; Shalev-Shwartz et al., 2011).

Even after 70 years, stochastic approximation algorithms are still widely used for handling large amounts of data (Robbins and Monro, 1951); the most well-known is presumably the Stochastic Gradient (SG) method, which has led to numerous extensions*.

*Email address: nicklas.werge@upmc.fr

*Bottou et al. (2018) reviews these stochastic algorithms for large-scale machine learning, including noise reduction and second-order methods, among others.

An essential extension is the Polyak-Ruppert averaging (ASG) proposed by Polyak and Juditsky (1992) and Ruppert (1988), which guarantees optimal statistical efficiency without jeopardizing the computational cost.

A fundamental aspect of this paper is to explore how changing data streams affect these optimization methods. This is a direct extension of Bach and Moulines (2011) to a streaming setting. We examine two different kinds of data streams: constant and varying streaming-batches. These data streams includes everything from vanilla SG and ASG, mini-batch SG and ASG, to more exotic learning designs. Our main theoretical contribution is the non-asymptotic analysis of the SG and ASG (Sections 2 and 3) methods in this streaming framework. Our findings show a noticeable improvement in the convergence rates by selecting the *learning rate* according to the expected data streams. In particular, we show how to obtain *optimal* convergence rates *robust* to any data streaming rate.

1.1 Problem Formulation

The objective of stochastic approximation algorithms is to minimize convex functions of the form $L(\theta) = \mathbb{E}[l_t(\theta)]$ with respect to $\theta \in \mathbb{R}^d$ for $d \geq 1$. The minimization of L is achieved without evaluating it directly, but instead by using unbiased functions $l_t : \mathbb{R}^d \to \mathbb{R}$. Let $(l_t)_{t\geq 1}$ constitute a sequence of independent differentiable random functions and let their gradients unbiased estimates of the gradient of L.

Let us outline our streaming setting: at each time $t \in \mathbb{N}$, we consider $n_t \in \mathbb{N}$ random functions $(l_{t,1}, \ldots, l_{t,n_t})$ with $l_t = n_t^{-1} \sum_{i=1}^{n_t} l_{t,i}$. The Stochastic Streaming Gradient (SSG) is defined as

$$\theta_t = \theta_{t-1} - \frac{\gamma_t}{n_t} \sum_{i=1}^{n_t} \nabla_\theta l_{t,i}(\theta_{t-1}), \tag{1}$$

with $\theta_0 \in \mathbb{R}^d$ where (γ_t) is the *learning rate* satisfying $\sum_{i=1}^t \gamma_i = \infty$ and $\sum_{i=1}^t \gamma_i^2 < \infty$ for $t \to \infty$. The accumulated sum is denoted by $N_t = \sum_{i=1}^t n_i$. To guarantee optimal convergence, we introduce the Averaged Stochastic Streaming Gradient (ASSG), defined for all $t \in \mathbb{N}$ by

$$\bar{\theta}_t = \frac{1}{N_t} \sum_{i=0}^{t-1} n_{i+1}\theta_i, \quad \bar{\theta}_0 = 0. \tag{2}$$

1.2 Strongly Convex Objectives

Following Sridharan et al. (2008), we make the following assumptions on L: the function L is μ-strongly convex for some $\mu > 0$, that is, for all $\theta, \theta' \in \mathbb{R}^d$ the following inequality holds, $L(\theta) \geq L(\theta') + \langle \nabla_\theta L(\theta'), \theta - \theta' \rangle + \frac{\mu}{2}\|\theta - \theta'\|^2$. Next, the function $\nabla_\theta L$ is C_∇-Lipschitz continuous, i.e., there is a constant $C_\nabla > 0$ such that for all $\theta, \theta' \in \mathbb{R}^d$, $\|\nabla_\theta L(\theta) - \nabla_\theta L(\theta')\|^2 \leq C_\nabla^2 \|\theta - \theta'\|^2$. The *true* optimizer of L is denoted by $\theta^* = \arg\min_{\theta\in\mathbb{R}^d} L(\theta)$.

2 Stochastic Streaming Gradient

In this section, we consider the SSG method with streaming-batches arriving in constant and varying streams. Our aim is to provide bounds on the quadratic mean $\mathbb{E}[\|\theta_t - \theta^*\|^2]$, which depends explicitly upon the problem's parameters.

Assumption 1. *For each $t \geq 1$, the random variable $\nabla_\theta l_{t,i}(\theta)$ is square-integrable and $\mathbb{E}[\nabla_\theta l_{t,i}(\theta)] = \nabla_\theta L(\theta)$ for all $\theta \in \mathbb{R}^d$ and $i = 1, \ldots, n_t$.*

Assumption 2-p (C_l-expected smoothness)**.** *The functions $(l_{t,i})$ is almost surely (a.s. for short) differentiable, and for $p \geq 1$, there exists $C_l > 0$ such that $\forall t \geq 1$, $\mathbb{E}[\|\nabla_\theta l_{t,i}(\theta) - \nabla_\theta l_{t,i}(\theta')\|^p] \leq C_l^p \mathbb{E}[\|\theta - \theta'\|^p]$ for all $\theta, \theta' \in \mathbb{R}^d$ and $i = 1, \ldots, n_t$.*

Assumption 3-p (σ-finite gradient noise)**.** *For $p \geq 1$, there exists $\sigma > 0$ such that $\forall t \geq 1$, $\mathbb{E}[\|\nabla_\theta l_{t,i}(\theta^*)\|^p] \leq \sigma^p$ with $i = 1, \ldots, n_t$.*

These assumptions are modified versions of the standard assumptions for stochastic approximations (Kushner and Yin, 2003; Bach and Moulines, 2011). Our setup include classic examples such as stochastic approximation (Robbins-Monro setting) and learning from i.i.d. data, such as linear regression or general ridge regressions.

Throughout this paper, we will consider learning rates on the form $\gamma_t = C_\gamma n_t^\beta t^{-\alpha}$ with hyper-parameters $C_\gamma > 0$, $\beta \in [0, 1]$, and α chosen accordingly to the expected streaming-batches denoted by n_t. We start by considering constant streaming-batches (i.e., mini-batch SG) where n_t follows the constant streaming-batch size $C_\rho \in \mathbb{N}$:

Theorem 1 (SSG, constant streaming-batches)**.** *Suppose $\gamma_t = C_\gamma C_\rho^\beta t^{-\alpha}$ such that $\alpha \in (1/2, 1)$. Under Assumptions 1 to 3-p with $p = 2$, we have*

$$\delta_t \leq \exp\left(-\frac{\mu C_\gamma N_t^{1-\alpha}}{2^{1-\alpha} C_\rho^{1-\alpha-\beta}}\right) \pi_\infty^c + \frac{2^{1+\alpha}\sigma^2 C_\gamma}{\mu C_\rho^{1-\alpha-\beta} N_t^\alpha}, \tag{3}$$

where π_∞^c is a finite constant depending on the model parameters.

The bound in Theorem 1 depends on the initial condition δ_0 and the variance σ^2 in the noise term. The initial condition δ_0 vanish sub-exponentially fast for $\alpha \in (1/2, 1)$. Thus, the asymptotic term is $2^{1+\alpha}\sigma^2 C_\gamma / \mu C_\rho^{1-\alpha-\beta} N_t^\alpha$, i.e., $\delta_t = \mathcal{O}(N_t^{-\alpha})$. This matches the findings made by Bach and Moulines (2011) for $C_\rho = 1$. Obviously, the hyper-parameter β only comes into play if the streaming-batch size C_ρ is larger than one. The asymptotic term is divided by $C_\rho^{1-\alpha-\beta}$, implying we could achieve noise reduction by taking $\alpha + \beta \leq 1$.

However, fixed-sized streaming-batches are not the most realistic streaming setting. These streaming-batches are far more likely to vary in size depending on the data streams. For the sake of simplicity, we consider varying streaming-batches where n_t are on the form $C_\rho t^\rho$ with $C_\rho \geq 1$ and $\rho \in (-1, 1)$ such that $n_t \geq 1$ for all t. We will refer to ρ as the *streaming rate*. For the convenience of notation, let $\tilde{\rho} = \rho \mathbb{1}_{\{\rho \geq 0\}}$.

Theorem 2 (SSG, varying streaming-batches)**.** *Suppose* $\gamma_t = C_\gamma n_t^\beta t^{-\alpha}$ *where* $n_t = C_\rho t^\rho$ *with* $C_\rho \geq 1$ *and* $\rho \in (-1, 1)$*, such that* $\alpha - \beta\tilde{\rho} \in (1/2, 1)$*. Under Assumptions 1 to 3-p with* $p = 2$*, we have*

$$\delta_t \leq \exp\left(-\frac{\mu C_\gamma N_t^{1-\phi}}{2^{(2+\rho)(1-\phi)} C_\rho^{1-\beta-\phi}}\right)\pi_\infty^v + \frac{2^{1+(2+\rho)\phi}\sigma^2 C_\gamma}{\mu C_\rho^{(1-\beta)\mathbb{1}_{\{\rho\geq 0\}}-\phi} N_t^\phi}, \tag{4}$$

where $\phi = ((1-\beta)\tilde{\rho} + \alpha)/(1+\tilde{\rho})$ *and* π_∞^v *a finite constant depending on the model parameters.*

As mentioned for Theorem 1, the condition of having $\alpha - \beta\tilde{\rho} \in (1/2, 1)$ is a natural restriction coming from Robbins and Monro (1951), which relaxes the usual condition of having $\alpha \in (1/2, 1)$ for ρ non-negative. Moreover, for $\rho \geq 0$, the sub-exponential and asymptotic term is scaled by $C_\rho^{1-\beta-\phi}$, implying we should take $\alpha + \beta \leq 1$ for having $1-\beta-\phi$ positive. We could accelerate the convergence by, e.g., setting $\alpha = 1$ and $\beta = 1/2$, whatever the streaming-batch rate $\rho \in (0, 1)$, would give us the rate of convergence $\delta_t = \mathcal{O}(N_t^{-(1+\rho/2)/(1+\rho)})$. Likewise, setting $\alpha = 2/3$ and $\beta = 1/3$ would give convergence rate $\delta_t = \mathcal{O}(N_t^{-2/3})$ for any $\rho \in (-1, 1/2)$.

3 Averaged Stochastic Streaming Gradient

In what follows, we consider the averaging estimate $\bar{\theta}_n$ given in (2) derived with use of (θ_t) from (1). Besides having Assumptions 2-p and 3-p to hold for $p = 4$ to bound the fourth-order moment of (1), an additional assumption is needed for bounding the *rest* term of the averaging estimate ASSG. Note that the operator Σ always exists when σ is finite for order $p = 4$ in Assumption 3-p.

Assumption 4. *The function* L *is twice differentiable with Lipschitz-continuous Hessian operator* $\nabla_\theta^2 L$*, and Lipschitz constant* $C_\delta > 0$*, that is, for all* $\theta, \theta' \in \mathbb{R}^d$*,* $\|\nabla_\theta L(\theta) - \nabla_\theta^2 L(\theta')(\theta - \theta')\| \leq C_\delta \|\theta - \theta'\|^2$*. Moreover, there exists a non-negative self-adjoint operator* Σ *such that* $\forall t \geq 1$*,* $\mathbb{E}[\nabla_\theta l_{t,i}(\theta^*)\nabla_\theta l_{t,i}(\theta^*)^\top] \preceq \Sigma$*.*

To ease notation, we make use of the functions $\psi_x^y(t) : \mathbb{R} \to \mathbb{R}$, given as

$$\psi_x^y(t) = \begin{cases} t^{(1-x)/(1+y)}/(1-x) & \text{if } x < 1, \\ (1+y)\log(t) & \text{if } x = 1, \\ x/(x-1) & \text{if } x > 1, \end{cases}$$

with $y \in \mathbb{R}_+$. We will use that $\psi_x^y(t)/t = \mathcal{O}(t^{-(x+y)/(1+y)})$ if $x < 1$, $\psi_x^y(t)/t = \mathcal{O}(\log(t)t^{-1})$ if $x = 1$, and $\psi_x^y(t)/t = \mathcal{O}(t^{-1})$ if $x > 1$.

Theorem 3 (ASSG, constant streaming-batches). *Suppose $\gamma_t = C_\gamma C_\rho^\beta t^{-\alpha}$ such that $\alpha \in (1/2, 1)$. Under Assumptions 1 to 4 with $p = 4$, we have*

$$\bar{\delta}_t^{1/2} \leq \frac{\Lambda^{1/2}}{N_t^{1/2}} + \frac{6\sigma C_\rho^{(1-\alpha-\beta)/2}}{\mu^{3/2} C_\gamma^{1/2} N_t^{1-\alpha/2}} + \frac{2^\alpha 6 C_\delta \sigma^2 C_\gamma}{\mu^2 C_\rho^{1-\alpha-\beta} N_t^\alpha} + \frac{2 C_l \sigma C_\gamma^{1/2}}{\mu^{3/2} C_\rho^{(1-\alpha-\beta)/2} N_t^{(1+\alpha)/2}} + \frac{C_\rho \Gamma_c}{\mu N_t}$$
$$+ \frac{C_\rho^{2-\alpha-\beta} \sqrt{\pi_\infty^c} A_\infty^c}{\mu C_\gamma N_t^{2-\alpha}} + \frac{(6 + 7\mathbb{1}_{\{C_\rho > 1\}}) 2^{3\alpha/2} C_\delta \sigma^2 C_\gamma^{3/2} C_\rho^{3\beta/2} \psi_{3\alpha/2}^{\tilde{\rho}}(N_t / C_\rho)}{\mu^{3/2} N_t},$$

with $\Lambda = \mathrm{Tr}(\nabla_\theta^2 L(\theta^*)^{-1} \Sigma \nabla_\theta^2 L(\theta^*)^{-1})$ and Γ_c consisting of the finite constants π_∞^c, Π_∞^c and A_∞^c, that only depends on the model parameters.

As noticed in Polyak and Juditsky (1992), the leading term Λ/N_t achieves the Cramer-Rao bound (Murata and Amari, 1999; Gadat and Panloup, 2017). Note that the leading term Λ/N_t is invariant of the learning rate. The two subsequent terms are the main remaining terms decaying at the rate $\mathcal{O}(N_t^{\alpha-2})$ and $\mathcal{O}(N_t^{-2\alpha})$, which suggests setting $\alpha = 2/3$. The remaining terms are negligible. Moreover, taking $\alpha = 2/3$ and $\beta \leq 1/3$ would be an *optimal* choice of hyper-parameters such that the streaming-batch size C_ρ have a positive or no impact.

Theorem 4 (ASSG, varying streaming-batches). *Suppose $\gamma_t = C_\gamma n_t^\beta t^{-\alpha}$ where $n_t = C_\rho t^\rho$ with $C_\rho \geq 1$ and $\rho \in (-1, 1)$, such that $\alpha - \beta\tilde{\rho} \in (1/2, 1)$. Under Assumptions 1 to 4 with $p = 4$, we have*

$$\bar{\delta}_t^{1/2} \leq \frac{\Lambda^{1/2}}{N_t^{1/2}} + \frac{2^{3+\phi(1+\tilde{\rho})} \sigma C_\rho^{(1-\phi-\beta)/2\mathbb{1}_{\{\rho \geq 0\}}}}{\mu^{3/2} C_\gamma^{1/2} N_t^{1-\phi/2}} + \frac{2^{(1+\phi)(1+\tilde{\rho})-2} C_\delta \sigma^2 C_\gamma}{\mu^2 C_\rho^{1-\phi-\beta} N_t^\phi} + \frac{2^{\phi(1+\tilde{\rho})/2} C_l \sigma C_\gamma^{1/2}}{\mu^{3/2} C_\rho^{(1-\phi-\beta)/2\mathbb{1}_{\{\rho \geq 0\}}} N_t^{(1+\phi)/2}}$$
$$+ \frac{C_\rho \Gamma_v}{\mu N_t} + \frac{C_\rho^{2-\phi-\beta} \sqrt{\pi_\infty^v} A_\infty^v}{\mu C_\gamma N_t^{2-\phi}} + \frac{2^{3(1+\phi)(1+\tilde{\rho})/2} C_\delta \sigma^2 C_\gamma^{3/2} C_\rho^{1+3\beta/2} \psi_{3(\alpha-\beta\tilde{\rho})/2}^{\tilde{\rho}}(N_t / C_\rho)}{\mu^{3/2} C_\rho^{\mathbb{1}_{\{\rho \geq 0\}}} N_t},$$

with $\Lambda = \mathrm{Tr}(\nabla_\theta^2 L(\theta^*)^{-1} \Sigma \nabla_\theta^2 L(\theta^*)^{-1})$ and Γ_v consisting of the finite constants π_∞^v, Π_∞^v and A_∞^v, that only depends on the model parameters.

Following the arguments above, the two main remainder terms reveal that $\phi = 2/3 \Leftrightarrow \alpha - \beta\tilde{\rho} = (2 - \tilde{\rho})/3$, e.g., by setting $\beta = 0$, we should pick $\alpha = (2 - \tilde{\rho})/3$. Likewise, if $\rho = 0$, we yield the same conclusion as in Theorem 3, namely $\alpha = 2/3$. However, these hyper-parameter choices are not resilient against any arrival schedule ρ. Nonetheless, we can *robustly* achieve $\phi = 2/3$ for any $\rho \in (-1, 1)$ by setting $\alpha = 2/3$ and $\beta = 1/3$. In other words, we can achieve *optimal* convergence for any data stream we may encounter by having $\alpha = 2/3$ and $\beta = 1/3$.

4 Conclusions

We introduced and studied the convergence rates of the stochastic streaming algorithms in a non-asymptotic manner. The theoretical results and our experiments showed a noticeable improvement in the convergence rate by choosing the learning rate according to the expected data streams. For ASSG, we showed that this choice of learning rate led to optimal convergence rates and was robust to any data stream rate we may encounter. Moreover, in large-scale learning problems, we know how to accelerate convergence and reduce noise through the learning rate and the treatment pattern of the data.

References

Bach, F., Moulines, E., 2011. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. Advances in Neural Information Processing Systems 24, 451–459.

Bottou, L., Curtis, F.E., Nocedal, J., 2018. Optimization methods for large-scale machine learning. SIAM Review 60, 223–311.

Bottou, L., LeCun, Y., 2004. Large scale online learning. Advances in Neural Information Processing Systems 16, 217–224.

Gadat, S., Panloup, F., 2017. Optimal non-asymptotic bound of the ruppert-polyak averaging without strong convexity. arXiv:1709.03342 .

Kushner, H.J., Yin, G.G., 2003. Stochastic Approximation and Recursive Algorithms and Applications. Springer-Verlag.

Murata, N., Amari, S., 1999. Statistical analysis of learning dynamics. Signal Processing 74, 3–28.

Polyak, B., Juditsky, A., 1992. Acceleration of stochastic approximation by averaging. SIAM Journal on Control and Optimization 30, 838–855.

Robbins, H., Monro, S., 1951. A stochastic approximation method. Annals of Mathematical Statistics 22, 400–407.

Ruppert, D., 1988. Efficient Estimations from a Slowly Convergent Robbins-Monro Process. Cornell University Operations Research and Industrial Engineering.

Shalev-Shwartz, S., Singer, Y., Srebro, N., Cotter, A., 2011. Pegasos: Primal estimated sub-gradient solver for svm. Mathematical programming 127, 3–30.

Sridharan, K., Shalev-Shwartz, S., Srebro, N., 2008. Fast rates for regularized objectives. Advances in Neural Information Processing Systems 21, 1545–1552.

Xiao, L., 2010. Dual averaging methods for regularized stochastic learning and online optimization. Journal of Machine Learning Research 11, 2543–2596.

Zhang, T., 2004. Solving large scale linear prediction problems using stochastic gradient descent algorithms, in: Proceedings of the twenty-first international conference on Machine learning, p. 116.

Survie

Machine learning for survival data prediction: Application of the super learner on pseudo-observations

Ariane Cwiling [1] & Olivier Bouaziz [1] & Vittorio Perduca [1]

[1] *Université de Paris, MAP5, UMR 8145, France*
ariane.cwiling@u-paris.fr *olivier.bouaziz@parisdescartes.fr*
vittorio.perduca@parisdescartes.fr

Résumé. La moyenne restreinte du temps de survie ("restricted mean survival time" ou RMST) est aisément interprétable, ce qui en fait un objet d'étude intéressant en analyse de survie. Sa prédiction par rapport aux caractéristiques d'un patient peut être très utile dans le domaine de la santé. Cependant peu de méthodes traitent de cette question en analyse de survie. Un article récent de Zhao (2021) propose d'appliquer un réseau de neurones profonds sur des pseudo-observations. Ces dernières peuvent être décrites comme une transformation des temps censurés en données pouvant être gérées comme non censurées. Dans ce travail, nous proposons une nouvelle méthode de prédiction pour le RMST basée sur les pseudo-observations et combinée avec le super learner, un algorithme de prédiction qui propose une combinaison pondérée optimale de différents algorithmes d'apprentissage. Nous avons évalué notre modèle à l'aide de simulations approfondies.

Mots-clés. RMST, pseudo-observations, super learner, prédiction.

Abstract. Because of its simple interpretation, the restricted mean survival time (RMST) is an interesting quantity of interest in survival analysis. Its prediction with regard to the attributes of a patient can be of great interest in health care. However few survival methods exist in practice for this purpose. To achieve this goal, a recent article applied a deep neural network on pseudo-observations, the latter being a transformation of the incomplete observed times into data that can be handled as uncensored (Zhao, 2021). In this work, we propose a new prediction model for the RMST based on pseudo-observations combined with the super learner, a prediction algorithm which fits a weighted combination of candidate learners. We evaluated our model through extensive simulations.

Keywords. RMST, pseudo-observations, super learner, prediction.

1 Introduction

Survival analysis has proven effective in drawing inference on long term studies where data are incompletely observed because of censoring. Survival analysis provides robust methods adapted to censored data, among which a majority aims at estimating outcomes

of interest such as the survival function and at making inference about the relative impact of covariates on the hazard rate. However, apart from the Cox model, few methods are used in practice for prediction purposes. In this work we focus on the prediction of the restricted mean survival time (RMST) defined as the expectation of the minimum between the time to event and a fixed time provided by the user. The RMST is of great clinical interest because of its interpretability and the purpose of this paper is to develop a new method for its prediction. Most classic survival methods for estimation or prediction fit a model directly on censored data. We adopt an alternative approach based on pseudo-observations to transform the incomplete observed times into data that can be handled as uncensored (Andersen and Perme, 2010). It is then possible to apply any standard prediction method to link these transformed outcomes with covariates. For instance, a recent article applied a deep neural network on pseudo-observations to this aim (Zhao, 2021). However data scientists are unlikely to know a priori which prediction algorithm will perform best on a specific dataset, this is why methods have been developed to select the optimal algorithm among a set of candidate learners. Yet theory showed that the super learner, fitting a weighted combination of these learners, provides at least equally performing predictions as the best candidates (Van der Laan et al., 2007). From these theoretical results we build a regression model for the prediction of the RMST, by computing pseudo-observations and feeding them into a super learner algorithm with a large library of candidate learners. We study the performance of this method through extensive simulations. Importantly, we focus on performance measures directly on censored data so as to take into account both the quality of the super learner prediction and the approximation made by the pseudo-observations.

2 Method and evaluation tools

2.1 Method

Survival analysis aims at analysing long term studies where data are often incomplete because of censoring. Only right-censoring has been considered here. Let Z_i denote the vector of covariates that describes subject i, and T_i^* the survival time. We assume that we only observe the minimum of T_i^* and C_i, where C_i represents a censoring time. We make the assumption that censoring is independent from time to event given the covariates. We define the follow-up time $T_i = \min(T_i^*, C_i)$, and the censoring indicator $\Delta_i = \mathbb{1}(T_i^* \leq C_i)$. We are interested in the RMST, which we define as the expectation of the minimum between the time to event and a fixed time τ: $\mu_\tau^* = \mathbb{E}[T^* \wedge \tau]$. In particular, we want to predict the conditional RMST: $\mu_\tau(Z) = \mathbb{E}[T^* \wedge \tau \mid Z]$.

Standard models and methods are applied directly to censored data. Alternatively, using the pseudo-observations approaches developed by Andersen and Perme (2010), we can transform censored data into data that can be handled as uncensored and apply classic prediction methods. Pseudo-observations aim at estimating the mean of a transformation

f of the survival time, which we denote $\theta := \mathbb{E}[f(T^*)]$. To estimate the RMST we set $f(T^*) = T^* \wedge \tau$, so that $\theta = \mathbb{E}[T^* \wedge \tau] = \mu_\tau^*$. Let $\hat{S}$ be the Kaplan-Meier estimator. Pseudo-observations, derived from jackknife methodology, are the $\hat{\mu}_{\tau,i}$'s defined by

$$\hat{\mu}_{\tau,i} = n \int_0^\tau \hat{S}(t)dt - (n-1) \int_0^\tau \hat{S}^{-i}(t)dt.$$

where $\hat{S}$ is the Kaplan-Meier estimator using all n data and $\hat{S}^{-i}$ is the same estimator without the i-th subject. It follows from Jacobsen and Martinussen (2016) that

$$\begin{cases} \hat{\mu}_{\tau,i} = \mu_\tau^* + \dot{\psi}(T_i, \Delta_i) + o_{\mathbb{P}}(1), \\ \mathbb{E}[\dot{\psi}(T_i, \Delta_i) \mid Z_i] = \mu_\tau(Z_i) - \mu_\tau^*. \end{cases}$$

The idea is then to replace the incompletely observed (because of censoring) $T_i^* \wedge \tau$ by $\hat{\mu}_{\tau,i}$ and regress them against the covariates. Any prediction algorithm can be used, such as neural networks (Zhao, 2021). The optimal method can be selected through cross-validation. However, one candidate algorithm may not be flexible enough to perform best on all data sets and a researcher is unlikely to know a priori which candidate learner will work best. A solution is then to fit a weighted combination of many candidate learners: this is the motivation behind the so called super learner. Figure 1 in Van der Laan et al. (2007) illustrates the steps of the algorithm in a flow diagram. Theoretical results in Van der Laan et al. (2007) show that such a super learner will perform asymptotically as well as or better than any of the candidate learners. The main advantage of the proposed super learner is its adaptivity to different data distributions across many studies. Golmakani and Polley (2020) adapted the super learner to survival analysis. This survival super learner is able to handle censored data and apply survival estimation methods. However, it is built on the Cox model and it requires the candidate learners to verify the proportional hazard assumption. By using pseudo-observations and classic super learner, we do not need to make this assumption and we can take advantage of a large variety of learners usually used in the context of uncensored data.

2.2 Evaluation tools

A close formula for the true value of the conditional RMST $\mu_\tau(Z)$ is not always available for simulation studies. When we can access it, the MSE is a simple way to evaluate the performance of an estimator. We assume that an estimator $\hat{\mu}_\tau$ is a function obtained from a train set $D_{\text{train}} = \{(T_i, \Delta_i, Z_i)\}_{i=1}^{n_{\text{train}}}$ independent from the testing set, whose covariates are denoted $\{Z_i\}_{i=1}^{n_{\text{test}}}$ (or Z for a vector of covariates in the general case). The MSE can be written

$$MSE(\hat{\mu}_\tau) = \mathbb{E}_Z\Big[\big(\mu_\tau(Z) - \hat{\mu}_\tau(Z)\big)^2 \;\Big|\; D_{\text{train}}\Big]$$

and is estimated on the testing set by

$$\widehat{MSE}(\hat{\mu}_\tau) = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \big(\mu_\tau(Z_i) - \hat{\mu}_\tau(Z_i)\big)^2.$$

However when a close formula is not available we can replace $\mu_\tau(Z_i) = \mathbb{E}[T_i^* \wedge \tau | Z_i]$ in the above formula by $T_i^* \wedge \tau$, which we have access to in simulations. In the following, we quantify the difference between these two approaches. We consider the model $T^* \wedge \tau = \mu_\tau(Z) + \epsilon$, where the random error term ϵ verifies $\mathbb{E}[\epsilon \mid Z] = 0$. By fixing the train set we get the following decomposition

$$\mathbb{E}_{Z,\epsilon}\Big[\big(T^* \wedge \tau - \hat{\mu}_\tau(Z)\big)^2 \;\Big|\; D_{\text{train}}\Big] = \mathbb{E}_Z\Big[\big(\mu_\tau(Z) - \hat{\mu}_\tau(Z)\big)^2 \;\Big|\; D_{\text{train}}\Big] + \mathbb{E}[\epsilon^2]$$

Then the difference between the two approaches is equal to the variance of the error term, which is constant and independent of the estimator. Note that when it comes to true data, we do not have access to the real event times for every subjects. This is why performance measures have been adapted to take censored values into account, such as the C-index (Heagerty and Zheng, 2005).

3 Simulation study

The model and evaluation methods previously described are illustrated in this section with two simulation studies. Each simulation model includes a tuning parameter, for instance the size of the training set (for the first simulation) or the type of generating model (for the second simulation). For each value of this parameter, we simulated $M = 10$ training sets on which to apply the model and evaluate performances. The performance values are obtained using one of the following two MSE's:

$$\widehat{MSE}_{\text{rmst}}(\hat{\mu}_\tau^m) = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \Big(\mu_\tau(Z_i) - \hat{\mu}_\tau^m(Z_i)\Big)^2,$$

$$\widehat{MSE}_{\text{min}}(\hat{\mu}_\tau^m) = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \Big(T_i^* \wedge \tau - \hat{\mu}_\tau^m(Z_i)\Big)^2,$$

with $m = 1, \ldots, M$. We represented the distribution of those MSE's (for $m = 1, \ldots, M$) using boxplots for each value of the tuning parameter. We applied the super learner using as a default learner the mean (that assigns to each instance the empirical average calculated from the whole sample), and as other predictors the linear model, elastic net, CART, random forest and one-layer neural network. We use this selection of candidate learners to estimate the RMST at the value of τ corresponding to the 80th percentile of the true event times distribution. Boxplots illustrate the performances of the super learner in comparison with each candidate learner trained alone. We also display the empirical distribution of the β coefficients, i.e. the weights of candidate learners in the super learner prediction.

3.1 First model

We adapted the simulation setting of Van der Laan et al. (2007), which uses a classic linear model including interactions. The varying parameter in this model is the size of the training set, $n \in \{100, 500, 1000\}$. A close formula of the RMST is available for this model, making possible the use of both MSE measures. Results in figure 1 show that these two measures are concordant. The visible difference between them, due to the error term described in section 2.2, appears constant. Furthermore, results show that the super learner performs at least as well as the best algorithm.

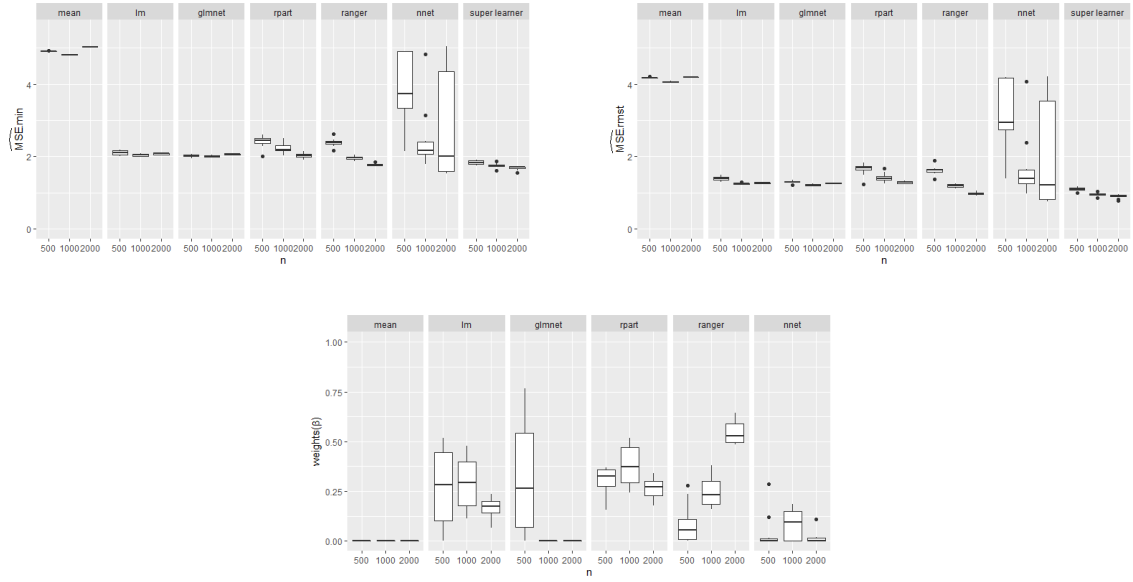


Figure 1: Prediction of the RMST with the super learner using a selection of libraries and trained on $n \in \{500, 1000, 2000\}$ data. Weights (β) on the candidate learners are also displayed.

3.2 Second model

We applied our method to the simulation model used in Golmakani and Polley (2020). It is a classic Cox model, and additional models are studied by adding quadratic or interaction effects. Here, the training sets have a fixed size. In this model we do not have access to the true RMST thus we can not use $\widehat{MSE}_{\text{rmst}}$ as indicator of performance. Results displayed in figure 2 show once again that the super learner provides a prediction accuracy very close to the best algorithms, if not better. It is also important to notice that the linear model and the elastic net are preferred to the other algorithms in this simulation setting. In the previous simulation, random forests or the CART algorithm showed higher weights in the super learner (at least for high values of n). This enlightens the flexibility of the super learner and its ability to combine all learners in an optimal way.

Figure 2: Prediction of the RMST with the super learner using a selection of libraries. The model can be a classic linear model or have higher order effects (quadratic, interaction). Weights (β) on the candidate learners are also displayed.

4 Conclusion

We described a new method for predicting the RMST of an individual given a set of covariates. We applied the super learner to pseudo-observations, considered as a substitute for the unobserved censored data. The undertaken simulations seemed promising concerning the validity of this approach. The proposed procedure shows performances that are very close to the best candidate algorithms, if not better. These empirical results are encouraging and warrant future theoretical developments: while it is known that the super learner on non-censored data performs at least as well as the candidate learners, it is an open question whether similar theoretical results hold for the performance of combining pseudo-observations with super learning.

Bibliographie

Andersen, Per Kragh and Maja Pohar Perme (2010), Pseudo-observations in survival analysis, *Statistical Methods in Medical Research*, 19, pp. 71-99.
Golmakani, Marzieh K. and Eric C. Polley (2020), Super Learner for Survival Data Prediction, *The International Journal of Biostatistics*, 16.2.
Heagerty, Patrick J. and Yingye Zheng (2005), Survival Model Predictive Accuracy and ROC Curves, *Biometrics* 61, pp. 92-105.
Jacobsen, Martin and Torben Martinussen (2016), A note on the large sample properties of estimators based on generalized linear models for correlated pseudo-observations, *Scandinavian Journal of Statistics*, 43.3, pp. 845-862.
Van der Laan, Mark J., Eric C. Polley, and Alan E. Hubbard (2007), Super Learner, *Statistical Applications in Genetics and Molecular Biology*, 6.1,25.
Zhao, Lili (2021), Deep neural networks for predicting restricted mean survival times, *Bioinformatics*, 36.24, 5672-5677.

Comparaisons de méthodes pour données de survie en grande dimension sur de petits échantillons : optimisation des hyperparamètres et validation.

Audrey Lavenu[1], Juliette Murris[2*], Alexis Mareau[3], Timothé Rouzé[4], Magalie Fromont[5], Valérie Gares[6] & Sandrine Katsahian[7]

[1] *CIC1414 Inserm, IRMAR, Université de Rennes 1. audrey.lavenu@univ-rennes1.fr,*
[2,3,4,7] *CIC1418 Inserm, HeKA-Inria, Université de Paris. juliette.murris-ext@aphp.fr, alexis.mareau5@gmail.com, timothe.rouze1@gmail.com, sandrine.katsahian@aphp.fr,*
[5] *IRMAR, Université de Rennes 2. magalie.fromont@univ-rennes2.fr,*
[6] *IRMAR, INSA, Rennes, France. valerie.gares@insa-rennes.fr,*
** Subvention de l'ANRT et Pierre Fabre (CIFRE 2020/1701).*

Résumé. Avec l'augmentation du nombre de données sur les patients dans les domaines de l'imagerie médicale ou encore de la génomique, les méthodes d'analyses classiques sont souvent inadéquates dans les cas où il y a moins d'observations que de variables. L'objectif de notre travail est d'étudier différents critères de performance et leur estimation de la méthode Cox Boost pour analyser des données de survie en grande dimension sur petits échantillons. Nous nous intéressons à la prédiction et la discrimination des variables pronostiques, mais aussi à la "tunabilité" du modèle Cox Boost (gain par optimisation des hyperparamètres). Nous simulons les temps de survie avec une loi exponentielle et les temps de censure avec une loi uniforme. Pour fixer le taux de censure à un taux prédéfini, nous montrons comment calculer le paramètre de la distribution de censure. Avec un schéma de simulation faisant varier : la taille d'effet des covariables, la taille de l'échantillon, et le taux de variables actives, nous comparons différents critères de performance de la méthode (C de Harrell et mesure d'importance de variable) estimés par validation croisée à 2 et 5 blocs, avec trois méthodes de choix des hyperparamètres. Nous montrons la difficulté d'optimiser les hyperparamètres pour de petits échantillons, et les lacunes des mesures d'importance des variables à détecter les variables simulées actives, même quand la performance du modèle en termes de prédiction est correcte.
Mots-clés. Méthodes d'apprentissage supervisé, Survie, Grande dimension.

Abstract. With an increasing amount of patient data in the fields of medical imaging and genomics, conventional analysis methods are often inadequate for modeling cases with fewer samples than features. The objective of the present work is to study different performance criteria of the Cox Boost method for analysing high-dimensional survival data on small sample sizes. Here we focus on prediction and discrimination of prognostic features, as well as the tunability of the model (potential score increase by optimization of hyperparameters). We simulate survival times with exponential distribution, and censoring times with uniform distribution. To fix the censoring rate at a predefined value, we show how to derive the parameter of the censoring distribution. The simulation scheme modifies the effect size of the covariates, the sample size, and the rate of active features. We compare

the different performance criteria (Harrell's C and variable importance measure) by 2 and 5-fold cross-validation, with three ways of selecting the hyperparameters. We show the complexity of optimising hyperparameters for small sample size data. We also illustrate that the feature importance measure does not always allow detection of active simulated features, although the predictive performance of the model is correct.
Keywords. Supervised machine learning, Survival, High-dimensional.

1 Introduction

L'utilisation des mégadonnées dans la recherche et la pratique en santé publique nécessite de nouvelles compétences pour gérer et analyser ces données. La particularité des données de survie est principalement la présence de réponses éventuellement censurées. Sur la base d'une revue de la littérature des méthodes statistiques couramment utilisées et des techniques d'apprentissage automatique développées pour l'analyse de survie, une taxonomie détaillée des méthodes existantes a été proposée (Wang, 2019).

Dans les études pronostiques, nous avons deux objectifs : la prédiction (c'est-à-dire la précision de l'estimation du risque) et la détection des facteurs de risque (c'est-à-dire l'interprétation des résultats en fonction de l'importance des covariables). Cette dernière est nécessaire pour développer un meilleur diagnostic et des stratégies de traitement optimales (Pittman, 2004). Ces deux objectifs sont rarement étudiés simultanément.

Tous les algorithmes d'apprentissage impliquent un plus ou moins grand nombre d'hyperparamètres (ou "paramètres de réglage"). La quantité de gain de performance qui peut être obtenue en optimisant les hyperparamètres par rapport aux valeurs par défaut définit la "tunabilité" de ces algorithmes ou modèles. Il a été montré que la méthode d'optimisation des hyperparamètres et le moment où elle se fait peut engendrer des biais sur les petits échantillons, par exemple un fort taux de bien classés d'une variable binaire alors que les covariables étaient simulées sans effet (Vabalas, 2019).

L'objectif de notre travail est d'étudier les performances des méthodes pour analyser des données de survie de grande dimension, donc souvent sur petits échantillons avec beaucoup de variables. Nous étudierons ici les performances de la méthode Cox Boost (Binder, 2008) en termes de prédiction et de discrimination des variables pronostiques, mais aussi la tunabilité en fonction des tailles d'échantillon et du taux de covariables informatives.

2 Critère de performance en analyse de survie et critère d'importance des variables

En analyse de survie, nous notons $T_1, \ldots, T_n$ les temps auxquels surviennent l'évènement (ou plus précisément les durées d'apparition de l'évènement) pour n individus. Ces variables aléatoires sont supposées positives, indépendantes et de même densité f, telle que f est continue, bornée et strictement positive sur $\mathbb{R}^+$. La spécificité des analyses de survie est que nous ne pouvons pas observer $(T_1, ..., T_n)$ mais $(T_1^S, \delta_1), (T_2^S, \delta_2), \ldots, (T_n^S, \delta_n)$, où

$$\text{pour tout } 1 \leq i \leq n, \quad T_i^S = \min(T_i, C_i), \quad \delta_i = \mathbf{1}_{\{T_i \leq C_i\}},$$

et C_i est une variable aléatoire positive qui représente le temps de censure. Soit $X_i = (X_i^1, \ldots, X_i^p)$ un vecteur de p covariables.

Risque. Soit g : $\mathbb{R}^d \to \mathbb{R}$ une règle de décision. La performance d'une règle de décision est définie par le risque ou l'erreur de généralisation $\mathcal{R}(g)$. L'objectif est de construire une règle de décision qui soit aussi performante que la règle optimale $g^* \in \text{argmin}_g \mathcal{R}(g)$.
Le **C de Harrell** est le plus souvent utilisé en analyse de données de survie pour mesurer le pouvoir discriminant et la capacité prédictive des modèles. Le risque considéré ici est :

$$\mathcal{R}(g) = 1 - \mathbb{P}[M_i^g < M_j^g | T_i > T_j, T_j < \min(C_i, C_j)],$$

où M_i^g représente la probabilité d'avoir eu l'évènement d'intérêt à la fin de l'étude pour l'individu i prédite par la règle g. $\mathbb{P}[M_i^g < M_j^g | T_i > T_j, T_j < \min(C_i, C_j)]$ est estimée par le **C de Harrell** qui est une mesure de concordance :

$$\text{CIndex}^n(g) = \frac{\sum_{i=1}^n \sum_{j=1, i \neq j}^n \mathbf{1}_{\left\{T_i^S > T_j^S\right\}} . \mathbf{1}_{\left\{M_j^g > M_i^g\right\}} . \delta_j}{\sum_{j=1, i \neq j}^n \mathbf{1}_{\left\{T_i^S > T_j^S\right\}} \delta_j}.$$

Estimation du risque. Le risque estimé sur les données servant à l'apprentissage conduit à une sous-estimation du risque. Plusieurs solutions existent pour construire des estimations sans biais du risque telles que le partage de l'échantillon en deux (apprentissage/test) ou validation croisée holdout, où le risque est alors estimé sur les données test. Une autre méthode utilisée est la validation croisée par blocs. Sur de petits échantillons, la première est difficile à envisager, et les méthodes de validation croisée à K blocs sont alors privilégiées. On considère une partition $\{\mathcal{I}_1, \ldots \mathcal{I}_K\}$ de $\{1, \ldots, n\}$. Pour chaque $k = 1, \ldots, K$, la réglè g est construite avec l'échantillon d'aprentissage $\left\{(X_i, T_i^S, \delta_i), i \in \{1, \ldots, n\} \setminus \mathcal{I}_k\right\}$, on la note $\widehat{g}^{n,k}$. Le C de Harrell $\text{CIindex}_{n,k}(\widehat{g}^{n,k})$ est calculé sur l'échantillon test $\left\{(X_i, T_i^S, \delta_i), i \in \mathcal{I}_k\right\}$. On considère ensuite le risque moyen sur les K blocs $\mathcal{R}_n(g) = 1 - \frac{1}{K} \sum_{k=1}^K \text{CIindex}_{n,k}(\widehat{g}^{n,k})$. Plus le nombre de blocs augmente, plus la taille de l'échantillon d'apprentissage est grande, plus celle de l'échantillon test est petite, et dans le cas du "leave-one-out" avec un seul individu par bloc, le C de Harell ne

peut pas être calculé.

Choix des hyperparamètres. Les règles de décision (méthodes) peuvent dépendre d'hyperparamètres $h \in \mathcal{H}$. Soit $\mathcal{G} = \{g_h, h \in \mathcal{H}\}$, la famille des règles g dépendant de l'hyperparamètre h. L'objectif est de sélectionner la règle g_h qui minimise le risque estimé par validation croisée $\mathcal{R}_n(g_h)$. Nous avons choisi deux approches d'optimisation des hyperparamètres. La méthode HalvingGridSearch (Jamieson, 2016) consiste en l'évaluation à partir d'un sous-ensemble des données plutôt que sur toutes les données. A chaque étape du processus, les combinaisons d'hyperparamètres les plus performantes sont évaluées et la quantité de données utilisées augmente en se concentrant sur les meilleures combinaisons. On dit qu'il s'agit d'halving successifs. La méthode Optuna (Akiba, 2019) est un système d'optimisation automatique par recherche dynamique des hyperparamètres. A chaque itération, cet algorithme permet de tirer au hasard une combinaison d'hyperparamètres en se basant sur les performances de la meilleure combinaison des itérations précédentes, et stoppe les combinaisons les moins prometteuses.

Critère d'importance des variables. Pour mesurer l'importance de la variable j, on considère la partition $\{\mathcal{I}_1, \dots \mathcal{I}_K\}$. Pour chaque $k = 1, \dots, K$, $\mathcal{I}_k^j$ est construit à partir de l'échantillon $\mathcal{I}_k$ où les valeurs de la variable j sont permutées aléatoirement. La règle $\widehat{g}^{n,k}$ est définie précédemment. On calcule le C de Harrell sur l'échantillon dont les valeurs de la variable j sont perturbées $\left\{(X_i, T_i^S, \delta_i), i \in \mathcal{I}_k^j\right\}$ que l'on note $\text{CIindex}_{n,k}^j(\widehat{g}^{n,k})$. La mesure d'importance de la variable j est donnée par $\text{Imp}(j) = \frac{1}{K}\sum_{k=1}^{K}(\text{CIndex}_{n,k}(\widehat{g}^{n,k}) - \text{CIndex}_{n,k}^j(\widehat{g}^{n,k}))$.

3 Méthode CoxBoost et hyperparamètres associés

L'ajustement des modèles de survie par la vraisemblance partielle de Cox est la méthode la plus largement utilisée pour la régression des risques proportionnels linéaires. Le modèle de Cox exprime la fonction de risque instantané de décès λ en fonction du temps t et des covariables $X_1, \cdots, X_n$. On a alors : $\lambda\left(t, X_1, \cdots, X_n\right) = \lambda_0\left(t\right) exp\left(\Sigma_{i=1}^n \beta_i X_i\right)$.
Le boosting s'adapte aussi aux modèles de régression non linéaires et aux modèles de Cox en construisant une famille d'estimateurs qui sont ensuite agrégés. Les 3 hyperparamètres que nous optimiserons pour cette méthode sont : le taux d'apprentissage (learning rate, la valeur par défaut est à 0,1), le nombre d'estimateurs (100 par défaut), et le taux d'abandon (dropout rate, 0 par défaut).

4 Simulation

Génération de données. Rappelons nos notations : nous considérons un échantillon de n individus, pour chaque inidividu i, soit X_i vecteur de p covariables, T_i temps de

survie, C_i temps de censure. Nous observons : $T_i^S = \min(T_i, C_i)$ et $\delta_i = \mathbf{1}_{T_i \leq C_i}$. Soit $X_i \sim \mathcal{N}_p(\mu, \Sigma)$ où $\mu = (a, \ldots, a)$ et Σ ayant une structure de corrélation autoregressive telle que $\text{corr}(X_i^k, X_i^{k'}) = \rho^{|k-k'|}$ ($\rho = 0.7$) avec $1 \leq i, j \leq p$. Basé sur un modèle de Cox avec composantes linéaires, nous associons les coefficients $\beta = (\beta_1, \ldots, \beta_p)$ aux p covariables. p' coefficients sont fixés à b (indicés $k_1, \ldots, k_{p'}$) et $p - p'$ sont fixés à 0. Le taux appelé $\text{sr} = \frac{p'}{p}$ (sparse rate) est alors le taux de variables actives. Prenons T un temps d'évènement suivant une distribution exponentielle $\mathcal{E}(\lambda_x)$ où $\lambda_x = \alpha \exp\left(\sum_{j=1}^{p'} b x_i^{k_j}\right)$. Si nous fixons m_0 le temps de survie médian de base souhaité et m_1 celui pour une augmentation de une unité d'une covariable active, nous pouvons alors calculer $b = \ln\left(\frac{m_0}{m_1}\right)$ et $\alpha = \frac{\ln(2)}{m_0}$. Nous supposons que la censure C suit une distribution uniforme sur $[0, \theta]$, $C \sim \mathcal{U}([0, \theta])$ et est indépendante de T conditionnellement à X. Soit $\tau = \mathbb{P}(T < C) = \mathbb{P}(\delta = 1)$ le taux de censure, θ est solution de l'équation adaptée de la méthode de Wan (2017) :

$$\gamma(\theta) = \int_0^{+\infty} \frac{1}{\theta u} \exp(-u\theta) \frac{1}{u\sqrt{2\pi}\sqrt{(\beta)^T \Sigma \beta}} \exp\left(-\frac{(\ln(u) - \ln(\alpha) - p'ba)^2}{2(\beta)^T \Sigma \beta}\right) du - \tau = 0.$$

Schémas de simulation. Pour chaque scénario, nous simulerons $M = 100$ jeux de données avec un taux de censure de $\tau = 50\%$. Nous choisirons $a = 0$ et $m_0 = 10$. Nous ferons varier : le gain de médiane pour l'augmentation de une unité : $m_1 - m_0 \in \{0.5, 1\}$; la taille de l'échantillon : $n \in \{50, 100, 150, 200, 500\}$; et le taux de variables actives : $\text{sr} \in \{0, 0.25, 0.5\}$.

Evaluation de la méthode. Pour les M jeux de données, nous estimerons le C de Harrell de la méthode CoxBoost par la validation croisée à 2 et 5 blocs (par la moyenne des C de Harrell dans chaque bloc). Nous évaluerons trois méthodes de choix des hyperparamètres (sans optimisation donc avec les valeurs par défaut, et les deux méthodes précédemment citées : HalvingGridSearch et Optuna).

Sur des jeux de données simulés, nous connaissons le vrai statut actif ou passif des p covariables, nous notons Y la variable aléatoire associée. Soit $Y^j = 1$ si la variable X^j est active (et a un effet sur la durée de vie) et $Y^j = 0$ sinon. Pour $s \in [0, 1]$ on définit :

$$\widehat{x}(s) = \frac{\sum_{j=1}^p \mathbf{1}_{\{\text{Imp}(j) > s\}}(1 - Y^j)}{\sum_{j=1}^p (1 - Y^j)} \text{ et } \widehat{y}(s) = \frac{\sum_{j=1}^p \mathbf{1}_{\{\text{Imp}(j) \geq s\}} Y^j}{\sum_{j=1}^p Y^j}, \quad \forall s \in [0, 1].$$

Nous appelons AUC^{Imp} l'aire sous la courbe paramétrée définie par $(\widehat{x}(s), \widehat{y}(s)), \forall s \in [0, 1]$. Si AUC^{Imp} est proche de 1, nous conclurons que la méthode de prédiction aura plutôt utilisé les variables actives sur ce jeu de données.

En notant $j_1, \ldots, j_p$ les indices des covariables tels que $\text{Imp}_{j_1} \geq \ldots \geq \text{Imp}_{j_p}$ on calculera en outre $\sum_{k=1}^{p'} \mathbf{1}_{\left\{Y^{j_k} = 1\right\}}$ le nombre de variables actives dans les p' variables dont la mesure d'importance est la plus grande, avec $p' = p.sr$ le nombre de variables simulées actives.

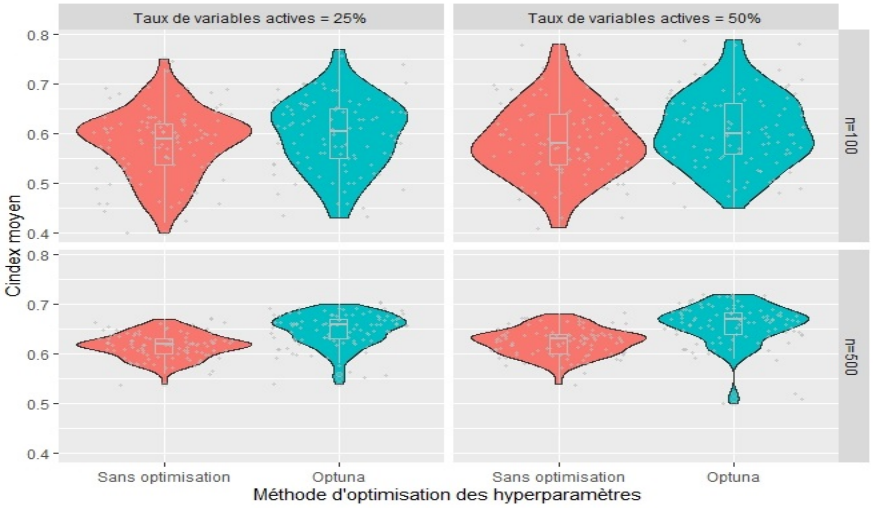


FIGURE 1 – Distribution des CIndex moyens sur les 5 blocs de la validation croisée pour les $M = 100$ jeux de données simulées, $n \in \{100, 500\}$, $m_1 - m_0 = 1$, sr $\in \{0.25, 0.5\}$.

Résultats. La figure 1 montre un exemple de simulations où $n \in \{100, 500\}$, $m_1 - m_0 = 1$, sr $\in \{0.25, 0.5\}$. Les validations croisées sont réalisées en 5 blocs et le choix des hyperparamètres se fait soit sans optimisation soit avec Optuna. L'optimisation des hyperparamètres est efficace pour $n = 500$ avec un C de Harrell estimé au dessus de 0.65.
Nous montrerons la difficulté d'optimiser les hyperparamètres pour de petits échantillons, et que l'importance des variables dans le modèle utilisé ne permet pas toujours de détecter les variables simulées actives, même quand la performance du modèle en termes de prédiction est correcte.

Bibliographie

Akiba T et al. (2019). Optuna : A next-generation hyperparameter optimization framework. *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery and data mining.*
Binder H and Schumacher M (2008). Allowing for mandatory covariates in boosting estimation of sparse high-dimensional survival models. *BMC Bioinformatics.* 9 :14.
Jamieson K and Talwalkar A (2016). Non-stochastic best arm identification and hyperparameter optimization. *Artificial intelligence and statistics.* 240-248.
Pittman J, et al. (2004). Integrated modeling of clinical and gene expression information for personalized prediction of disease outcomes. *Proc Natl Acad Sci USA.* 101(22) :8431-6.
Vabalas A, Gowen E, Poliakoff E, Casson A. (2019). Machine learning algorithm validation with a limited sample size. *PLoS ONE.* 14(11) :e0224365
Wan F. (2016). Simulating survival data with predefined censoring rates for proportional hazards models. *Stat Med.* 36(5) :838-854.
Wang P, Li Y, Reddy C.K. (2017). Machine Learning for Survival Analysis : A Survey. *ACM Computing Surveys* 51(6)

Modèle de Cox avec des données hétérogènes

Eliz Peyraud [1,2], Julien Jacques [1], Guillaume Metzler [1] et Alexandre Lopez [2]

[1] *Univ Lyon, Univ Lyon 2,*
ERIC UR 3083, Lyon, France
{eliz.peyraud,julien.jacques,guillaume.metzler}@univ-lyon2.fr.
[2] *Institut Georges Lopez (IGL)*
Lissieu, France
{epeyraud,alopez}@igl-transplantation.com

Résumé. Dans un contexte médical, l'analyse de survie se fait essentiellement à l'aide de modèles à risques proportionnels de Cox. Le présent papier cherche à construire un tel modèle afin de prédire la probabilité de survie d'un patient à un temps t après un acte chirurgical lourd. Nous mettons en évidence la nécessité de prendre en compte l'hétérogénéité des patients lors de la modélisation et expliquons comment nous pouvons prendre en compte des variables explicatives de natures variées.

Mots-clés. Modèle de Cox, Analyse de survie, Données médicales, Données hétérogènes.

Abstract. In a medical context, survival analysis is mainly done using Cox proportional hazards models. The present paper seeks to build such a model in order to predict the probability of survival of a patient at a time t after a heavy surgical procedure. We highlight the need to take into account the heterogeneity of the patients during the modeling and explain how we can take into account explanatory variables of various natures.

Keywords. Cox Proportional Hazards Model, Survival Analysis, Medical Data, Heterogeneous Data.

1 Introduction

L'analyse de survie est une méthode couramment utilisée dans le domaine médical pour, entre autre, estimer la durée de vie d'un patient suite à une opération lourde ou à la prise d'un traitement conséquent. Elle peut également se retrouver pour étudier la durée de vie d'un composant suite à une modification effectuée sur ce dernier. Dans un contexte exclusivement médical, ces modèles sont importants pour guider les médecins dans le choix des traitements à prescrire en estimant le taux de survie du patient avec la prise en compte de ce nouveau facteur.

A ce titre, les modèles de Cox [6] constituent les principaux outils de modélisation de ce phénomène via la prise en compte d'informations relatives aux patients. En revanche, l'hétérogénéité des profils de patients (représentés par les différentes covariables) peut souvent mettre à mal l'efficacité de tels modèles.

Dans cette étude, nous montrons la nécessité de prendre en compte cette hétérogénéité des données dans l'apprentissage d'un modèle de Cox. Plus précisément, nous montrons que les courbes de survie peuvent fortement varier selon la typologie du patient (*e.g.* distinction homme/femme, antécédant médicaux, origine ethnique, ...). Enfin, nous montrons également, sur un ensemble restreint de descripteurs, que leur significativité peut dépendre de cette typologie de patients.

Ces observations sont à l'origine de la perspective présentée à la fin de ce travail.

2 Analyse de Survie

On cherche à analyser la probabilité de survie du patient après un évènement ayant eu lieu à un temps t_0 (*e.g.* un acte chirurgical lourd). Cela passe par l'étude d'une variable aléatoire T qui représente la *durée de survie* du patient à partir de l'évènement qui s'est produit à t_0. On définit alors la *fonction de survie* d'un individu X, notée $S(t \mid X)$, comme la probabilité que le temps de survie du patient soit supérieur à $t > 0$:

$$S(t \mid X) = \mathbb{P}(T \geq t \mid X),$$

où X représente l'ensemble des covariables de l'individu concerné.

Deux approches historiques permettent de représenter cette fonction de survie. Une première, non-paramétrique, basée sur la courbe de Kaplan-Meier et qui ne permet pas de prendre en compte des covariables, et une deuxième, paramétrique cette fois, reposant sur les modèles de Cox.

2.1 Approche non paramétrique : Kaplan-Meier

Une première estimation non-paramétrique de la fonction de survie $S(t)$ est donnée par l'estimateur de Kaplan-Meier :

$$\hat{S}(t) = \prod_{t_i < t} \frac{n_i - d_i}{n_i}$$

où t_i ($i \geq 1$) représente les instants où surviennent des évenements (une défaillance ou une censure), d_i représente le nombre d'évènements qui se sont produits jusqu'à l'instant t_i, et n_i le nombre d'individus ayant survécus jusqu'au temps t_i (y compris les individus censurés).

La censure intervient lorsque l'évènement n'a pas été observé durant le temps de l'étude. Cela peut signifier que l'évènement ne s'est pas produit, ou que l'individu est sorti de l'étude avant que celui-ci se produise. Nous faisons ici l'hypothèse que la censure est non-informative : le mécanisme de censure est indépendant de l'évènement observé. Pour prendre en compte la censure sur cet estimateur, lorsque l'on rencontre un individu censuré, l'effectif n_i au premier temps t_i après la date de censure diminue de 1.

La courbe de Kaplan-Meier a une forme caractéristique en escalier. Cette approche non-paramétrique ne tient cependant pas compte des attributs des différents observations, or ces dernières peuvent se révéler importantes dans l'estimation de la probabilité de survie d'un patient.

2.2 Modèle à risques proportionnels de Cox

Le modèle de Cox [1, 6] est un modèle semi-paramétrique servant à modéliser le phénomène d'apparition d'évènements (comme la mort d'un patient) au cours du temps, en fonction des covariables.

Ces phénomènes interviennent selon une fonction $\lambda(t, X)$, appélée *fonction de risque instantané* ou *taux de défaillance*, qui selon le modèle de Cox peut être modélisée comme suit :

$$\lambda(t, X) = \lambda_0(t) \exp(\beta^t X)$$

où λ_0 est appelé la fonction de risque de base (autrement dit le risque instantané de l'évènement lorsque toutes les covariables sont nulles), $\beta \in \mathbb{R}^p$ est un vecteur de paramètres et $X \in \mathbb{R}^p$ un vecteur de covariables.

Dans cette expression le premier terme $\lambda_0(t)$ dépend uniquement du temps tandis que le second terme $\exp(\beta^T X)$ ne dépend que des covariables.

Une hypothèse majeure qui en découle pour pouvoir utiliser le modèle de Cox est l'hypothèse des risques proportionnels : le rapport de risque doit être constant en temps.

A partir de l'observation d'un échantillon $(X_1, \ldots, X_n, \delta_1, \ldots, \delta_n)$, où δ_i est l'indicatrice de censure de l'individu i ($\delta_i = 0$ si l'individu est censuré, 1 sinon), l'estimateur $\hat{\beta}$ est obtenu en maximisant la log-vraisemblance partielle [5] :

$$\log\left(L(\beta, X_1, \ldots, X_n, \delta_1, \ldots, \delta_n)\right) = \sum_{i=1}^{n} \delta_i \left[\beta^T X_i - \log\left(\sum_{j \in R(t_i)} \exp(\beta^T X_j) \right) \right]$$

où $R(t_i)$ est l'ensemble des individus susceptibles de subir une défaillance (ou censure) au temps t_i, i.e. l'ensemble des individus encore en vie au temps $t_i - \epsilon$ (ou non censurés).

Cette estimation nous permet alors de retrouver la fonction de survie associée. En effet, la fonction de survie est liée à la fonction de risque instantané par la relation suivante :

$$S(t \mid X) = \exp\left(- \int_0^t \lambda(u, X)\, du \right)$$

3 Expériences

Les données utilisées pour les expériences suivantes sont issues d'une base américaine qui retrace le suivi des patients post-opératoire (500 000 patients environ). Ces données

contiennent des informations relatives aux patients : comme la description des individus ou encore des indicateurs biologiques aux différents moments du suivi (on dénombre environ 1000 indicateurs) sur une échelle de temps pouvant varier. Les variables descriptives sélectionnées pour cette étude sont toutes de nature quantitatives.

3.1 Protocole expérimental

L'objectif de ce travail préliminaire est de démontrer la nécessité de construire des analyses de survies par sous-groupes homogènes de patients. Nous nous sommes intéressés à l'ensemble des patients sans restriction. Nous cherchons à estimer le temps de survie après opération (l'évènement observé étant donc la mort du patient). Dans notre cas, 70% de nos données étaient censurées, et les dates d'opérations étaient différentes pour chaque patient : nous avons donc calculé le temps de survie comme le temps écoulé entre la date d'opération et la date de dèces si l'évènement était observé, la date de dernier suivi du patient sinon (avec l'indicatrice de censure δ_i).

Dans un premier temps nous avons tracé les courbes de Kaplan-Meier [1]. Ensuite nous avons testé le modèle de Cox sur nos données.

Notre cohorte étant relativement grande et hétérogène, nous pouvons envisager que plusieurs populations se distinguent à l'intérieur de celle-ci. Pour confirmer cela, nous avons séparé nos observations en plusieurs groupes de population. Par exemple, nous savons que les critères d'opérations ne sont pas exactement les mêmes en fonction du genre du patient ou encore de son groupe d'âge (enfant/adulte/personne âgée).

Nous avons séparé notre population globale en plusieurs groupes successivement et de manière indépendante : séparation de la population en deux groupes selon le genre (Homme/Femme), selon la présence ou l'absence de diabète, selon l'ethnie (latine ou non), et selon le groupe d'âges (5 ans et moins, 6-17 ans, 18-49 ans, 50 ans et plus). Nous avons tracé, pour chaque groupe, les courbes de survie de Kaplan-Meier correspondantes dans le but de les comparer graphiquement.

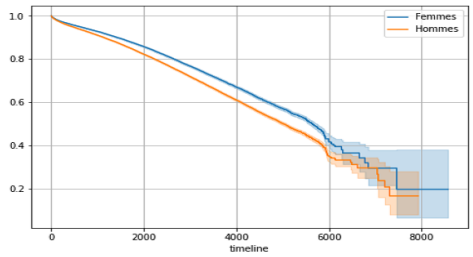
De plus, sur chacun de ces groupes nous avons également testé un modèle de Cox à 7 covariables (identiques pour chaque groupe) de nature quantitatives de manière à comparer la significativité des paramètres du modèle.

3.2 Résultats

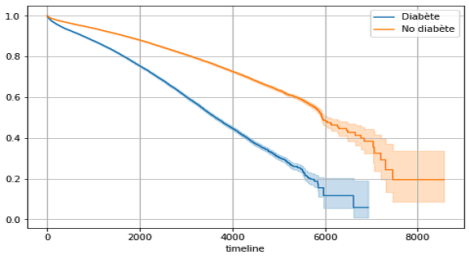
Les courbes de survie de Kaplan-Meier (probabilité de survie en fonction du temps en jours, estimée par l'approche de la section 2.1) obtenues sur nos données, et représentées en Figure 1, confirment l'existence de distinction entre les groupes étudiés en terme de probabilité de survie.

Si l'on s'intéresse en particulier à la Figure 1b on remarque que la fonction de survie décroit bien plus rapidement pour les personnes atteintes de diabète. Cette observation

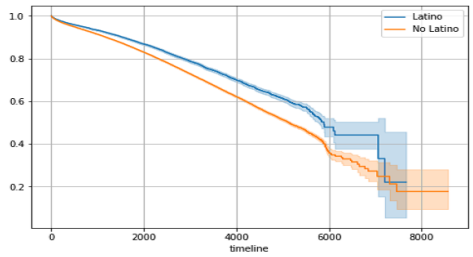
1. les courbes sont obtenues avec la librairie *lifelines* sous Python.



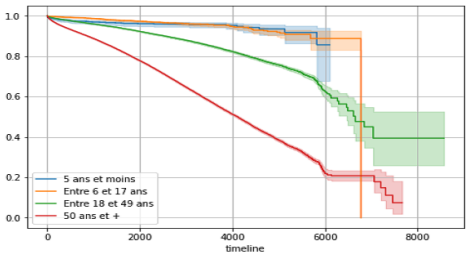
(a) Courbes de survie en fonction du genre (Homme/Femme)



(b) Courbes de survie en fonction du diabète (Présence/Absence)



(c) Courbes de survie en fonction de l'ethnie (Latine ou non)



(d) Courbes de survie en fonction de l'âge (5 ans et moins / 6-17 ans / 18-49 ans / 50 ans et plus

FIGURE 1 – Courbes de Kaplan-Meier avec intervalle de confiance à 95% sur différents sous-groupes de la population((a) Genre ; (b) Diabète ; (c) Ethnie ; (d) Âge).

rejoint les études menées en France mettant en lumière les difficultés d'obtenir les soins médicaux adéquats chez les personnes diabétiques [3]. De même, on constate également un écart important entre les courbes de survie selon les différents groupes d'âge. Ainsi, ces expériences préliminaires démontrent qu'il existe des sous-groupes de patients pour lesquels les courbes de survies sont très différentes.

	Modèle 5 ans et moins	Modèle 6-17 ans	Modèle 18-49 ans	Modèle 50 ans et plus
Âge (en mois)	0.04	0.07	< 0.005	<0.005
Creatinine donneur	0.52	0.97	0.17	0.08
Créatinine receveur	0.65	<0.005	<0.005	<0.005
Opérations antérieures	<0.005	<0.005	<0.005	<0.005
Creatinine élevée chez le donneur	0.59	0.95	0.69	0.22
Episodes de rejet aigu	0.44	0.27	<0.005	<0.005
Creatinine à la sortie de l'hôpital	<0.005	0.01	<0.005	<0.005

TABLE 1 – p-valeurs des covariables utilisées dans le modèle de Cox en fonction du groupe d'âges. *La creatinine est un déchet naturel de l'organisme. Lorsque la capacité de l'organisme à éliminer les déchets diminue, le taux de creatinine dans le sang augmente.*

En se concentrant sur la classe d'âge des patients, nous avons également réalisé quatre modèles de Cox distincts par groupes d'âges (en gardant les mêmes covariables). Ici encore les modèles pour chaque groupes se distinguent, en particulier par le nombre et le choix des variables significatives dans chaque modèle (Table 1). Par exemple, le taux de créatine du patient n'est une variable siginificative qu'à partir de l'âge 6 ans.

4 Conclusion et Perspectives

Les résultats précédents nous confirment qu'un modèle unique ne peut convenir à l'ensemble de notre jeu de données. Les modèles existants aujourd'hui, basés sur la régression de Cox classique, sont limités en ne représentant qu'une partie de la population (adultes, sans antécédants médicaux, etc). Ces critères restrictifs sont identifiés manuellement par les médecins [2].

Pour palier ce problème nous voulons détecter automatiquement des sous-groupes homogènes de la population d'un point de vue de la modélisation de la survie. Pour cela, la suite de ce travail consistera à considérer un mélange de modèles de Cox [4] :

$$\lambda(t, X) = \sum_{k=1}^{K} \pi_k \lambda_{0,k}(t) \exp(\beta_k^t X).$$

Références

[1] Breslow N. E. Analysis of survival data under the proportional hazards model. *International Statistical Review*, 43(1) :45–57, 1975.

[2] Foucher Y. et al. A clinical scoring system highly predictive of long-term kidney graft survival. *Kidney international*, 78(12) :1288–1294, 2010.

[3] Imhoff O. et al. «le receveur limite» : existe-t-il encore des freins à l'inscription des patients sur liste d'attente de transplantation rénale ? *Néphrologie & thérapeutique*, 3 :282–288, 2007.

[4] Rosen O. and Tanner M. Mixtures of proportional hazards regression models. *Statistics in Medicine*, 18(9) :1119–1131, 1999.

[5] Cox D. R. Regression models and life-tables. *Journal of the Royal Statistical Society : Series B (Methodological)*, 34(2) :187–202, 1972.

[6] Cox D. R. Partial likelihood. *Biometrika*, 62(2) :269–276, 1975.

Cox regression with linked data

Thanh Huan Vo [1], Guillaume Chauvet [2], André Happe [3], Emmanuel Oger [4], Stéphane Paquelet [5] & Valérie Garès [6]

[1] *INSA (IRMAR) and IRT b<>com, Rennes, France, E-mail: than-huan.vo@insa-rennes.fr*
[2] *ENSAI (IRMAR), Campus de Ker Lann, Bruz, France, E-mail: guillaume.chauvet@ensai.fr*
[3] *EA 7449 REPERES, France, E-mail: andre.happe@chu-brest.fr*
[4] *EA 7449 REPERES, France, E-mail: emmanuel.oger@univ-rennes1.fr*
[5] *IRT b<>com - Institute of Research and Technology b<>com, France, E-mail: stephane.paquelet@b-com.com*
[6] *Univ Rennes, INSA Rennes, CNRS, IRMAR - UMR 6625, F-35000 Rennes, France, E-mail: valerie.gares@insa-rennes.fr*

Résumé. Les méthodes d'appariement sont couramment utilisées pour combiner des variables provenant de plusieurs bases de données, mais associées à la même entité. Cependant, l'appariement est rarement parfait, et ne pas prendre en compte les erreurs afférentes peut conduire à des estimations biaisées. Plusieurs méthodes ont été proposées pour traiter ce problème dans le cas du modèle linéaire généralisé, mais à notre connaissance le modèle de Cox n'a pas été beaucoup étudié dans la littérature alors que c'est le modèle statistique le plus cité. Dans cet article, nous proposons une équation estimante ajustée pour tenir compte des erreurs dues à l'appariement. Cette méthode est utilisable pour un utilisateur secondaire de données appariées, quand l'appariement a été réalisé par un opérateur tiers et quand la personne réalisant l'analyse n'a pas d'information sur les variables d'appariement. Les résultats de nos simulations montrent que la méthode proposée permet de réduire significativement le biais d'estimation pour les paramètres du modèle de Cox causé par les erreurs d'appariement.

Mots-clés. Appariement, données de survie, erreur d'appariement, régression de Cox.

Abstract. Record linkage is increasingly used to combine data from different databases which refer to the same entities. This process can bring analysts valuable knowledge. However, linkage errors are usually unavoidable regardless of record linkage methods and ignoring these errors may lead to bias estimates. While different methods have been developed to deal with the linkage errors in the generalized linear model, there is not much interest on Cox regression model though this is the most cited statistical model in literature. In this article, we propose an adjusted estimating equation for secondary Cox regression analysis, where linked data have been prepared by someone else and no information on matching variables are available to the analyst. Through a Monte Carlo simulation study, the proposed method has significantly improved the bias of the parameter estimates of the Cox model caused by false links.

Keywords. Cox regression, linkage error, record linkage, survival data

1 Introduction

Record linkage, also known as data matching, is a process of combining data from different sources that refer to the same entities. In this era, data are collected everywhere by different sectors, and the ability of combining information on individuals from different databases can lead to novel knowledge for analysts. For example, record linkage is widely used in epidemiology and medical studies to enrich cohort data (e.g. Harron et al., 2016). The record linkage process is straightforward if unique identifiers (e.g. Social Security Number) are available in both databases. However, this information is often not available or can sometimes not be used due to ethical reasons. In such cases, record linkage methods can only use partial identifying information called matching variables such as name, address and gender. Over the few decades, different methods have been developed to link data efficiently (Herzog et al., 2007; Christen, 2012).

Because the matching variables are not unique identifiers and are likely to contain errors, linkage errors are unavoidable regardless of record linkage methods. The two kinds of record linkage errors are false links (false positive), i.e. a non-matched pair predicted as a link, and missed link (false negative), i.e. a matched pair failed to be predicted as a link. Ignoring these errors may cause substantial bias in the analysis model.

In the literature, different approaches have been considered to account for record linkage errors in generalized linear model (e.g. Lahiri and Larsen, 2005; Chambers, 2009; Zhang and Tuoto, 2020). Although the Cox proportional hazard model (Cox, 1972) is of extended use for survival analysis, comparatively very few papers have focused on accounting for record linkage errors in this context. Baldi et al. (2010) performed a simulation study emphasizing the impact of incomplete record linkage errors on the estimation of the parameters of the Cox model, but did not propose any solution. Hof et al. (2017) proposed a joint modeling for survival analysis and probabilistic record linkage. However, this analysis model is developed under a primary analysis viewpoint, and requires knowledge on the record linkage process. In many applications however, the record linkage process and the analysis of linked data are performed independently. In the present article, we reason from the secondary analysis viewpoint. We propose a model to account for the record linkage errors, and an estimation method to correct for the bias caused by record linkage errors in the Cox regression model.

2 Cox regression model

The proportional hazard model (Cox, 1972) is the most popular method to assess the effect of covariates $\mathbf{X}$ on the survival time. Let $\tilde{T}$ be a non-negative random variable

which denotes the duration between a time origin and the time of occurrence of some event of interest. $\tilde{T}$ is assumed to be right censored: we observe the event only if it occurs before a certain time C. Suppose that we have a random sample of n observations. Let $\tilde{T}_i$ and C_i be the latent survival time and the censoring time respectively of the unit i. For units $i = 1, \ldots, n$, we observe $T_i = \min(\tilde{T}_i, C_i)$ and $\delta_i = \mathbb{1}_{\{\tilde{T}_i \leq C_i\}}$ where $\mathbb{1}$ is the indicator function. The vector of covariates is denoted as $\mathbf{X}_i = (X_i^1, \ldots, X_i^p)^T$.

According to the Cox model, the hazard function of an event at time t is given by

$$\lambda(t|\mathbf{X}_i) = \lambda_0(t) \exp\left(\mathbf{X}_i^T \beta\right), \tag{1}$$

where $\beta = (\beta_1, \ldots, \beta_p)^T$ is a vector of $p \times 1$ unknown parameters and $\lambda_0(t)$ is a common baseline hazard function. Assuming that C is independent of T conditionally on $\mathbf{X}$, and that the data are observed on a finite interval $[0, \tau]$, a consistent estimator of β may be obtained by solving the estimating equation:

$$\sum_{i=1}^{n} \delta_i \left\{ \mathbf{x}_i - \frac{\sum_{j=1}^{n} Y_j(T_i) \exp\left(\mathbf{x}_j^T \beta\right) \mathbf{x}_j}{\sum_{j=1}^{n} Y_j(T_i) \exp\left(\mathbf{x}_j^T \beta\right)} \right\} = 0, \tag{2}$$

where $Y_j(t) = \mathbb{1}_{(T_i \geq t)}$. We call (2) the theoretical estimating equation. This is also the partial likelihood score equation.

3 A secondary Cox regression analysis for linked data

3.1 Linkage error model

Suppose that we have a dataset A of n_A time-to-event data. If the covariates x_i were known for any unit $i \in A$, the parameter of the Cox model would be estimated by solving the theoretical estimating equation. Unfortunately, if the covariates are not known for the units in A, equation (2) may not be solved in practice.
In order to obtain the needed covariates, a linkage is performed with a dataset B of size $n_B \geq n_A$, which contains in particular the auxiliary variables x_i. For some unit i in A, we note z_i for the auxiliary values resulting from the linkage process. The notations are summarized in Table 1.

We assume that the linkage error is non-informative, i.e. may depends on the errors in the matching process, but not on the model covariates nor on the survival time. Adopting the modelling approach of Copas and Hilton (1990), we suppose that both databases are partitioned into Q blocks A_q and B_q, $q = 1, \ldots, Q$, where the record linkage is performed independently. Also, we suppose that for any entity $i \in A_q$, we have:

$$\mathbf{Z}_i = \begin{cases} \mathbf{X}_i & \text{with probability } \alpha_q, \\ \mathbf{X}_{(j)} & \text{with probability } 1 - \alpha_q, \end{cases} \tag{3}$$

(a) File of interest A

	T	δ	$\mathbf{X} \in R^p$
$i = 1$			$\mathbf{Z}_1$
$\vdots$			
$i = n_A$			$\mathbf{Z}_n$

(b) Linking data file B.

	$\mathbf{X} \in R^p$
$j = 1$	
$\vdots$	
$\vdots$	
$j = n_B$	

Table 1: Record linkage context.

In other words, it is supposed that for any $i \in A_q$, the correct entity is linked to i with probability α_q. Otherwise, we suppose that the unit j linked to i is randomly selected in B_q.

In practice, the probabilities α_q are unknown. They may be estimated from a validation sample, or may be (somewhat arbitrarily) specified by the record linkage practitioner.

3.2 Adjusted estimating equation

By treating the linked covariates $\mathbf{z}_i$ as if they were the true covariates $\mathbf{x}_i$, the estimator of β is obtained by solving the following equation:

$$\sum_{q=1}^{Q} \sum_{i \in A_q} \delta_i \left\{ \mathbf{z}_i - \frac{\sum_{q=1}^{Q} \sum_{j \in A_q} Y_j(T_i) \exp(\mathbf{z}_j^\top \beta) \mathbf{z}_j}{\sum_{q=1}^{Q} \sum_{j \in A_q} Y_j(T_i) \exp(\mathbf{z}_j^\top \beta)} \right\} = 0. \tag{4}$$

We call (4) the naive estimating equation. Since some units are incorrectly linked, it may lead to biased estimates, see the simulation results.

Therefore, we propose a bias-corrected estimating equation, accounting from the fact that from the *hit-miss* model (3), the covariates may be incorrectly linked. The proposed estimating equation is:

$$\sum_{q=1}^{Q} \sum_{i \in A_q} \delta_i \left(\frac{\mathbf{z}_i}{\alpha_q} - \left(\frac{1}{\alpha_q} - 1 \right) \bar{\mathbf{x}}_q - \frac{\sum_{q=1}^{Q} \alpha_q^{-1} \sum_{j \in A_q} Y_j(T_i) \psi_j^Z(\beta) - (\alpha_q^{-1} - 1) \sum_{j \in A_q} Y_j(T_i) \bar{\psi}_q^X(\beta)}{\sum_{q=1}^{Q} \alpha_q^{-1} \sum_{j \in A_q} Y_j(T_i) \mu_j^Z(\beta) - (\alpha_q^{-1} - 1) \sum_{j \in A_q} Y_j(T_i) \bar{\mu}_q^X(\beta)} \right) = 0, \tag{5}$$

with $\mu_i^Z(\beta) = \exp(\mathbf{z}_i^\top \beta)$, $\mu_i^X(\beta) = \exp(\mathbf{x}_i^\top \beta)$, $\psi_i^Z(\beta) = \exp(\mathbf{z}_i^\top \beta)\mathbf{z}_i$ and $\psi_i^X(\beta) = \exp(\mathbf{x}_i^\top \beta)\mathbf{x}_i$, and where $\bar{\mathbf{x}}_q$, $\bar{\mu}_q^X(\beta)$ and $\bar{\psi}_q^X(\beta)$ are the means of $\mathbf{x}_i$, $\mu_i^X(\beta)$ and $\psi_i^X(\beta)$ over B_q respectively. Note that the computation of the terms in (5) requires access to the $\mathbf{x}_j$'s for all the units $j \in B$.

4 Simulation study

We generate two databases A with $n_A = 1000$ individuals and B with $n_B = 2000$ individuals. We first generate the units in database B with two covariates $X^1 \sim \mathcal{N}(0, 1)$ and

$X^2 \sim$ Bernoulli(0.7). Given the vector of coefficients $\beta = (0.5, -0.5)$, the true survival time $\tilde{T}^B$ is generated as $\tilde{T}^B = -\frac{\log(U)}{\lambda \exp(\mathbf{x}^\top \beta)}$ where U is a variable following a uniform distribution on the interval from 0 to 1 (Bender et al., 2005) and λ is fixed equal to 1 for simplicity. We chose a fixed censoring C to have a proportion of non-censoring equal to 0.75.

Without loss of generality, we suppose that a pair of individuals (a_i, b_j) is a match if $i = j = 1, \ldots, n_A$. Therefore, we obtain the survival times T^A for the individual in database A by taking the n_A first survival times in B. We use $Q = 3$ blocks with $(n_{A_1}, n_{A_2}, n_{A_3}) = (200, 300, 500)$, $(n_{B_1}, n_{B_2}, n_{B_3}) = (400, 600, 1000)$ and $(\alpha_1, \alpha_2, \alpha_3) = (0.75, 0.85, 0.95)$. The linked covariates Z are obtained from a record linkage process between A and B, as described in model (3).

We repeat the generating process 1,000 times. For each of the datasets, we obtain an estimator of the β parameter by solving the `Theoretical` estimating equation (2), the `Naive` estimating equation (4) or the `Proposed` estimating equation (5). For this last method, the coefficients α_q are estimated from a 10% random validation set.

The simulation results are presented in Table 2, in terms of Monte Carlo bias ($\mathrm{B}_{\mathrm{MC}}(\cdot)$), and Monte Carlo standard deviation ($\mathrm{Sd}_{\mathrm{MC}}(\cdot)$). As expected, the `Theoretical` estimating equation leads to almost unbiased estimators. The `Naive` estimating equation leads to biased estimators, but the `Proposed` method enables to significantly reduce this bias, while only slightly increasing the variability.

Methods	$\mathrm{B}_{\mathrm{MC}}(\beta_1)$	$\mathrm{Sd}_{\mathrm{MC}}(\beta_1)$	$\mathrm{B}_{\mathrm{MC}}(\beta_2)$	$\mathrm{Sd}_{\mathrm{MC}}(\beta_2)$
`Theoretical`	0.003	0.038	0.001	0.083
`Naive`	0.072	0.040	0.075	0.083
`Proposed`	0.015	0.073	0.012	0.116

Table 2: Summary of 1000 Monte-Carlo simulations when there are three blocks with three levels of linkage quality

References

Ileana Baldi, Antonio Ponti, Roberto Zanetti, Giovannino Ciccone, Franco Merletti, and Dario Gregori. The impact of record inkage bias in the cox model. *Journal of evaluation in clinical practice*, 16:92–6, 02 2010.

Ralf Bender, Thomas Augustin, and Maria Blettner. Generating survival times to simulate cox proportional hazards models. *Statistics in Medicine*, 24(11):1713–1723, 2005.

Ray Chambers. Regression analysis of probability-linked data. *Statistics New Zealand*, 2009.

Peter Christen. *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Springer Publishing Company, Incorporated, 2012.

J. B. Copas and F. J. Hilton. Record linkage: Statistical models for matching computer records. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 153 (3):287–320, 1990. ISSN 09641998, 1467985X.

D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972. ISSN 00359246.

Katie Harron, Ruth Gilbert, David Cromwell, and Jan van der Meulen. Linking data for mothers and babies in de-identified electronic health data. *PLOS ONE*, 11(10):1–18, 10 2016. doi: 10.1371/journal.pone.0164667.

Thomas Herzog, Fritz Scheuren, and William Winkler. *Data Quality and Record Linkage Techniques*. Springer-Verlag New York, 2007.

Michel H. Hof, Anita C. Ravelli, and Aeilko H. Zwinderman. A probabilistic record linkage model for survival data. *Journal of the American Statistical Association*, 112 (520):1504–1515, 2017.

P. Lahiri and Michael D. Larsen. Regression analysis with linked data. *Journal of the American Statistical Association*, 100(469):222–230, 2005.

Li-Chun Zhang and Tiziana Tuoto. Linkage-data linear regression. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 2020.

SFdS - AMIES

Lights: a generalized joint model for high-dimensional multivariate longitudinal data and censored durations

Van Tuan Nguyen [1]

[1]University de Paris – LPSM – France

Résumé

This paper introduces a prognostic method calledlightsto deal with the problemof joint modeling of longitudinal data and censored durations, where a large numberof both longitudinal and time-independent features are available. In the literature,standard joint models are either of type shared random-effect or joint latent classones ; where the association structure between the longitudinal and the time-to-eventsubmodels takes respectively the form of either shared association features learnedfrom the longitudinal processes and included as potential risk factor in the survivalmodel, or latent classes modeling population heterogeneity. We pick modeling ideasfrom both worlds and use appropriate penalties during inference for being able tolearn from a high-dimensional context. The statistical performance of the method isexamined on an extensive Monte Carlo simulation study, and finally illustrated on apublicly available dataset. Our proposed method significantly outperforms the state-of-the-art joint models regarding risk prediction in terms of C-index in a so-calledreal-time prediction paradigm, with a computing time orders of magnitude faster. Inaddition, it provides powerful interpretability by automatically pinpointing significantfeatures being relevant from a practical perspective. Thus, we propose a powerfulltool with the ability of automatically determining significant prognostic longitudinalfeatures, which is of increasing importance in many areas: for instance personalizedmedicine, or churn prediction in a customer profile and activity monitoring setting, toname but a few.

L'agence pour les mathématiques en interaction avec l'entreprise et la société (AMIES)

Véronique Maume-Deschamps [1,2]

[1]Institut Camille Jordan [Villeurbanne] – Université Claude Bernard Lyon 1 – France
[2]AMIES – INSMI (CNRS), , Universitey of Grenoble Alpes (UGA) – France

Abstract

Comment AMIES agit au niveau national pour développer les relations avec les entreprises ?

Prédiction des niveaux de risque pollinique à partir de données historiques multi-sources

Esso-Ridah Bleza[1], **Valérie Monbet**[2] **et Pierre-François Marteau**[1]

[1]*IRISA, Université Bretagne Sud.*
e-mail : esso-ridah.bleza@univ-ubs.fr *and e-mail :* pierre-francois.marteau@univ-ubs.fr
[2]*IRMAR, Université de Rennes*
e-mail : valerie.monbet@univ-rennes1.fr

Résumé: Dans la littérature scientifique, de nombreuses études montrent que les conditions météorologiques ont un impact sur l'émission, la dispersion et la suspension des pollens dans l'air menaçant la santé des millions de personnes en France. L'objectif est d'étudier la capacité à prédire à 3 jours (J+3) les niveaux de risques de présence de pollens dans l'air sur un territoire donné (en France Métropolitaine) avec des techniques d'apprentissage statistique exploitant des données historiques, et les paramètres météorologiques jusqu'au jour (J). Nous nous sommes intéressés à la prévision de risque pour 3 familles de pollens qui font partie des espèces les plus allergisantes (ambroisie, cupressacées et graminées). L'agrégation de modèles de régression logistiques par un classifieur de type Forêt aléatoire a permis de prédire les niveaux du risque pollinique ('nul', 'faible', 'moyen' et 'fort') avec des performances de l'ordre de 75% à 90% d'AUC et 70% de précision et de rappel, avec quelques confusions pour les niveaux faible et moyen.

Abstract: In the scientific literature, numerous studies show that meteorological conditions have an impact on the emission, dispersion and suspension of pollens in the air threatening the health of millions of people in France. The objective is to study the ability to predict 3 days (D+3) in advance the risk levels of pollen presence in the air on a given territory (in Metropolitan France) with statistical learning techniques exploiting historical data, and the meteorological parameters of the day (D). We were interested in risk prediction for 3 families of pollens that are among the most allergenic species (ragweed, cupressaceae and grasses). The aggregation of binary logistic regression models by a Random Forest classifier allowed us to predict the levels of pollen risk ('null', 'low', 'medium' and 'high') with performances in the range of 75% to 90% AUC and 70% precision and recall, with some confusions for the low and medium levels.

Mots-clés : pollen, aérobiologie, apprentissage automatique, classification

1. Introduction

Plusieurs études récentes montrent que les populations, tout particulièrement en France, souffrent de plus en plus d'allergies à un ou plusieurs pollens. On estime à ce jour que 25 % des français sont concernés et que la population allergique passera la barre des 50 % en 2050 (https://lejournal.cnrs.fr). À ce jour, les mesures de concentration des pollens dans l'air sont délivrées au public *a posteriori* avec un décalage temporel important (Cassagne, 2009). En effet, le réseau national de surveillance aérobiologique (RNSA), organisme de référence, diffuse les informations issues des capteurs HIRST dont la technique est basée sur l'analyse et l'identification au microscope des pollens induisant un délai dans la diffusion d'information. Ces constats incitent à développer des modèles prédictifs du risque pollinique à courte échéance pour compléter l'information diffusée par le RNSA et améliorer la prévention des allergies.

Différentes études (Cassagne, 2009; Spieksma et al., 1995; Castellano-Méndez et al., 2005; Muzalyova et al., 2021) se sont intéressées au développement de modèles prédictifs (présence d'une espèce de pollen, saison pollinique, variation inter-annuelle, niveau du risque ou de concentration pollinique) à partir des données météorologiques. Dans (Castellano-Méndez et al., 2005) les auteurs ont ramené le problème à une classification binaire pour chaque niveau de risque. Des modèles (régression ARIMA et dynamiques, réseaux neuronaux artificiels) (Iglesias-Otero et al., 2015; Cordero et al., 2021; Muzalyova et al., 2021), la méthode de Gradient Boosting (Cordero et al., 2021), les machines à vecteurs de support (Zewdie et al., 2019), les forêts aléatoires (Zewdie et al., 2019; Nowosad et al., 2018) sont utilisés dans d'autres études aérobiologiques, pour faire des prévisions de concentration quotidienne de pollen en ajoutant aux variables météorologiques, des paramètres phénologiques, des données caractéristiques du site et l'historique des concentrations de pollen.

Notre étude se distingue principalement sur deux aspects : premièrement nous avons choisi de prédire le risque pollinique assimilé à une variable discrète qui correspond à des niveaux d'émission (Thibaudon, 2003) seulement

Soumis au
File: output.tex, compiled with jsfds, version : 2009/12/09
date: 15 avril 2022

à partir de trois paramètres météorologiques (température, précipitation, humudité). Deuxièmement, nous nous intéressons à l'ensemble du territoire français en considérant au total 68 sites (sur 74 disponibles). Finalement, nous étudions avec une même méthodologie plusieurs pollens parmi les plus allergisants (21 espèces étudiées au total dont trois sont présentées dans cet article) répertoriés en France métropolitaine.

Les données sont introduites dans la section 2. Dans la section 3, les algorithmes d'apprentissage sont décrits. La section 4 regroupe les expérimentations : la capacité de généralisation des algorithmes de prédiction proposés est évaluée en temps et en espace. Enfin la section 5 présente quelques conclusions.

2. Données

Les données d'émission de pollen sont produites par le RNSA [1]. En pratique, les grains de pollen se collent sur une lame adhésive qui est relevée puis analysée. On dispose ainsi de données journalières de concentration de pollen pour au moins une vingtaine d'espèces allergisantes et 74 sites pour une période couvrant les années 2000 à 2017.

Comme le montre la figure 1, les concentrations de pollens présentent une forte saisonnalité annuelle avec une variation inter-annuelle qui peut être assez marquée selon les espèces. De plus, la distribution des pollens est dissymétrique. Elle est caractérisée par un nombre important de zéros et une queue lourde.

Trois espèces choisies pour leurs niveaux d'émissions, leurs caractères allergisants et leurs saisonnalités polliniques distincts sont présentées : l'ambroisie, les poacées (graminées) et les cupréssacées avec les seuils de niveaux de risques détaillés dans le tableau 1.

pollen	seuil faible	seuil moyen	seuil fort
ambroisie	2	3	11
cupressacées	70	142	284
poacées	2	5	36

TABLE 1. *Seuils de risque d'exposition allergique pour l'ambroisie, les cupréssacées et les poacées, en nombre de grains par m^3/jour. Valeurs issues de (Thibaudon, 2003).*

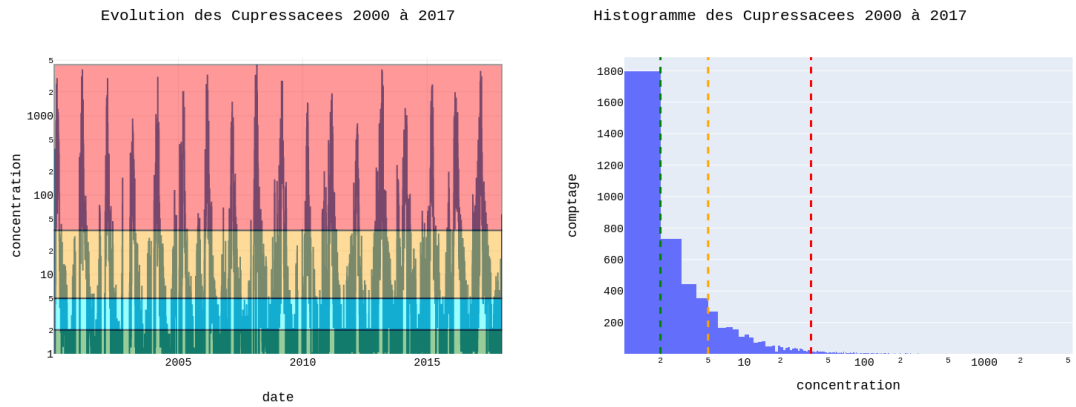


FIGURE 1. *exemple d'un historique de données de concentration des cupréssacées de 2000 à 2017 (à gauche), avec les niveaux de risque "nul", "faible", "moyen", "fort" matérialisés par les bandes verte, bleue, orange et rouge . Histogramme du même historique de données (à droite) ; Les barres verticales matérialisent les seuils "faible", "moyen" et "fort" de niveaux de risque (de la gauche vers la droite)*

[1] https://www.pollens.fr/reports/database

Soumis au
File: output.tex, compiled with jsfds, version : 2009/12/09
date: 15 avril 2022

Les données météorologiques sont issues de la base du European Climate Assessment & Dataset project (ECAD) [2]. Il s'agit de données enregistrées par des capteurs au sol situés à proximité des aéroports (68 sites). Les stations météorologiques ne se trouvent pas à proximité immédiate des capteurs HIRST souvent situés dans les centres-villes (voir figure 2). Nous avons néanmoins choisi de travailler avec ces données d'observation parce qu'elles sont facilement accessibles en temps réel. En outre, à l'échelle journalière, les champs de variables météorologiques sont assez lisses et la variabilité sur quelques dizaines de kilomètres est faible. Dans cette étude nous considérons des données journalières de la température moyenne à 2 mètres, de cumul de précipitations et d'humidité. Les variables introduites dans le modèle de prédiction sont :
- la température du jour t et ses écarts entre deux jours successifs pour les jours t et $t-1$;
- les écarts entre les températures maximum et minimum aux jours t, $t-1$ et $t-2$;
- les degrés jour $D_t = \sum_{j=t-30}^{t} \frac{Tmax_j - Tmin_j}{2} \mathbf{1}_{Tmin_j > 10}$ où $\mathbf{1}$ est la fonction indicatrice et T_{min} et T_{max} les min et max journaliers de la température ;
- l'humidité du jour t et ses écarts entre deux jours successifs pour les jours t et $t-1$;
- le cumul de précipitations ;
- la saisonnalité annuelle avec le numéro de la semaine et une variable de jour $\cos(\frac{2\pi t}{365})$;
- la latitude et la longitude des capteurs HIRST et une variable qualitative de région qui prend les valeurs Nord-Est, Nord-Ouest, Sud-Est et Sud-Ouest.

3. Modèles

Ici, nous proposons une approche dans laquelle la concentration en pollen est discrétisée. Ainsi, la variable cible $Y_t^{(p)}$ représente le risque allergique lié au pollen. Elle est construite de la façon suivante.

$$\begin{aligned} \text{si } C_t^{(p)} < s_{\text{faible}} \rightarrow Y_t^{(p)} &= \text{"null"} \\ \text{si } s_{\text{faible}} \leq C_t^{(p)} < s_{\text{moyen}} \rightarrow Y_t^{(p)} &= \text{"low"} \\ \text{si } s_{\text{moyen}} \leq C_t^{(p)} < s_{\text{fort}} \rightarrow Y_t^{(p)} &= \text{"medium"} \\ \text{si } C_t^{(p)} \geq s_{\text{fort}} \rightarrow Y_t^{(p)} &= \text{"high"} \end{aligned}$$

où $C_t^{(p)}$ est la concentration du pollen p au temps t. Les seuils s_{faible}, s_{moyen}, s_{fort} sont définis a priori et dépendent de l'espèce de pollen (voir table 1). Ainsi, $Y_t^{(p)}$ est une variable ordinale et on peut envisager différent modèles de prédiction.

Nous proposons ici de calibrer des modèles de régression multinomiale. Les paramètres sont estimés en maximisant la vraisemblance. Plus précisément, nous proposons un modèle d'agrégation qui consiste à combiner les prédictions de 3 régressions logistiques pour les seuils "faible", "moyen" et "fort" du niveau de risque. En pratique, les classes sont déséquilibrées et les prédictions sont réalisées en comparant la probabilité prédite à un seuil. Pour chaque modèle le seuil est déterminé de façon à maximiser le F1-score sur l'ensemble d'apprentissage.

Dans une seconde version du modèle d'agrégation, les probabilités $\hat{\pi}_t^{R_{(r)}}$ des modèles binomiaux associées au risque ${R_t}^r$, $r \in \{faible, moyen, fort\}$ sont utilisées en entrée d'un algorithme de forêt aléatoire pour obtenir un classement final à quatre modalités de risque : "nul", "faible", "moyen" ou "fort".

4. Résultats

Deux types de modèles sont entraînés : modèles dits "locaux" lorsque les données d'apprentissage utilisées sont restreintes à un site donné et "spatiaux" lorsqu'elles intègrent les données de l'ensemble des sites. Les modèles sont entraînés sur 80 % de données et évalués sur les 20 % restant. Six sites ont été retirés de la base d'apprentissage

[2] https://www.ecad.eu/

Soumis au
File: output.tex, compiled with jsfds, version : 2009/12/09
date: 15 avril 2022

Pollen	Risk	AUC		Precision		Recall	
		model	AR model	model	AR model	model	AR model
ragweed	**low**	89.2	90.06	85.6	97.45	88.5	94.52
	medium	89.4	88.78	86	97.95	89	96.33
	high	89.7	84.64	88.7	99.05	91.9	98.7
cupressaceae	**low**	79.1	85.75	70.4	89.29	70	80.65
	medium	78.7	81.64	72.6	90.59	71.7	88.57
	high	83.5	85.54	80.8	97.74	83.9	94.94
poaceae	**low**	84.7	85.88	77.6	87.14	76.6	82.31
	medium	82.7	85.65	75.3	89.47	74.2	89.25
	high	79	89.49	72.2	94.74	74.5	91.87

TABLE 2. *AUC, precision and recall obtained on the test data set by the aggregation of binary logistic regression models as well as the auto-regressive (AR) versions (% values)*

.

pour évaluer la capacité de généralisation en espace. Les différents modèles entraînés sont comparés sur la base de critères (l'aire sous la courbe ROC (AUC), la précision et le rappel et F1-scores) calculés sur les échantillons tests et reportés dans le tableau 4.4).

4.1. Performances globales des modèles binaires

Les performances globales des modèles binaires agrégés sont assez bonnes, comme le montre le tableau 2. Les résultats du tableau 2 concernant les cupressacées et les poacées montrent que les modèles autorégressifs (AR) incorporant les risques prédits à $t+1$ et $t+2$ sont significativement meilleurs que les modèles de base simples selon les mesures AUC, précision et rappel. Pour l'ambroisie, l'AUC des modèles AR est moins bonne que celle des modèles simples pour les seuils moyens et élevés. L'ambroisie est l'une des espèces qui n'émet de très faibles quantités de pollen. Elle atteint donc difficilement les risques moyens et élevés.

4.2. Analyse par site géographique

Une cartographie des AUC du modèle des poacées (figure 2) montre que les sites du Nord-Est et du Centre sont légèrement mieux prédits que les sites de l'Ouest qui peut être dû la variabilité de quantités polliniques. On remarque aussi que les sites voisins ont des AUC similaires. Par ailleurs, on observe en général que les prédictions sont meilleures quand les stations météorologiques sont proches des sites des capteurs HIRST (voir Rennes et Dinan par exemple).

4.3. Comparaison des modèles spatiaux et locaux

Une étude comparative entre modèle spatial et local est proposée sur les sites de Rennes et d'Ajaccio. Par comparaison des AUC obtenus pour chaque type de modèle, un classement est présenté dans le tableau 3. Dans 87.5 % des cas le modèle local est meilleur en AUC que le modèle spatial à Ajaccio contrairement au site de Rennes pour lequel 85.71 % des modèles spatiaux sont meilleurs que les modèles locaux. Ceci confirme les premiers résultats montrés sur la figure 2, le score du modèle spatial à Ajaccio correspond à l'un des plus mauvais scores.

Pour le cas des sites de Rennes et d'Ajaccio nous avons comparé la classification finale des poacées, en effectuant une combinaison par forêts aléatoires des modèles locaux, d'une part et des modèles spatiaux d'autre part (voir tableau 4) . La classification par combinaison des modèles spatiaux permet d'obtenir de meilleurs scores avec environ 2 % d'écart.

La comparaison des modèles spatiaux et locaux a montré de façon générale que les écarts entre les AUC des deux types de modèles sont inférieurs à 10 %. Dans environ 70 % des cas, les modèles spatiaux conduisent à

Soumis au
File: output.tex, compiled with jsfds, version : 2009/12/09
date: 15 avril 2022

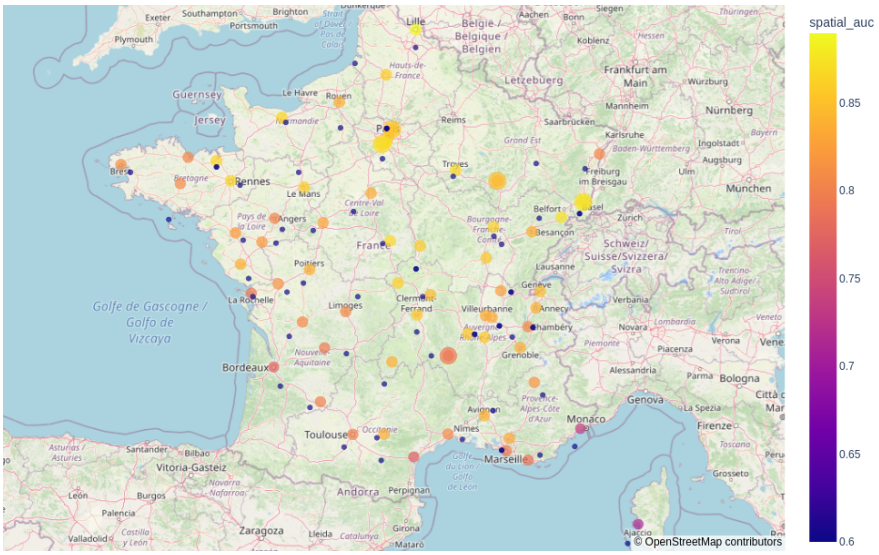


FIGURE 2. *Carte des AUC des prédictions du risque moyen des Poacees des données tests (points colorés suivant l'échelle spaltial$_a$uc et indiquant la position des capteurs HIRST). Les points bleus matérialisent la position des capteurs météos utilisés*

sites	modèle	nb	taux
Ajaccio	Local	7	87.5 %
	Spatial	1	12.5 %
Rennes	Local	1	12.29 %
	Spatial	6	85.71 %

TABLE 3. *Comparaison des modèles spatiaux et locaux sur Rennes et d'Ajaccio. nb est le nombre de fois que le type de modèle est meilleur. taux est le pourcentage de meuilleurs modèles*

sites	modèle	précision	rappel	F1-score
Ajaccio	spatial	0.79	0.68	0.72
	local	0.77	0.65	0.69
Rennes	spatial	0.74	0.68	0.70
	local	0.72	0.64	0.67

TABLE 4. *Comparaison des classifications finales par combinaison des modèles (de régression logistique binaire) spatiaux et locaux*

de meilleures performances. Ces derniers présentent l'avantage de réduire le nombre de modèles à entraîner d'un facteur 62 et de permettre l'extrapolation spatiale.

La généralisation en espace des modèles en faiant les prédictions sur les données des six sites retirés montre que les performances sont dans les mêmes ordres de grandeurs pour l'AUC et la précision que pour les autres sites. Par exemple, pn observe respectivement pour l'ambroisie, les cupréssacées et les poacées des minmums des AUC de 81,84%, 69,3% et 79,23% et de précision de 76,54%, 70,1% et 73,65%. Les résultats d'AUC des modèles de poacées sont confirmés sur la carte de la figure 2 où les points les plus gros représentent ces sites retirés. Les performances moins bonnes pour Le-Puy-en-Velay (comparativement à Saint-Etienne ou Lyon) s'expliquent par des conditions météorologiques très locales.

4.4. comparaison de l'agrégation des modèles binomiaux et d'autres approches de classification

Nous comparons plusieurs autres modèles avec notre approche d'agrégation de modèles de classification binaire dans le tableau 5 : Agrégation de modèles binaires (MetaModel), Régression logistique multinominale (MNL), Régression logistique ordinale (OL), Forêt aléatoire (RF) et Arbre de décision (DT).

Le MetaModel présente un léger avantage par rapport à tous les autres modèles en termes de performance en AUC sur les données de test.

Soumis au
File: output.tex, compiled with jsfds, version : 2009/12/09
date: 15 avril 2022

pollen	DT	RF	MNL	OL	MetaModel
ragweed	89.52	90.89	86.00	90.73	91.81
cupressaceae	67.55	76.27	72.51	73.89	77.50
poaceae	77.66	80.57	78.32	81.33	80.91

TABLE 5. *omparaison de l'AUC en% obtenue par les différents modèles spatiaux pour les trois espèces. DT, RF, MNL et OL sont entraînés à classer les niveaux de risque "nul", "faible", "moyen" et "élevé" avec les mêmes données d'entrée que celles utilisées pour les modèles de régression logistique binaire. MetaModel a comme entrées les 3 scores de probabilité des modèles binaires plus les données temporelles (numéro de jour) et les données géographiques (coordonnées GPS et région).*

5. Conclusion

Nous avons présenté une étude de faisabilité sur la prédiction à trois jours des niveaux de risque d'émission de pollen à partir de données historiques aérobiologiques, météorologiques et de géolocalisation. L'originalité de l'approche réside dans la prise en compte conjointe des informations spatiales et temporelles. Notre étude montre qu'il est généralement possible de prédire les épisodes polliniques de manière acceptable en prenant en compte des niveaux de risque discrets. Les perspectives à court terme de cette étude sont : i) le développement de modèles hybrides pour agréger les modèles locaux et globaux, ii) l'enrichissement des données d'apprentissage en intégrant des informations telles que le rayonnement ou le taux d'ensoleillement, les données de couverture végétale ou les données satellitaires. (ceci permettra d'expérimenter des modèles basés sur des approches neuronales (CNN spatio-temporel, etc.), iii) le couplage des données historiques HIRST et météorologiques avec de nouveaux types de données en temps réel provenant de nouveaux capteurs optiques développés par la société Lify Air et qui sont dédiés à la détection des émissions de pollen.

Références

Cassagne, E. (2009). Revue bibliographique des principaux seuils de détermination et méthodes de prévision de la date de début de pollinisation (ddp). *Revue Francaise d'Allergologie*, 49(8) :571–576.

Castellano-Méndez, M., Aira, M., Iglesias, I., Jato, V., and González-Manteiga, W. (2005). Artificial neural networks as a useful tool to predict the risk level of betula pollen in the air. *International Journal of Biometeorology*, 49(5) :310–316.

Cordero, J. M., Rojo, J., Gutiérrez-Bustillo, A. M., Narros, A., and Borge, R. (2021). Predicting the olea pollen concentration with a machine learning algorithm ensemble. *International Journal of Biometeorology*, 65(4) :541–554.

Iglesias-Otero, M., Fernández-González, M., Rodríguez-Caride, D., Astray, G., Mejuto, J., and Rodríguez-Rajo, F. (2015). A model to forecast the risk periods of plantago pollen allergy by using the ann methodology. *Aerobiologia*, 31(2) :201–211.

Muzalyova, A., Brunner, J. O., Traidl-Hoffmann, C., and Damialis, A. (2021). Forecasting betula and poaceae airborne pollen concentrations on a 3-hourly resolution in augsburg, germany : toward automatically generated, real-time predictions. *Aerobiologia*, pages 1–22.

Nowosad, J., Stach, A., Kasprzyk, I., Chłopek, K., Dąbrowska-Zapart, K., Grewling, Ł., Latałowa, M., Pędziszewska, A., Majkowska-Wojciechowska, B., Myszkowska, D., et al. (2018). Statistical techniques for modeling of corylus, alnus, and betula pollen concentration in the air. *Aerobiologia*, 34(3) :301–313.

Spieksma, F. T. M., Emberlin, J., Hjelmroos, M., Jäger, S., and Leuschner, R. (1995). Atmospheric birch (betula) pollen in europe : Trends and fluctuations in annual quantities and the starting dates of the seasons. *Grana*, 34(1) :51–57.

Thibaudon, M. (2003). The pollen-associated allergic risk in france. *European annals of allergy and clinical immunology*, 35(5) :170–172.

Zewdie, G. K., Liu, X., Wu, D., Lary, D. J., and Levetin, E. (2019). Applying machine learning to forecast daily ambrosia pollen using environmental and nexrad parameters. *Environmental monitoring and assessment*, 191(2) :1–11.

Soumis au
File: output.tex, compiled with jsfds, version : 2009/12/09
date: 15 avril 2022

Deep Calibration of Interest Rates Models

Djibril SARR [1], Mohamed BEN ALAYA [2] & Ahmed KEBAIER[3]

[1] *LAGA, Univ. Sorbonne Paris Nord, 99 Av. Jean Baptiste Clément, 93430 Villetaneuse*
sarr@math.univ-paris13.fr
[2] *LMRS, Univ. de Rouen Normandie, 1 Rue Thomas Becket, 76130 Mont-Saint-Aignan*
mohamed.ben-alaya@univ-rouen.fr
[3] *LAMME, Univ. d'Evry, 23 Bd. de France, 91037 Évry Cedex*
ahmed.kebaier@univ-evry.fr

Abstract. For any financial institution it is a necessity to be able to apprehend the behaviour of interest rates. Despite the use of Deep Learning that is growing very fastly, for many reasons (expertise, ease of use, ...) classic rates models such as CIR, or the Gaussian family are still being used widely. We propose to calibrate the five parameters of the G2++ model using Neural Networks. To achieve that, we construct synthetic data sets of parameters drawn uniformly from a reference set of parameters calibrated from the market. From those parameters, we compute Zero-Coupon and Forward rates and their covariances and correlations. Our first model is a Fully Connected Neural network and uses only covariances and correlations. We show that covariances are more suited to the problem than correlations. The second model is a Convolutional Neural Network using only Zero-Coupon rates with no transformation. The methods we propose have performed better than classic calibration methods on a comparison data set, with a global error, 2 times lower and a calibration in less than a second, compared to 164 seconds for the classic method.

Keywords. Quantitative Finance, Calibration, Neural Networks ...

1 Introduction and State Of the Art

Whether it is to manage risks or to optimize investments returns, it is necessary to be able to understand, forecast and stress drivers of fair values assets for example. Among those drivers are interest rates (IR). For this purpose, IR models are widely used, among the more popular are, the Vasicek [Vasicek, 1977], the Cox Ingersoll Ross Model (CIR) [Cox et al., 1985], and the 2 factors Gaussian model G2++ [Brigo and Mercurio, 2006].

This document focuses on the calibration of the G2++ using deep learning (DL) techniques. More precisely we will introduce 2 different approaches that are expected to be at least as accurate as the *classic* ones , to perform faster and above all, to be easier to use regarding the work necessary on the data. We recall the equations defining the G2++ model:

$$
\begin{aligned}
&dr(t) = \phi(t) + x(t) + y(t) \\
&dx(t) = -K_x x(t)dt + \sigma_x dW_t^x,\ x(0) = x_0 \\
&dy(t) = -K_y y(t)dt + \sigma_y dW_t^y,\ y(0) = y_0 \text{ with } k_x, k_y, \sigma_x, \sigma_y > 0 \text{ and } [dW_t^x, dW_t^y] = \rho dt
\end{aligned}
$$

For the one factor Hull-White model, [Gurrieri et al., 2009] define the calibration as these 3:

1. The choice of constant or time-dependent mean reversion and volatility.
2. The choice of products to calibrate to, and whether to calibrate locally or globally.
3. Whether to optimize on the mean reversion $a(t)$ and the volatility $\sigma(t)$ together or separately, and in the latter case, how to estimate one independently of the other.

[Hull and White, 2001] for example uses numerical optimization techniques to find the set of volatility parameters minimizing the error between the model's prices for cap & floor options and swaptions. The error used is the sum of the squares of the differences (SSD).

Regarding the suse of DL in finance, [Pironneau, 2019] for example calibrated the Heston model for European options using (Feed Forward Neural Network) FNN. [Hernandez, 2016] also uses FNNs to calibrate the one factor Hull-White model using swaptions prices.
The approach we present is different than the two aforementioned in both the input data, the architectures we use and above all, the models we introduce are applied to the simultaneous calibration of the five parameters of the G2++ model but can be applied to any model.

2 A systematic approach with Deep Calibration (DC)

In what follows, we will present two different approaches to calibrate IR models. Both the approaches actually are not model-dependent (thus systematic) and could be applied to any model (financial or not) as long as what we call the Observable Quantity Of Interest (OQOI) which has to be a relevant information bearer, can be both observed (in the market in our financial context) and obtained through an analytical expression from the model to calibrate.

We recall the expression of ZC $Z(t,T)$ and FWD $f(t,T)$ rates at time t and maturity T (see [Brigo and Mercurio, 2006] for the proof):

$$
\begin{aligned}
Z(t,T) &= -\frac{1}{T} ln \left\{ \frac{P^M(0,T)}{P^M(0,t)} \right\} - \frac{1}{T}\{\mathcal{A}(t,T)\} \\
f(t,T) &= -\frac{\sigma_x^2}{2K_x^2}\left[1 - 2e^{-K_x(T-t)} + e^{-2K_x(T-t)}\right] + \frac{\sigma_y^2}{2K_y^2}\left[1 - 2e^{-K_y(T-t)} + e^{-2K_y(T-t)}\right] \\
&\quad + \rho\frac{\sigma_x\sigma_y}{k_x k_y}\left[1 - e^{-K_x(T-t)} + e^{-K_y(T-t)} - e^{-(K_x+K_y)(T-t)}\right] + e^{-k_x(T-t)}x(t) + e^{-K_y(T-t)}y(t)
\end{aligned} \tag{1}
$$

With $P^M(0,T)$, the initial ZC market curve at maturity T and $V(.,.)$ and $\mathcal{A}(.,.)$ verifying:

$$
\begin{aligned}
V(t,T) =& \frac{\sigma_x^2}{K_x^2}\left[T - t + \frac{2}{K_x}e^{-K_x(T-t)} - \frac{1}{2K_x}e^{-2K_x(T-t)} - \frac{3}{2K_x}\right] + \frac{\sigma_y^2}{K_y^2}\left[T - t + \frac{2}{K_y}e^{-K_y(T-t)} - \frac{1}{2K_y}e^{-2K_y(T-t)} - \frac{3}{2K_y}\right] \\
&+ 2\rho\frac{\sigma_x\sigma_y}{K_x K_y}\left[T - t + \frac{e^{-K_x(T-t)} - 1}{K_x} + \frac{e^{-K_y(T-t)} - 1}{K_y} - \frac{e^{-(K_x+K_y)(T-t)} - 1}{K_x + K_y}\right] \\
\mathcal{A}(t,T) =& \frac{1}{2}[V(t,T) - V(0,T) + V(0,t)] - \frac{1 - e^{-K_x(T-t)}}{K_x}x(t) - \frac{1 - e^{-K_y(T-t)}}{K_y}y(t)
\end{aligned}
$$

2.1 Indirect Deep Calibration of the G2++ model

2.1.1 Covariances and Correlations as OQOI

We start from the observation that ZCs and FWDs rates indeed are easily observable in the market and an analytical expression can be drawn from most if not all IR models. However, if we want to keep a simple Fully Connected Network (FCN) architecture, directly using them as our OQOI might be time consuming and less efficient. We decide to use intermediary quantities (therefore the *Indirect Deep Calibration*) that will bear enough compressed information from the ZCs and FWDs curves to allow the DC. Good candidates for the OQOI could be correlation (COR) and covariance (COV) matrices of ZCs - and FWDs - variations.

We can introduce the discrete instantaneous covariance (and correlation) expressions as follows (see [Brigo and Mercurio, 2006]):

$$\mathrm{Cov}\left(\Delta G(t,T_i),\Delta G(t,T_j)\right) = \left[X(t,T_i)X(t,T_j) + Y(t,T_i)Y(t,T_j) + \rho\left(X(t,T_i)Y(t,T_j) + X(t,T_j)Y(t,T_i)\right)\right]\Delta t$$

$$\mathrm{Cor}\left(\Delta G(t,T_i),\Delta G(t,T_j)\right) = \frac{\mathrm{Cov}\left[\Delta G(t,T_i),\ \Delta G(t,T_j)\right]}{\sqrt{\left[X^2(t,T_i)+Y^2(t,T_i)+2\rho X(t,T_i)Y(t,T_i)\right].\left[X^2(t,T_j)+Y^2(t,T_j)+2\rho X(t,T_j)Y(t,T_j)\right]}} \tag{2}$$

$$X(t,T) = \begin{cases} \sigma_x e^{-K_x(T-t)} & \text{if } G(t,T) \ = \ f(.,T) \\ \sigma_x \frac{1-e^{-K_x(T-t)}}{K_x T} & \text{if } G(t,T) \ = \ Z(.,T) \end{cases} \tag{3}$$

Where $\Delta G(t,T) = G(t+\Delta t,T) - G(t,T),\ \Delta t > 0$.

Data set construction The data set construction is made in 5 steps:

1. Calibrating reference parameters for the G2++ model using a MSE minimization;
2. Extending the reference parameters to define a range of acceptance and drawing the parameters within the corresponding intervals;
3. Using equations (2) and (3), computing the set of COVs and CORs for ZCs and FWDs;
4. Choosing the scaling transformations needed;
5. And finally splitting the data set for training and validation.

The neural network architecture: As said previously we want here to keep a simple FCN architecture. Our NN has 5 linear layers : 1 input layer with nf - the number of features - neurons, 3 hideen layers, the first one with $h_1 = 1\,000$ neurons, the second with $h_2 = 1\,500$, the last hidden layer with $h_3 = 1\,000$ neurons and finally 1 output layer of np - the number of parameters to calibrate ($np = 5$) - neurons. The first 3 layers are activated using a ReLU function and the last one, as a prediction layer has no activation. To help prevent overfitting, we also include a dropout of probability 25% between the last hideen layer and the prediction layer.

2.2 Direct Deep Calibration of the G2++ model

2.2.1 ZCs as OQOI

We now want to consider calibration directly using ZCs curves which are direct market observations unlike the CORs and COVs.

The neural network architecture The architecture combining convolution and linear layers is still quite shallow. We start with a convolution layer of dimension (C=1, H=106, nf=28), with a filter of size (7 x 7). The stride of the convolution is 2 and we also allow padding. Then we add a pooling layer of stride 2 also. Finally follow 2 linear layers, the first one with 100 neurons and the second one (which is the prediction layers) with still $np = 5$ layers. Here also, we include in the second to last layer a dropout of probability 25%.

3 Results obtained

Regarding accuracy, both algorithms lead to pretty good results, whilst time efficiency is also remarkable as calibration on the test set (2 000 entries) is performed in the indirect DC model in less than 0.30 seconds and less than 0.10 seconds for direct DC. Table 1 below recapitulates the MSEs obtained on the validation by parameter.

Method	OQOI	K_x	K_y	σ_x	σ_y	ρ
Indirect DC	COV - ZC	3.5×10^{-4}	5.6×10^{-4}	7.7×10^{-4}	8.3×10^{-4}	8.2×10^{-2}
	COV - FWD	3.5×10^{-4}	5.5×10^{-4}	7.6×10^{-4}	8.3×10^{-4}	8.2×10^{-2}
	COR - ZC	7.5×10^{-4}	1.1×10^{-3}	1.3×10^{-3}	1.4×10^{-3}	2.0×10^{-1}
	COR - FWD	7.4×10^{-4}	1.1×10^{-3}	1.3×10^{-2}	1.4×10^{-2}	2.4×10^{-2}
Direct DC	ZCs	4.0×10^{-4}	6.3×10^{-4}	9.8×10^{-4}	1.0×10^{-4}	15.2×10^{-2}

Table 1: Calibration errors on the validation set

Indirect DC: As one can expect, knowing the similarities between the FWDs and ZCs curves, the result with the indirect DC are really close for FWDs and ZCs.

The really interesting result we get is the fact that the error made using COVs is about half the error with CORs. It appears that they bring significantly more information to calibrations. We can verify this assertion numerically by observing the derivative curves of covariances and correlations w.r.t. to the parameters, computed by finite differences.
We directly see that the correlations' derivatives quickly vanish for the three parameters. Meanwhile, for covariances, we only have the vanishing derivatives for K_x but it occurs less quicker. For σ_y and ρ, we don't even have that behaviour.

Direct DC: Compared, to the indirect DC with ZC COVs, the direct method leads to slightly less accurate calibrations. Even though the differences are not significant, it actually makes sense: the indirect DC benefits from information already extracted from the ZC curve.

While MSEs analysis allows to compare the different models, visualization allow to observe the fitting and assess whether the results seem good or not.

Figure 1: Indirect DC on ZCs COVs: Fitting curve on K_x, σ_x, and ρ (original and zoom)

Figure 2: Indirect DC on FWDs CORs: Fitting curve on K_y, σ_y, and ρ

4 Deep Calibration vs Classic Calibration

Comparison Methodology

1. First of all, we choose N_{sets} sets of parameters $\{k_{x_i}, k_{y_i}, \sigma_{x_i}, \sigma_{y_i}, \rho_i\}_{i\in[|1,N_{sets}|]}$.
2. We compute ZCs with equation (1) and we add white noise and jumps to replicate market data behaviour;
3. Then we compute (numerically) COVs and CORs from the ZCs computed;
4. We proceed to the direct DC and to the CC;

Computational performances: The time needed for the calibration is remarkably reduced by the DC. The $N_{sets} = 50$ CCs are made in 2 minutes and 44 seconds (164 seconds), meanwhile, the DC finished in less than a second, more precisely, in 0.063 seconds. This drastic reduction of the processing time easily justifies the time spent beforehand to train the neural network.

Calibration accuracy: Overall (i-e by taking the mean of all MSEs), the DC results in a global error of 0.10 which is 2.5 lower than the 0.25 scored by the CC.

5 Conclusion

Despite the growing use of DL and related AI techniques in finance, analytical models, such as IR ones like the G2++ remain commonly used due to the expertise professional and researchers have developed over the years regarding these techniques. However, it doesn't quite mean that DL and AI in general, shouldn't try to improve our use of those models.

The errors and computational performances obtained by our model allow us to conclude that using DL leads to very fast calibration and with low errors for the G2++ model - and subsequently for any other model with data fitting our requirements for the OQOI. Indeed, we have achieved a global error, 2 times lower than classical optimization method in less than a second compared to the 164 seconds of the classic method. Also, as we aimed, the models are really effortless to use and only require very easily available data with few to none transformation to perform.

As a next step, it would be of high interest to apply our pre-trained models on other models such as option pricing ones for example or models with time-dependant parameters.

References

Oldrich Vasicek. An equilibrium characterization of the term structure. *Journal of financial economics*, 5(2):177–188, 1977.

John C. Cox, Jonathan E. Ingersoll, and Stephen A. Ross. A theory of the term structure of interest rates. *Econometrica*, 53(2):385–407, 1985. ISSN 00129682, 14680262. URL `http://www.jstor.org/stable/1911242`.

Damiano Brigo and Fabio Mercurio. *Interest rate models-theory and practice: with smile, inflation and credit*, volume 2. Springer, 2006.

Sébastien Gurrieri, Masaki Nakabayashi, and Tony Wong. Calibration methods of hull-white model. *Available at SSRN 1514192*, 2009.

John Hull and Alan White. The general hull–white model and supercalibration. *Financial Analysts Journal*, 57(6):34–43, 2001.

Olivier Pironneau. Calibration of heston model with keras. 2019.

Andres Hernandez. Model calibration with neural networks. *Available at SSRN 2812140*, 2016.

ML - Graphes - Réseaux

THE DEEP LATENT POSITION MODEL FOR NODE PARTITIONING IN GRAPHS

Dingge Liang [1] & Marco Corneli [1,2] & Charles Bouveyron [1] & Pierre Latouche [3]

[1] *Université Côte d'Azur, INRIA, CNRS, Laboratoire J.A. Dieudonné, Maasai team*
[2] *Université, Côte d'Azur, Center of Modelling, Simulation and Interactions (MSI)*
[3] *Université de Paris, CNRS, Laboratoire MAP5, UMR 8145*

Abstract. With the significant increase of interactions between individuals in different domains, the clustering of vertex in graphs has become a fundamental approach for analysing large and complex networks. We propose here the deep latent position model (DeepLPM), an end-to-end clustering approach which combines the widely used latent position model (LPM) for network analysis with a graph convolutional network (GCN) encoding strategy. Thus, DeepLPM can automatically assign each node to its group without using any additional algorithms and better preserves the network topology. To demonstrate its effectiveness in performing clustering task, an application on the ecclesiastical network in Merovingian Gaul is conducted. Moreover, we inspect the capability of DeepLPM to automatically select the number of clusters, without using external model selection tools.

Keywords. Network Analysis, Graph Clustering, Graph Convolutional Networks

1 Introduction and related work

Networks are used in many applications, from social media and email communications to protein-protein interactions, because they are simple structures yet capable of modelling complex systems. In this context, vertex clustering is a key branch of clustering which attempts to partition the nodes of the graph into different groups to extract patterns summarizing the data.

On the one hand, a long series of statistical methods (Schaeffer, 2007; Snijders, 2011) have been developed to discover the underlying communities in networks by learning the latent features of graph-structured data. The stochastic block model (SBM, Wang and Wong, 1987; Nowicki and Snijders, 2001) is widely used to detect communities or more general clusters of nodes (Lee and Wilkinson, 2019). It assumes that nodes are dispersed into different clusters and that the connection probability between each pair of nodes depends exclusively on their groups. Based on SBM, many extensions looking for overlapping clusters have been proposed, such as the mixed-membership stochastic blockmodel (MMSBM, Airoldi et al., 2008) and the overlapping stochastic blockmodel (OSBM, Latouche et al., 2011). A different approach to model network data relies on latent position models (LPMs), originally proposed by Hoff et al. (2002). The main assumption is that each node has an unknown position in a latent space and that the probability of a specific link between two nodes is modelled by some function of their

positions. Afterwards, the latent position cluster model (LPCM, Handcock et al., 2007) was introduced to incorporate a clustering structure into LPM by considering that the latent position of each node is drawn from a Gaussian mixture model (GMM). Further developments of LPMs exist and the reader is referred to Raftery (2017) for an extensive review.

On the other hand, deep neural network (DNN)-based techniques have recently shown to be effective for representation learning and have been actively explored in clustering (Aljalbout et al., 2018; Zhang et al., 2020). For the purpose of performing node clustering on networks, new models were introduced based on graph neural networks (GNNs). In this line of methods, the variational graph auto-encoder (VGAE, Kipf and Welling, 2016b) adopts a graph convolutional network (GCN, Kipf and Welling, 2016a) encoder to produce nodes embeddings in the latent space and a simple inner product decoder for graph reconstruction. As an extension, Wang et al. (2017) proposed a marginalization process that allows node content to interact with network features. By introducing adversarial learning into the generation process, ARVGA (Pan et al., 2018) enforced the latent representation to match a prior distribution. Lately, DGLFRM (Mehta et al., 2019) combined OSBM with GCN by positing each node of the graph to have an embedding modelled by a Beta-Bernoulli process. All the above mentioned approaches employ inner-product-based decoders, whereas we argue that a different solution, accounting for the Euclidean distance between nodes in the latent space might be more suited. Additionally, these and other existing approaches (Tian et al., 2014; Nie et al., 2017; Zhang et al., 2019) adopt a two-step clustering procedure, simply relying on *external* clustering algorithms (e.g. k-means) to group the embedded nodes, independently from the generative model.

In order to overcome the limitations of the methods listed above, while exploring their benefits, we introduce the deep latent position model (DeepLPM) for network data, allowing to simultaneously learn vertex representations and obtain node partitions. By combining a GCN encoder with a LPM-based decoder, our model aims at capturing the best of both worlds described so far: it is a flexible representation learning tool based on the deep learning architecture, yet comprehensive and interpretable thanks to the statistical model considered.

2 Deep latent position model

Notations. In this work, networks are modelled as undirected, unweighted graphs $G = (V; E)$ with $N = |V|$ nodes. We introduce an $N \times N$ adjacency matrix A, where $A_{ij} = 1$ if there is a link between node i and node j, 0 otherwise. The set of the edges E can be associated with an additional covariate information, collected into the matrix $Y \in \mathbb{R}^{|E| \times D}$. The generic entry of Y, denoted y_{ij}, is the D-dimensional feature associated with the edge connecting i to j. For instance, y_{ij} could encode the text that author i sends to author j in a communication network. We aim to learn well-represented, latent, node embeddings Z in a lower dimension P and to partition the nodes into K clusters.

Generative model. As in LPM (Hoff et al., 2002), we assume that each node i, $i = \{1, \cdots, N\}$, has an unknown position $z_i \in \mathbb{R}^P$ in a latent space and that the edges in the network are sampled independently given these positions. The probability of a link between two individuals is modelled as a function of the distance between the two nodes, in the latent space. The generative process is as follows. First, each node is assigned to a cluster via the random variable c_i

$$c_i \overset{iid}{\sim} \mathcal{M}(1, \pi), \quad \text{with} \quad \pi \in [0,1]^K, \sum_{k=1}^{K} \pi_k = 1. \tag{1}$$

Then, conditionally to its cluster, a latent embedding z_i is generated

$$z_i | (c_{ik} = 1) \sim \mathcal{N}(\mu_k, \sigma_k^2 I_P), \quad \text{with} \quad \sigma_k^2 \in \mathbb{R}^{+*}, \tag{2}$$

independently for each node. Finally, the distance between the latent positions is associated with the probability of a connection through Bernoulli random variables

$$A_{ij} | z_i, z_j \sim \mathcal{B}(f_{\alpha,\beta}(z_i, z_j)), \tag{3}$$

with

$$f_{\alpha,\beta}(z_i, z_j) = \sigma(\alpha + \beta^T y_{ij} - ||z_i - z_j||^2), \tag{4}$$

where $f_{\alpha,\beta}$ can be seen as a *decoding*, one-layer, neural network parametrized by α and β and σ is the logistic sigmoid function and y_{ij} is the covariate of the edge connecting i with j.

3 Model inference

By denoting $\Theta = \{\pi, \mu_k, \sigma_k^2, \alpha, \beta\}$ the set of the model parameters introduced so far, a natural procedure would consist in maximizing $\log p(A|\Theta)$. Unfortunately, this (integrated) log-likelihood is not tractable and we rely on a variational approach to approximate it

$$\log p(A|\Theta) = \mathcal{L}(q(Z, C); \Theta) + D_{KL}(q(Z, C) || p(Z, C|A, \Theta)), \tag{5}$$

where D_{KL} denotes the Kullback-Leibler divergence between the true and approximate posterior distributions of (Z, C). Then, in order to deal with a tractable family of distributions, $q(Z, C)$ is assumed to factorize into two terms (*mean-field* assumption)

$$q(Z, C) = q(Z)q(C) = \prod_{i=1}^{N} q(z_i)q(c_i), \tag{6}$$

where it is also assumed that

$$q(z_i) = \mathcal{N}(z_i; \tilde{\mu}_\phi(\overline{A})_i, \tilde{\sigma}_\phi^2(\overline{A})_i I_P), \tag{7}$$

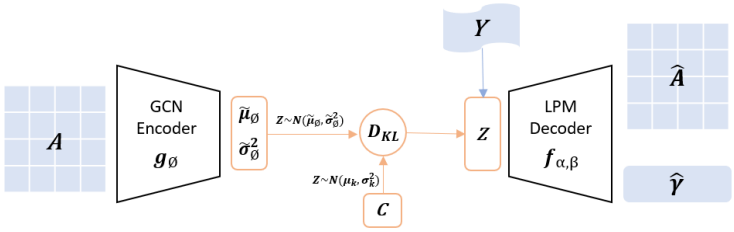


Figure 1: A deep-learning-like model view of DeepLPM.

where $\tilde{\mu}_\phi(\cdot) : \mathbb{R}^{N\times N} \mapsto \mathbb{R}^{N\times P}$ (respectively $\tilde{\sigma}^2_\phi(\cdot) : \mathbb{R}^{N\times N} \mapsto \mathbb{R}^{+N}$) is the function mapping the normalized adjacency $\overline{A} = D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$ (Kipf and Welling, 2016a) into the matrix of the variational means (vector of the standard deviations). In the above equation, $\tilde{\mu}_\phi(\overline{A})_i$ denotes the i-th row of $\tilde{\mu}_\phi(\overline{A})$, corresponding to the variational mean for the latent position z_i (similarly for $\tilde{\sigma}^2_\phi(\overline{A})_i$). The functions $\tilde{\mu}_\phi(\cdot)$ and $\tilde{\sigma}^2_\phi(\cdot)$ are parametrized by the GCN *encoder* g_ϕ.

Finally, a standard assumption is made for the variational clustering probabilities

$$q(C) = \prod_{i=1}^{N} \mathcal{M}(c_i; 1, \gamma_i), \tag{8}$$

where γ_{ik} represents the variational probability that node i is in cluster k, with $\sum_{k=1}^{K} \gamma_{ik} = 1$.

Thanks to Equations (6)-(7)-(8), the evidence lower bound can be further developed as

$$\begin{aligned}\mathcal{L} &= \int_Z \sum_C q(Z,C) \log \frac{p(A|Z,\alpha,\beta)p(Z|C,\mu_k,\sigma^2_k)p(C|\pi)dZ}{q(Z,C)} \\ &= \mathbb{E}\left[\sum_{i\neq j} A_{ij}\log\eta_{ij} + (1-A_{ij})\log(1-\eta_{ij})\right] - \sum_{i=1}^{N}\sum_{k=1}^{K}\gamma_{ik} D_{KL}(\mathcal{N}(\tilde{\mu}_\phi(\overline{A})_i, \tilde{\sigma}^2_\phi(\overline{A})_i I_P)||\mathcal{N}(\mu_k, \sigma^2_k I_P)) \\ &\quad + \sum_{i=1}^{N}\sum_{k=1}^{K}\gamma_{ik}\log(\frac{\pi_k}{\gamma_{ik}}),\end{aligned} \tag{9}$$

where $\eta_{ij} = \sigma(\alpha + \beta^T y_{ij} - ||z_i - z_j||^2)$, $D_{KL}(\cdot)$ denotes the KL divergence and the expectation is taken with respect to the variational probability $q(\cdot)$.

For more clarity, the variational structure of DeepLPM is shown in Figure 1. Within the framework of VAE, first the graph adjacency matrix A is taken as the model input and normalized; then, through the two-layer GCN encoder, we obtain the variational mean and variance of each node; next, by minimizing the KL divergence between the variational and the posterior distributions, we get the learned latent representations; finally, by using the LPM-based decoder, we can reconstruct the matrix $\hat{A}$ and obtain the cluster probability matrix $\hat{\gamma}$.

4 Analysis of a medieval network

As an illustration of the practical use of DeepLPM, it is now fit to a real-world dataset coming from historical science. We consider the data set proposed by Jernite et al. (2014), which reports the ecclesiastical councils that took place in Merovingian Gaul during the 6th century. A council is an ecclesiastical meeting, usually called by a bishop, where issues regarding the church or the faith are addressed. The composition of these councils is known thanks to the acts written at the end of the meeting, and which were signed by all the attending members. The network contains $N = 1287$ individuals who held one or several offices in Gaul between the years 480 and 614, and who either have been related or have at least met during their lifetime. The number of edges is equal to 33,384. Figure 2 shows a visualisation of the network highlighting the importance of the temporality in the relationships. In addition to the interaction data, the data set also contains information about the individuals: period of activity, type of position and location. From this covariate information, we were able to build a 3-dimensional tensor Y encoding the similarities and differences between individuals. Thus, $Y_{ij}^{(1)}$ is equal to the number of years for which i and j have been active at the same time or, alternatively, the negative time lag (in years) between their period of activity; $Y_{ij}^{(2)} = 1$ if i and j were in the same region, -1 otherwise; $Y_{ij}^{(3)} = 1$ if i and j held a similar social position (noble, ecclesiastical or other), -1 otherwise.

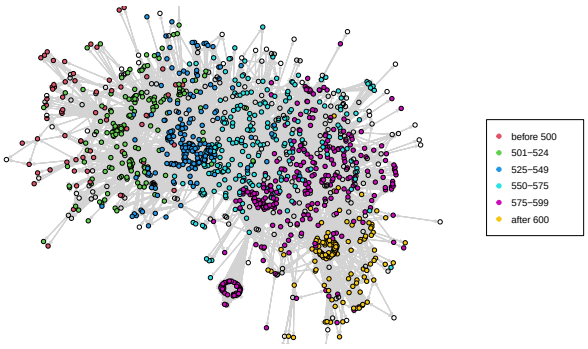


Figure 2: Visualisation of the ecclesiastical network highlighting the temporality of relationships.

Model selection. DeepLPM was fit to this network, with and without considering the covariate tensor Y, for different numbers of groups ranging between 2 and 10, and a fixed dimension ($P = 16$) of the latent space. When we ignore the covariate information, the evolution of the training loss shows a clear minimum at $K = 9$ number of groups. Instead, when edge features are added to the model, the number of groups is estimated to be $K = 8$ with minimal loss. We also plotted the confusion matrix between the predicted labels at $K = 9$ and $K = 8$ to investigate the fusion or dispersion between multiple clusters, as shown in Figure 3. One can see that, by introducing the covariates Y, the cluster $N1$ gathered people from clusters $O3$, $O6$ and

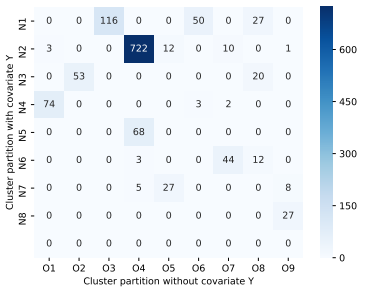


Figure 3: Confusion matrix between cluster partitions with and without covariates.

$O8$ together; on the contrary, it separates the individuals from $O8$ into $N1$, $N3$ and $N6$; and all nodes from $N5$ were coming from $O4$. It is worth mentioning that the exploitation of the covariate information allows DeepLPM to focus on patterns that are not explained by the covariates alone.

Results. We analyse hereafter the clustering results in 8 groups by taking into account the covariates Y. Figure 5 shows the clustering results obtained by DeepLPM. First, it is worth noticing that DeepLPM has not been influenced by the temporality since it was able to detect communities that played a similar role in the network at different periods. Indeed, the groups #1 (black), #3 (green), #5 (cyan) and #7 (yellow) gather people who lived at different and not overlapping time periods. The red group #2 is particularly representative of this since it covers the whole period (480-614 of our era). Figure 6 also shows that group #2 gathers individuals from all strata of society, who do not play a central role in the network. Conversely, combining Figure 6 with the confusion matrix in Figure 3, the group #1 is specific in the sense that it only gathers people from clergy and civilians, who were probably discussing some central questions about the faith during different periods. Moreover, the group #5 is specially separated from the big group $O4$, which is made of a relatively significant proportion of nobles, in particular kings and queens, which suggests that this group was discussing political or nobility matters.

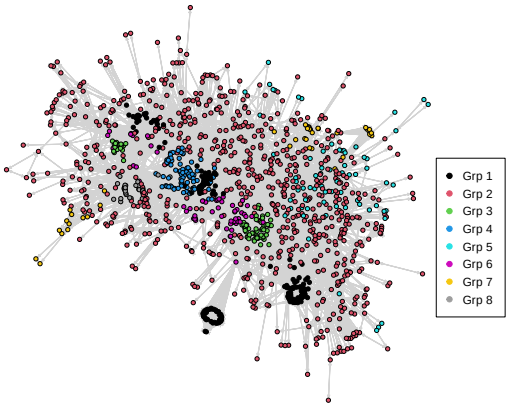


Figure 5: Visualisation of cluster partitions.

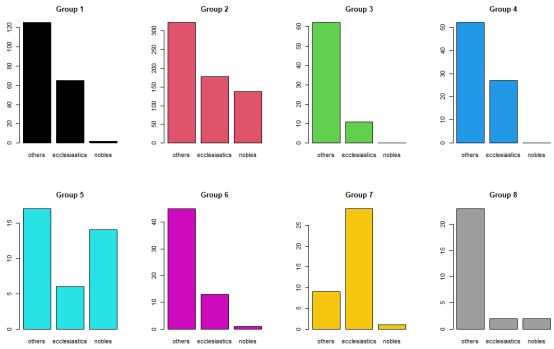


Figure 6: Partitions taking into account types in each group.

5 Conclusion

We introduced DeepLPM to perform node clustering in an end-to-end manner by integrating the GCN encoder with the LPM-based decoder. An application on a historical network illustrates the interests of our methodology for unsupervised learning and highlights its self-regularization ability for selecting the number of groups appropriately.

References

Airoldi, E. M., Blei, D. M., Fienberg, S. E., and Xing, E. P. (2008). Mixed membership stochastic blockmodels. *Journal of machine learning research*.

Aljalbout, E., Golkov, V., Siddiqui, Y., Strobel, M., and Cremers, D. (2018). Clustering with deep learning: Taxonomy and new methods. *arXiv preprint arXiv:1801.07648*.

Handcock, M. S., Raftery, A. E., and Tantrum, J. M. (2007). Model-based clustering for social networks. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(2):301–354.

Hoff, P. D., Raftery, A. E., and Handcock, M. S. (2002). Latent space approaches to social network analysis. *Journal of the american Statistical association*, 97(460):1090–1098.

Jernite, Y., Bouveyron, C., Jegou, L., Lamassé, S., Latouche, P., and Rivera, P. (2014). The random subgraph model for the analysis of an ecclesiastical network in merovingian gaul. *The Annals of Applied Statistics*, 8(1).

Kipf, T. N. and Welling, M. (2016a). Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Kipf, T. N. and Welling, M. (2016b). Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308*.

Latouche, P., Birmelé, E., and Ambroise, C. (2011). Overlapping stochastic block models with application to the french political blogosphere. *The Annals of Applied Statistics*, pages 309–336.

Lee, C. and Wilkinson, D. J. (2019). A review of stochastic block models and extensions for graph clustering. *Applied Network Science*, 4(1):1–50.

Mehta, N., Duke, L. C., and Rai, P. (2019). Stochastic blockmodels meet graph neural networks. In *International Conference on Machine Learning*, pages 4466–4474. PMLR.

Nie, F., Zhu, W., and Li, X. (2017). Unsupervised large graph embedding. In *Thirty-first AAAI conference on artificial intelligence*.

Nowicki, K. and Snijders, T. A. B. (2001). Estimation and prediction for stochastic blockstructures. *Journal of the American statistical association*, 96(455):1077–1087.

Pan, S., Hu, R., Long, G., Jiang, J., Yao, L., and Zhang, C. (2018). Adversarially regularized graph autoencoder for graph embedding. *arXiv preprint arXiv:1802.04407*.

Raftery, A. E. (2017). Comment: Extending the latent position model for networks. *Journal of the American Statistical Association*, 112(520):1531–1534.

Schaeffer, S. E. (2007). Graph clustering. *Computer science review*, 1(1):27–64.

Snijders, T. A. (2011). Statistical models for social networks. *Annual review of sociology*, 37:131–153.

Tian, F., Gao, B., Cui, Q., Chen, E., and Liu, T.-Y. (2014). Learning deep representations for graph clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 28.

Wang, C., Pan, S., Long, G., Zhu, X., and Jiang, J. (2017). Mgae: Marginalized graph autoencoder for graph clustering. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 889–898.

Wang, Y. J. and Wong, G. Y. (1987). Stochastic blockmodels for directed graphs. *Journal of the American Statistical Association*, 82(397):8–19.

Zhang, X., Liu, H., Li, Q., and Wu, X.-M. (2019). Attributed graph clustering via adaptive graph convolution. *arXiv preprint arXiv:1906.01210*.

Zhang, Z., Cui, P., and Zhu, W. (2020). Deep learning on graphs: A survey. *IEEE Transactions on Knowledge and Data Engineering*.

Etude de la Pénalisation GraphNet en Analyse de Données Multi-blocs

Annabelle Beaudoin [1] & Natalia Pietrosemoli [1] & Cathy Philippe [2] & Hervé Abdi [3] & Vincent Guillemot [1]

[1] *Institut Pasteur, Université Paris Cité, Bioinformatics and Biostatistics Hub, F-75015 Paris, France*
[2] *Neurospin, CEA, Gif-Sur-Yvette, FR*
[3] *The University of Texas at Dallas, Richardson, TX, USA*

Résumé

L'intégration de données multiblocs est maintenant incontournable pour analyser des données complexes allant, par exemple, des données multi-omiques, aux données d'imagerie génétique. Par ailleurs, les bases de données biologiques de référence contiennent maintenant une information très riche qu'il convient d'intégrer dans de telles analyses.

Nous proposons d'explorer la méthode netSGCCA qui permet l'intégration de réseaux dans le cadre de l'Analyse des Corrélations Canonique Généralisée pénalisée à l'aide d'une pénalité GraphNet. Plus particulièrement, nous souhaitons mettre en lumière un des désavantages de cette pénalité, qui est d'introduire des composantes "haute-fréquence".

L'exemple que nous étudions est issu d'une étude clinique sur la Spondylarthrite ankylosante et comprend trois blocs : deux blocs de données d'expression, et un bloc de données cliniques. Le réseau de référence que nous utilisons est extrait de la base de données STRING-DB. Nous montrons sur cet exemple un moyen de ne conserver que les éléments "basse fréquence" induits par l'introduction de la pénalité GraphNet.

Mots-clés. Analyse Multi-bloc, Analyse Factorielle, Graphes, Pénalité GraphNet

Abstract

The integration of multiblock data is now an essential tool to analyze the increasingly complex data that bionformaticians, neuroscientists and biostatisticians encounter on a daily basis: from multi-omics data, to imaging-genetic data. Moreover, biological reference databases now contain rich and complex information that should be integrated into such analyses.

Here we explore the netSGCCA method, which allows the integration of networks within the framework of the Generalized Canonical Correlation Analysis penalized using a GraphNet penalty. More particularly, we focus on one of the well-known disadvantages of this penalty, which is to introduce "high-frequency" components.

The example that we present corresponds to a clinical study on spondyloarthritis and comprises three blocks: two blocks of gene expression data and one block of clinical data. The reference network we use is taken from the STRING-DB database. We show on this example a strategy to keep only the "low frequency" elements induced by the introduction of the GraphNet penalty.

Keywords. Multi-block Analysis, Factor Analysis, Graphs, GraphNet Penalty

1 Introduction

L'Analyse Canonique des Corrélations (CCA) [2] ou les méthodes de projection sur variable latente (PLS) [1] sont les méthodes de choix pour analyser deux blocs de données multivariées quantitatives. Plus récemment, RGCCA [10] (*Regularized Generalized Canonical Correlation Analysis*) permet d'étudier les relations entre plus de deux groupes de variables, et de proposer, ainsi, un cadre unificateur pour de nombreuses méthodes d'analyse multiblocs. Lorsque ces blocs possèdent un nombre important de variables, il faut pouvoir sélectionner les variables les plus pertinentes afin de faciliter l'interprétation des résultats. C'est pour cela que SGCCA [9] (*Sparse General Canonical Correlation Analysis*) a été développée: pour permettre, dans le même cadre que RGCCA, d'identifier et de sélectionner des variables pertinentes. Pour ce faire, une contrainte de norme ℓ_1 est ajoutée sur les vecteurs poids du problème RGCCA. Afin de mettre en évidence les interactions entre ces variables, nous ajoutons à SGCCA la pénalité GraphNet [3]. SGCCA avec une pénalité de structure a auparavant été définie dans le cas de deux blocs de variables avec une variable réponse [4], nous appellerons cette méthode netSGCCA dans la suite.

Le but de cette étude est d'intégrer une pénalité de type GraphNet [3] dans une analyse multi-bloc inspirée de RGCCA [5, 4].

2 Méthodes et Données

2.1 Méthode

Soient 3 blocs de variables $\mathbf{X}$, $\mathbf{Y}$ et $\mathbf{Z}$ (standardisées) tels que $\mathbf{X} \in \mathbb{R}^{N\times P}$, $\mathbf{Y} \in \mathbb{R}^{N\times R}$, $\mathbf{Z} \in \mathbb{R}^{N\times Q}$, où N est le nombre d'individus, et P, R, Q sont les nombres de variables de chaque bloc. On note également $f(\mathbf{x},\mathbf{y},\mathbf{z}) = \mathbf{x}^\top\mathbf{X}^\top\mathbf{Y}\mathbf{y} + \mathbf{x}^\top\mathbf{X}^\top\mathbf{Z}\mathbf{z} + \mathbf{y}^\top\mathbf{Y}^\top\mathbf{Z}\mathbf{z}$ la fonction d'interaction entre les blocs de variables où $\mathbf{x}$, $\mathbf{y}$, $\mathbf{z}$, sont les vecteurs poids associés à chaque bloc. Dans la suite, on suppose que le bloc $\mathbf{Y}$ contient des "variables réponse" et on associe à chacun des blocs $\mathbf{X}$ et $\mathbf{Z}$ un graphe non dirigé représentant les interactions entre variables de ces blocs. Soient $\mathbf{L} \in \mathbb{R}^{P\times P}$ et $\mathbf{M} \in \mathbb{R}^{Q\times Q}$ les matrices Laplaciennes associées à ces deux graphes.
On note $\mathcal{P}_\mathbf{L}(\mathbf{x}) = \mathbf{x}^\top\mathbf{L}\mathbf{x}$ et $\mathcal{P}_\mathbf{M}(\mathbf{z}) = \mathbf{z}^\top\mathbf{M}\mathbf{z}$ les pénalités GraphNet associées aux blocs $\mathbf{X}$ et $\mathbf{Z}$ respectivement.

Le problème d'optimisation associé à la méthode netSGCCA est le suivant :

$$\begin{array}{l} \underset{\mathbf{x}\in\mathbb{R}^P,\mathbf{y}\in\mathbb{R}^R,\mathbf{z}\in\mathbb{R}^Q}{\operatorname{argmax}} \; (1-\lambda) f(\mathbf{x},\mathbf{y},\mathbf{z}) + \lambda\left(\gamma \mathcal{P}_{\mathbf{L}}(\mathbf{x}) + (1-\gamma)\,\mathcal{P}_{\mathbf{M}}(\mathbf{z})\right),\ \lambda,\gamma\in[0,1] \\ \text{sous les contraintes :}\ \ \|\mathbf{x}\|_2^2 \le 1, \|\mathbf{y}\|_2^2 \le 1, \|\mathbf{z}\|_2^2 \le 1,\ \|\mathbf{x}\|_1 \le s_1 \text{ et } \|\mathbf{z}\|_1 \le s_2 \end{array} \tag{1}$$

où s_1 et s_2 sont des paramètres permettant d'obtenir des vecteurs $\mathbf{x}$ et $\mathbf{z}$ parcimonieux, λ et γ permettent de réguler les pénalités GraphNet. Enfin, l'algorithme utilisé pour résoudre ce problème d'optimisation est donné ci-dessous. Il s'agit d'un algorithme des moindres carrés alternés consistant, à chaque itération, à fixer les poids de deux blocs pour mettre à jour les poids du bloc courant, en les projetant sur l'espace convexe défini par les contraintes (fonction proj) de sorte à obtenir des poids parcimonieux et normés. Comme la fonction objectif de netSGCCA reste convexe et différentiable, et que les contraintes sont les mêmes que celles de SGCCA, netSGCCA hérite des propriétés de convergence de SGCCA.

Algorithme 1 : Algorithme netSGCCA

Entrées : $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{Z}$, ε, λ, γ, $t_{\max}$;
Initialisation des poids : $\mathbf{x}^{(0)}, \mathbf{y}^{(0)}, \mathbf{z}^{(0)}$;
pour $t = 1, \cdots, t_{max}$ **faire**
  si $\|\mathbf{x}^{(t)} - \mathbf{x}^{(t-1)}\|_2 < \varepsilon$ *et* $\|\mathbf{y}^{(t)} - \mathbf{y}^{(t-1)}\|_2 < \varepsilon$ *et* $\|\mathbf{z}^{(t)} - \mathbf{z}^{(t-1)}\|_2 < \varepsilon$ **alors**
    retourner $\mathbf{x}^{(t)}, \mathbf{y}^{(t)}, \mathbf{z}^{(t)}$
  sinon
    $\nabla_{\mathbf{x}} = (1-\lambda)(\mathbf{X}^\top\mathbf{Y}\mathbf{y}^{(t-1)} + \mathbf{X}^\top\mathbf{Z}\mathbf{z}^{(t-1)}) + 2\lambda\gamma\mathbf{L}\mathbf{x}^{(t-1)}$
    $\mathbf{x}^{(t)} = \text{proj}(\nabla_{\mathbf{x}})$
    $\nabla_{\mathbf{y}} = (1-\lambda)(\mathbf{Y}^\top\mathbf{X}\mathbf{x}^{(t-1)} + \mathbf{Y}^\top\mathbf{Z}\mathbf{z}^{(t-1)})$
    si $\lambda = 1$ **alors**
      $\mathbf{y}^{(t)} = \mathbf{0}$
    sinon
      $\mathbf{y}^{(t)} = \nabla_{\mathbf{y}}/\|\nabla_{\mathbf{y}}\|_2$
    $\nabla_{\mathbf{z}} = (1-\lambda)(\mathbf{Z}^\top\mathbf{X}\mathbf{x}^{(t-1)} + \mathbf{Z}^\top\mathbf{Y}\mathbf{y}^{(t-1)}) + 2\lambda(1-\gamma)\mathbf{M}\mathbf{z}^{(t-1)}$
    $\mathbf{z}^{(t)} = \text{proj}(\nabla_{\mathbf{z}})$
fin
Sorties : $\mathbf{x}^{(t)}$, $\mathbf{y}^{(t)}$, $\mathbf{z}^{(t)}$

2.2 Données

La méthode netSGCCA a été appliquée sur des données issues d'une étude portant sur l'immunorégulation [6] de $N = 80$ patients souffrant de spondylarthrite ankylosante. Nous avons extrait 3 blocs de variables de cette étude : (1) un bloc $\mathbf{X}$ contenant l'expression de $P = 277$ gènes après une stimulation par lipopolysaccharides (LPS, qui représente, pour simplifier, l'immunité "innée") (2) un bloc $\mathbf{Y}$ composé de $R = 3$ scores d'évaluation clinique, et (3) un bloc $\mathbf{Z}$ contenant l'expression de $Q = 283$ gènes après une stimulation

par entérotoxine staphylococcique B (SEB, représentant, pour simplifier, l'immunité "acquise"). Il y a 277 gènes en commun entre les blocs **X** et **Z**. La base de données publique des interactions entre protéines utilisée pour construire les réseaux qui servent à pénaliser les blocs **X** et **Z** est STRING-DB [8].

3 Résultats

La Figure 1 représente le réseau de gènes pour chaque stimulation avec des valeurs de λ différentes mais aussi avec deux matrices laplaciennes **L** et **M** différentes pour la pénalité de structure. Les nœuds avec un poids nul ont été retirés.

Une seule valeur de γ ($= 0.5$) a été utilisée: nous avons considéré que les blocs **X** et **Z** devaient être tous les deux pénalisés de façon équivalente. De même, nous avons fixé $s_1 = 0.5\sqrt{P}$ et $s_2 = 0.5\sqrt{Q}$. De plus, nous avons utilisé trois valeurs de λ différentes afin de dissocier la situation où il n'y a pas de pénalités de graphes ($\lambda = 0$) de la situation où il n'y a pas d'interaction entre les blocs de variables (λ=1) et d'une situation compromis avec $\lambda = 0.9998$ (une valeur inférieure ne donne pas de résultats très différents de ceux obtenus pour $\lambda = 0$). On remarque que lorsque λ est nul, de nombreuses variables s'éteignent et donc que la solution est bien parcimonieuse. Pour $\lambda = 0.9998$ et pour $\lambda = 1$, nous distinguons deux cas. Dans le premier cas, nous utilisons les matrices laplaciennes normalisées associées à chaque bloc de variables dans la pénalité de graphe et dans le second cas nous filtrons les valeurs propres de ces matrices [7] de sorte à ne retenir que celles qui sont inférieures à 1, ce qui a pour effet de privilégier les variations basse fréquence des poids sur le graphe STRING-DB.

4 Conclusion et perspectives

La méthode que nous utilisons, netSGCCA, permet d'obtenir des résultats prometteurs garantissant d'une part la parcimonie des solutions obtenues ainsi qu'une plus grande interprétabilité grâce aux propriétés de la pénalité GraphNet, après avoir filtré le Laplacien du graphe concerné par la pénalité.

De nouvelles pistes d'étude incluent: (i) comment définir des valeurs optimales des paramètres de la pénalité GraphNet, ou encore (ii) comment enrichir les critères permettant de choisir le nombre de valeurs propres à sélectionner pour avoir des résultats basses fréquences.

References

[1] Hervé Abdi. Partial least square regression, projection on latent structure regression, PLS-Regression. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2:97–

106, 2010.

[2] Hervé Abdi, Aida Eslami, and Vincent Guillemot. Canonical correlation analysis (cca). In R. Alhajj and J. Rokne, editors, *Encyclopedia of Social Networks and Mining (2nd Edition)*. Springer Verlag, New York, New York, USA, 2018.

[3] Logan Grosenick, Brad Klingenberg, Kiefer Katovich, Brian Knutson, and Jonathan E. Taylor. Interpretable whole-brain prediction analysis with GraphNet. *NeuroImage*, 72:304–321, 2013.

[4] N. Guigui, C. Philippe, A. Gloaguen, S. Karkar, V. Guillemot, T. Löfstedt, and V. Frouin. Network regularization in imaging genetics improves prediction performances and model interpretability on alzheimer's disease. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 1403–1406, 2019.

[5] Tommy Löfstedt, Fouad Hadj-Selem, Vincent Guillemot, Cathy Philippe, Edouard Duchesnay, Vincent Frouin, and Arthur Tenenhaus. Structured Variable Selection for Regularized Generalized Canonical Correlation Analysis. *Springer Proceedings in Mathematics and Statistics*, 173:129–139, 2014.

[6] Silvia Menegatti, Vincent Guillemot, Eleonora Latis, Hanane Yahia-Cherbal, Daniela Mittermüller, Vincent Rouilly, Elena Mascia, Nicolas Rosine, Surya Koturan, Gael A. Millot, Claire Leloup, Darragh Duffy, Aude Gleizes, Salima Hacein-Bey-Abina, Jérémie Sellam, Francis Berenbaum, Corinne Miceli-Richard, Maxime Dougados, Elisabetta Bianchi, and Lars Rogge. Immune response profiling of patients with spondyloarthritis reveals signalling networks mediating TNF-blocker function in vivo. *Annals of the Rheumatic Diseases*, 80(4):475–486, apr 2021.

[7] Franck Rapaport, Andrei Zinovyev, Marie Dutreix, Emmanuel Barillot, and Jean-Philippe Vert. Classification of microarray data using gene networks. *BMC bioinformatics*, 8:35, feb 2007.

[8] Damian Szklarczyk, Annika L. Gable, Katerina C. Nastou, David Lyon, Rebecca Kirsch, Sampo Pyysalo, Nadezhda T. Doncheva, Marc Legeay, Tao Fang, Peer Bork, Lars J. Jensen, and Christian von Mering. The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic acids research*, 49(D1):D605–D612, jan 2021.

[9] Arthur Tenenhaus, Cathy Philippe, Vincent Guillemot, Kim-Anh Le Cao, Jacques Grill, and Vincent Frouin. Variable selection for generalized canonical correlation analysis. *Biostatistics (Oxford, England)*, 15(3):569–83, jul 2014.

[10] Arthur Tenenhaus and Michel Tenenhaus. Regularized Generalized Canonical Correlation Analysis. *Psychometrika*, 76(2):257–284, mar 2011.

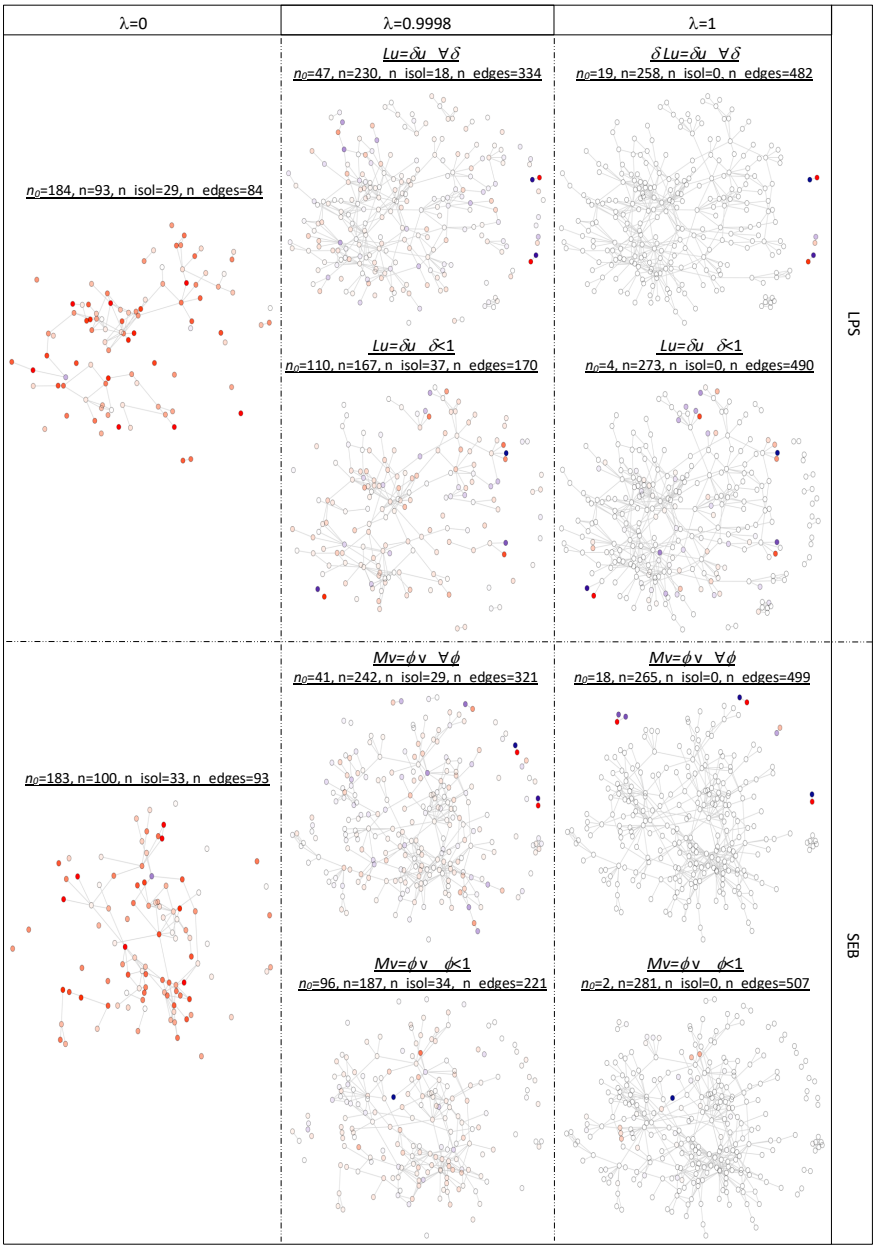


Figure 1: Réseaux de gènes pour différentes valeurs de λ avec les stimulations LPS (bloc $\mathbf{X}$, $P = 277$) et SEB (bloc $\mathbf{Z}$, $Q = 283$). Les variables n_0, n, n_isol et n_edges sont respectivement le nombre de poids nuls dans le vecteur de poids, le nombre de poids non nuls, le nombre de noeuds isolés et le nombre d'arêtes. δ et ϕ sont respectivement les valeurs propres des laplaciens $\mathbf{L}$ et $\mathbf{M}$. Le gradient de couleur représente la proximité des poids par rapport au zéro. Plus le noeud est foncé, plus son poids est éloigné de zéro. Plus il est rouge, plus son poids est positif. Plus il est bleu, plus son poids est négatif.

Don't fear the unlabelled: safe deep semi-supervised learning via simple debiasing

Hugo Schmutz [1,2,4,5], Olivier Humbert [1,3,4] & Pierre-Alexandre Mattei [1,2]

[1] *Universite Côte d'Azur (UCA)*
[2] *Inria (Maasai team), Laboratoire J.A. Dieudonné, CNRS*
[3] *Department of Nuclear Medicine, Centre Antoine-Lacassagne*
[4] *TIRO-UMR E 4320, CEA*
[5] *hugo.schmutz@inria.fr*

Abstract. Semi supervised learning (SSL) provides an effective means of leveraging unlabelled data to improve a model's performance. Nevertheless, most methods present the common drawback of being unsafe. By safeness we mean the quality of not degrading a fully supervised model when including unlabelled data. Our starting point is to notice that the estimate of the risk that most discriminative SSL methods minimise is biased, even asymptotically. This bias makes these techniques untrustable without a proper validation set, but we propose a simple way of removing the bias. Our debiasing approach is straightforward to implement, and applicable to most deep SSL methods. We provide simple theoretical guarantees on the safeness of these modified methods, without having to rely on the strong assumptions on the data distribution that SSL theory usually requires. We show experimentally that debiasing can compete with classic deep SSL techniques in various classic settings and even performs well when traditional SSL fails.

Keywords. Semi-supervised learning, deep learning theory, debiasing, controle variates

Resumé. L'apprentissage semi-supervisé (SSL) offre un moyen efficace d'exploiter les données non étiquetées pour améliorer les performances d'un modèle. Néanmoins, la plupart des méthodes présentent l'inconvénient commun de manquer de securité. Par sécurité, nous entendons la qualité de ne pas dégrader un modèle entièrement supervisé lors de l'inclusion de données non étiquetées. Notre idée a été de remarquer que l'estimateur du risque que la plupart des méthodes SSL discriminatives minimisent est biaisé, même asymptotiquement. Ce biais rend ces techniques non fiables sans un jeu de validation approprié, mais nous proposons un moyen simple de supprimer ce biais. Notre approche de débiasage est simple à mettre en œuvre et applicable à la plupart des méthodes SSL profondes. Nous fournissons des garanties théoriques simples sur la sécurité de ces méthodes modifiées, sans avoir à s'appuyer sur les hypothèses fortes sur la distribution des données que le SSL exige habituellement. Nous montrons expérimentalement que le débiaisage peut rivaliser avec les techniques SSL profondes classiques dans divers contextes classiques et qu'il donne même de bons résultats lorsque le SSL traditionnel échoue.

Mots clés. Apprentissage semi-supervisé, Apprentissage profond, débiaisage, variables de contrôles

1 Introduction

An obvious SSL baseline is to simply throw away the unlabelled data. We will call such a baseline the *complete case*, following the missing data literature (e.g. Tsiatis, 2006). As reported in van Engelen and Hoos (2020), the main risk of SSL is the potential degradation caused by the introduction of unlabelled data. To overcome this issue, previous works introduced the notion *safe* semi-supervised learning for techniques which never reduce predictive performance by introducing unlabelled data (Trapp et al., 2017; Guo et al., 2020). Our loose definition of safeness is as follows: *a SSL algorithm is safe if it has theoretical guarantees that are similar or stronger to the complete case baseline.* The "theoretical" part of the definition is motivated by the fact that any empirical assessment of generalisation performances of an SSL algorithm is jeopardised by the scarcity of labels. Indeed, with few labels, it is risky to rely too much on a very small validation set.

We distinct three types of methods for deep SSL:

- Entropy minimisation: Grandvalet and Bengio (2004) simply use the Shannon entropy as a penalisation. It can be interpreted as favouring parameters inducing small class overlap (Chapelle et al., 2006, Chapter 9).

- Pseudo-labelling: It consists in picking the class with the maximum predicted probability as a pseudo-label for the unlabelled data (Scudder, 1965; Lee, 2013; Sohn et al., 2020; Rizve et al., 2021) .

- Consistency-based: Another range of SSL methods minimise a consistency objective that encourages invariant prediction for perturbations either on the data either on the model (Sajjadi et al., 2016; Laine and Aila, 2017; Tarvainen and Valpola, 2017; Miyato et al., 2018). In particular, consistency-based method aim to smooth the decision function of the models or have more stable predictions.

Unfortunately, popular deep SSL techniques are generally not safe. Indeed, they rely on strong and essentially untestable assumptions on the data distribution to ensure performance in entropy minimisation, pseudo-labelling and consistency-based methods. However, no proof is given that guarantees the effectiveness of state-of-the-art methods Tarvainen and Valpola (2017); Miyato et al. (2018); Sohn et al. (2020); Pham et al. (2021).

Contributions Rather than relying on the strong geometric assumptions usually used in SSL theory, we simply use the *missing completely at random (MCAR)* assumption, a standard assumption from the missing data literature (see e.g. Little and Rubin, 2019). With this only assumption on the data distribution, we propose a new safe SSL method derived from simply debiasing common SSL risk estimates. Our main contributions are:

- We introduce debiased SSL (DeSSL), a safe method that can be applied to most deep SSL algorithms;

- We propose a theoretical explanation of the intuition of the popular SSL methods. We provide theoretical guarantees on the safeness of using DeSSL both on consistency and calibration of the method. We also provide a generalisation error bound;
- We show how simple it is to apply DeSSL to existing methods and show empirically that DeSSL leads to models that are never worse than their classical counterparts, generally better calibrated and sometimes much more accurate.

2 Semi-supervised learning

The ultimate objective of most of the learning frameworks is to minimise a risk $\mathcal{R}$, defined as the expectation of a particular loss function L over a data distribution $p(x, y)$, on a set of models $f_\theta(x)$, parametrised by $\theta \in \Theta$. The distribution $p(x, y)$ being unknown, we generally minimise an estimate of the risk. Classicaly, it is preferable that this estimator is unbiased. Indeed, its unbiased nature is one of the basic properties that is used for the development of traditional learning theory and asymptotic statistics (van der Vaart, 2000; Shalev-Shwartz and Ben-David, 2014).

Complete case: throwing the unlabelled data away The natural estimator of the risk is then simply the empirical risk computed on the labelled data. Fortunately, in the MCAR setting, the complete case risk estimate is unbiased and converges pointwisely to $\mathcal{R}(\theta)$. Therefore, traditional learning theory holds for the complete case under MCAR.

$$\hat{\mathcal{R}}_{CC}(\theta) = \frac{1}{n_l} \sum_{i=1}^{n_l} L(\theta; x_i, y_i). \tag{1}$$

In real-world applications, validation sets are not large enough to ensure the generalisation of the trained model and therefore, the introduction of unlabelled data may come at the risk of degrading performances. Then, theoretical guarantees are required to use safely SSL algorithms. For this reason, in our work, we consider as *safe* a SSL algorithm that has theoretical guarantees that are similar or stronger than those of the complete case baseline.

Incorporating unlabelled data A class of SSL approaches generally aim to minimise a modified estimator of the risk by including unlabelled data. Therefore, the optimisation problem generally becomes finding $\hat{\theta}$ that minimises the SSL risk,

$$\hat{\mathcal{R}}_{SSL}(\theta) = \frac{1}{n_l} \sum_{i=1}^{n_l} L(\theta; x_i, y_i) + \frac{\lambda}{n_u} \sum_{i=1}^{n_u} H(\theta; x_i), \tag{2}$$

where H is a term that does not depend on the labels and λ is a scalar weight which balances the labelled and unlabelled terms. In the literature, H can generally be seen as a surrogate of L. Indeed, it looks like the intuitive choices of H are equal or equivalent to a form of expectation of L on a distribution given by the model. Our theory will indicate that it is desirable, indeed, that H is positively correlated with L.

DeSSL: Unbiased semi-supervised learning In order to overcome the issues introduced by the second term in the approximation of the risk for the semi-supervised learning approach, we propose DeSSL, an unbiased version of the SSL estimator using labelled data to annul the bias. The idea here is to retrieve the properties of classical learning theory. Fortunately, we will see that the proposed method can eventually have better properties than the complete case, in particular with regards to the variance of the estimate. The proposed DeSSL objective is

$$\hat{\mathcal{R}}_{DeSSL}(\theta) = \frac{1}{n_l}\sum_{i=1}^{n_l} L(\theta; x_i, y_i) + \frac{\lambda}{n_u}\sum_{i=1}^{n_u} H(\theta; x_i) - \frac{\lambda}{n_l}\sum_{i=1}^{n_l} H(\theta; x_i). \tag{3}$$

Under the MCAR assumption, this estimator is unbiased for any value of the parameter λ. This way of debiasing is closely related to the method of control variates (Owen, 2013, Chapter 8) which is a common variance reduction technique. The idea is to add an additional term to a Monte-Carlo estimator with a null expectation in order to reduce the variance of the estimator without modifying the expectation.

3 Experiments

We evaluate the performance of DeSSL against different classic methods. In the experiments, our goal is not to produce state-of-the-art results, but instead to compare DeSSL methods and their orginal counterparts. In particular, we perform experiments with simple SSL methods such as pseudo-label (PseudoLabel) and entropy minimisation (EntMIN) with varying λ on CIFAR-10 Krizhevsky (2009) and compare them to the debiased method, respectively DeEntMin and DePseudoLabel.

We trained a CNN-13 from Tarvainen and Valpola (2017) on 5 different splits of the training dataset into a labelled and unlabelled set. For this experiement, we use $n_l = 4000$ and use the rest of the dataset as unlabelled. Models are then evaluated using the standard $10,000$ test samples. For a more realistic validation set, we used 10% of n_l as the validation set. We test the influence of the hyperparameter λ and report the accuracy and the loss at the epoch of best validation accuracy, see Figure 1.

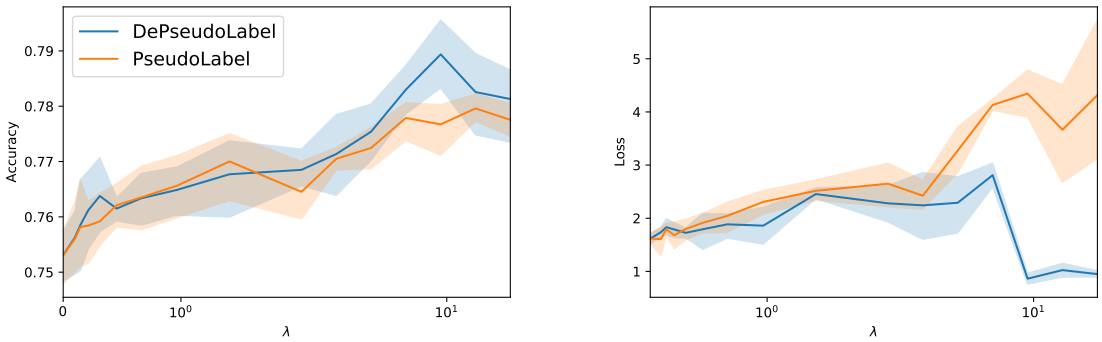


Figure 1: The influence of λ on Pseudo-label and DePseudo-label for a CNN trained on CIFAR with $n_l = 4000$: (Left) Mean test accuracy; (Right) Mean test loss, with 95% CI.

We observe DeSSL provides both a better loss with the same accuracy for small λ. For larger λ, DeSSL performs better in all the reported metrics. We perfomed a paired Student's t-test to ensure that our results are significant. The p-values indicates that for λ close to 10, DeSSL is often significantly better in all the metrics. Moreover, DeSSL for large λ provides a better loss than the complete case whereas SSL never does.

References

Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien. Semi-supervised learning. *MIT Press*, 2006.

Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. *Advances in Neural Information Processing Systems*, 2004.

Lan-Zhe Guo, Zhen-Yu Zhang, Yuan Jiang, Yu-Feng Li, and Zhi-Hua Zhou. Safe deep semi-supervised learning for unseen-class unlabeled data. *International Conference on Machine Learning*, 2020.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, MIT, NYU, 2009.

Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *International Conference on Learning Representations*, 2017.

Dong-Hyun Lee. Pseudo-Label : The simple and efficient semi-supervised learning method for deep neural networks. *Workshop on challenges in representation learning, International conference on machine learning*, 2013.

Roderick JA Little and Donald B Rubin. *Statistical Analysis with Missing Data*. John Wiley & Sons, 2019.

Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41:1979–1993, 2018.

Art B. Owen. *Monte Carlo theory, methods and examples.* 2013.

Hieu Pham, Zihang Dai, Qizhe Xie, and Quoc V Le. Meta pseudo labels. *Conference on Computer Vision and Pattern Recognition*, 2021.

Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. In Defense of Pseudo-Labeling: An Uncertainty-Aware Pseudo-label Selection Framework for Semi-Supervised Learning. *International Conference on Learning Representations*, 2021.

Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. *Advances in Neural Information Processing Systems*, 2016.

Henry Scudder. Probability of error of some adaptive pattern-recognition machines. *IEEE Transactions on Information Theory*, 11:363–371, 1965.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms.* Cambridge university press, 2014.

Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. FixMatch: Simplifying semi-supervised learning with consistency and confidence. *Avances in Neural Information Processing Systems*, 2020.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advancer in Neural Information Processing Systems*, 2017.

Martin Trapp, Tamas Madl, Robert Peharz, Franz Pernkopf, and Robert Trappl. Safe semi-supervised learning of sum-product networks. *Conference on Uncertainty in Artificial Intelligence*, 2017.

Anastasios A Tsiatis. *Semiparametric theory and missing data.* Springer, 2006.

Aad W van der Vaart. *Asymptotic statistics.* Cambridge university press, 2000.

Jesper E van Engelen and Holger H Hoos. A survey on semi-supervised learning. *Machine Learning*, 109:373–440, 2020.

Generalised Mutual Information for Discriminative Clustering

Louis Ohl [1] & Pierre-Alexandre Mattei [2] & Warith Harchaoui [3] & Charles Bouveyron [4] & Frédéric Precioso [5]

[1] *I3S - UMR7271 - CNRS, Maasai project-team / louis.ohl@inria.fr*
[2] *INRIA, Maasai project-team / pierre-alexandre.mattei@inria.fr*
[3] *DERAISON.ai / warith@deraison.ai*
[4] *LJAD - UMR7375 - CNRS, Maasai project-team / charles.bouveyron@inria.fr*
[5] *I3S - UMR7271 - CNRS, Maasai project-team / frederic.precioso@univ-cotedazur.fr*

Résumé. Au cours des dernières années, le problème de partionnement des données a été abordé avec succès par les méthodes d'apprentissage profond. Cependant, la plupart de ces méthodes font intervenir la maximisation de l'information mutuelle comme objectif d'entrainement et font appel à des techniques de régularisations sophistiquées pour corriger son instabilité plutôt que de chercher à l'améliorer. Nous prenons un autre point de vue sur l'information mutuelle et la considérons comme une distance moyenne entre les distributions des classes. Nous étendons sa définition pour inclure d'autres distances que celle de Kullback-Leibler entre les distributions: l'information mutuelle généralisée (GEMINI). Nos expériences soulignent qu'un choix judicieux de GEMINI peu améliorer le partitionnement lorsque l'information mutuelle est utilisée comme objectif tout en étant un outil efficace pour trouver un bon nombre de partitions.

Mots-clés. Partitionnement, Apprentissage Profond, Théorie de l'information ...

Abstract. In the last decade, deep learning methods started to extend to clustering with successful applications, notably regarding images. However, most of these methods rely on maximising the mutual information as a clustering objective and require strong regularisation to face its instability rather than improving it. We take here another point of view on the mutual information by considering it as an average distance between clusters and extend its definition to include multiple statistical distances: the GEneralised Mutual INformatIon (GEMINI). Our experiments highlight that well-selected GEMINI models can improve the clustering results when using mutual information as a training objective, while also implicitly providing an efficient tool to choose a relevant number of clusters.

Keywords. Clustering, Deep Learning, Information Theory...

1 Motivations

Clustering is a fundamental task in machine learning which aims at grouping a set of N elements into K groups named clusters.

To that end, model-based algorithms make assumptions on the true distribution of the data as a result of some latent distribution of clusters. These models are then able to find the most likely cluster assignment.

Such procedure is called generative: assumptions on the data distribution $p(x|y)$ given the cluster assignment $p(y)$ models a joint distribution $p(x, y)$. However, if we do not make assumptions on the distribution of the data given the latent cluster assignment we turn to discriminative modelling. In the discriminative school, we aim at modelling the distribution of the cluster assignment given the data $p(y|x)$. Such models avoid the burden of computing the $p(x)$ for inferring the cluster assignment of a sample x. For such model standard methods like maximum likelihood or Bayesian inference cannot be applied: the absence of hypothesis regarding the data distribution given a cluster assignment $p(x|y)$ makes the evidence $p(x)$ impossible to compute.

In the recent years, discriminative clustering was leveraged with neural networks as main models able to directly infer the cluster assignment of samples. To train such models authors often rely on the mutual information maximisation (Krause et al., 2010; Hu et al., 2017; Ji et al., 2019; Ghasedi Dizaji et al., 2017), Eq. (1), as an objective which describes how non independent the data distribution $p(\boldsymbol{x})$ and the cluster assignment $p(y)$ are:

$$\mathcal{I}(X;Y) = D_{\mathrm{KL}}(p(\boldsymbol{x}, y)||p(\boldsymbol{x})p(y)) \quad , \tag{1}$$

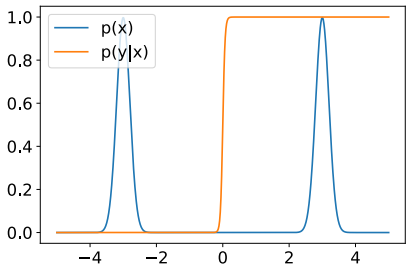
where D_{KL} is the Kullback-Leibler divergence.

One crucial question in clustering is to know how many clusters there are inside a dataset. This question has often been overlooked in related works, implying that clustering methods were tested with models trying to find the same number of clusters as the number of classes given by some manual annotations.

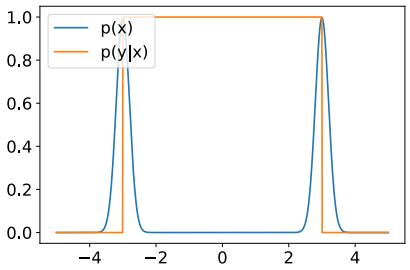
We argue that the success of these methods cannot solely be granted to the mutual information, and is more owing to careful designs of regularisation biasing the models enough to the data at hand.

2 Contributions

We focus on the mutual information by questioning the Kullback-Leibler distance at its core. Consider for example Fig. (1) where we try to find the correct decision boundary modelling by $p(y|x)$ to split two Gaussian distributions in a mixture model. We can propose two different boundaries that both maximise the mutual information. The first one (Fig. 1a) is deemed appropriate and can be obtained using a sigmoid posterior with



(a) The correct and expected decision boundary



(b) A non linear inappropriate decision boundary

Figure 1: Two different decision boundaries for separating a gaussian mixture with 2 disitributions, both maximising the mutual information

sharp parameters. The second one however is inappropriate (Fig. 1b) but still maximises the mutual information.

Therefore we introduce a generalisation of the mutual information by replacing the Kullback-Leibler divergence for other divergences and distances (Sriperumbudur et al., 2009). We call it the GEnaralised Mutual INformatIon (GEMINI). We motivate that the choice of some specific distances that carry knowledge about the data can enhance the notion of mutual information and make it a more relevant clustering objective for training discriminative models.

We can indeed rewrite the mutual information with Eq. (2) which reveals the expectation of a divergence between cluster distributions and the data distribution:

$$\mathcal{I}(X;Y) = \mathbb{E}_{y\sim p(y)}\left[D_{\mathrm{KL}}(p(x|y)||p(x))\right] \quad . \tag{2}$$

We replace in this equation D_{KL} by any other distance that we note D. We found the total variation (TV), squared Hellinger distance, maximum mean discrepancy (MMD) (Gretton et al., 2012) also known as kernel scoreGneiting and Raftery (2007) or Wasserstein distance to yield appropriate objectives.

Moreover we argue that the current definition of the mutual information can be more fine-grained for clustering purposes. Indeed the mutual information seeks to maximise the distance between the distribution of the data according to one cluster $p(x|y)$ and the overall data distribution $p(x)$ of which the first one is already part. We find more relevant to actually compare the distances between the data distribution according to one cluster $p(x|y_1)$ and the distribution of another cluster $p(x|y_2)$. We hence call the Eq. (3) the One-vs-All GEMINI and Eq. (4) the One-vs-One GEMINI:

$$\mathcal{I}_D^{\mathrm{ova}}(X:Y) = \mathbb{E}_{y\sim p(y)}\left[D(p(x|y)||p(x))\right] \quad , \tag{3}$$

$$\mathcal{I}_D^{\text{ovo}}(X;Y) = \mathbb{E}_{y_1,y_2\sim p(y)}\left[D(p(x|y_1)||p(x|y_2))\right] \quad . \tag{4}$$

Using these namings, the traditional mutual information of Eq. (2) is now referred to as KL-OvA.

3 Experiments

We experimented our new objectives using the protocol proposed by Hu et al. (2017) in Tab. 1, i.e. training a neural network to perform clustering on the MNIST dataset. Training is compared when using only the GEMINI as a clustering objective and when using both GEMINI with Virtual Adversarial Training (Miyato et al., 2018), a regularisation method that forces decision boundaries to be more consistent around dense areas in data space. However, we also compared the results when using a LeNet-5 architecture (LeCun et al., 1998) additional to the proposed Multi-Layered Perceptron of Hu et al. (2017). Eventually, we explored the performances of the models when we task it to find more than 10 clusters in Tab. 2, i.e. the number of labelled classes in MNIST.

Table 1: ARI for deep neural network trained with GEMINIs on MNIST for 500 epochs. Models were trained either with or without the Virtual Adversarial Training regularisation. The objective is 10 clusters.

GEMINI		MLP		LeNet-5	
		w/ VAT	VAT	w/ VAT	VAT
KL	OvA	0.320	0.601	0.132	0.816
	OvO	0.348	0.431	0.123	0.569
Squared Hellinger	OvA	0.301	0.712	0.207	**0.922**
	OvO	0.287	0.707	0.161	0.706
TV	OvA	0.299	0.665	0.171	0.820
	OvO	0.422	**0.765**	0.161	0.744
MMD	OvA	0.373	0.591	0.382	0.404
	OvO	0.361	0.506	0.373	0.399
Wasserstein	OvO	**0.445**	0.627	**0.383**	0.575

We further highlight that mutual information requires the specification of a *maximum number of clusters* instead of a strict number to find. The models thus converge to different varying number of clusters depending on the chosen GEMINI and neural network architecture. Overall, models trained using the mutual information without any regularisation possess cluster emptying properties.

Table 2: ARI for deep neural network trained with GEMINIs on MNIST for 500 epochs with more clusters than the number of classes: 20. For computational complexity reasons, we tasked the Wasserstein OvO with 15 clusters to find.

GEMINI		MLP		LeNet-5	
		w/ VAT	VAT	w/ VAT	VAT
KL	OvA	0.320	0.435	0.124	0.549
	OvO	0.413	0.581	0.154	0.551
Squared Hellinger	OvA	0.295	0.627	0.212	**0.708**
	OvO	0.280	0.669	0.166	0.400
TV	OvA	0.287	**0.743**	0.190	0
	OvO	0.284	0.589	0.175	0.682
MMD	OvA	0.375	0.509	0.379	0.402
	OvO	0.381	0.576	0.370	0.380
Wasserstein	OvO	**0.423**	0.734	**0.414**	0.599

4 Conclusion

We replaced the Kullback-Leibler divergence at the core of mutual information by other distances and extended the definition to include cluster-wise distribution comparisons, named OvO GEMINI. These extensions bring improvement in the mutual information and allow to use it as more expressive objective for discriminative clustering, i.e. without any assumption regarding the data distribution.

Our future works should focus on the cluster-emptying ability of the generalised mutual information when used as an objective since it leverages implicit model selection in clustering.

Bibliography

References

Kamran Ghasedi Dizaji, Amirhossein Herandi, Cheng Deng, Weidong Cai, and Heng Huang. Deep clustering via joint convolutional autoencoder embedding and relative entropy minimization. In *Proceedings of the IEEE international conference on computer vision*, pages 5736–5745, 2017.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexan-

der Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13 (1):723–773, 2012. Publisher: JMLR. org.

Weihua Hu, Takeru Miyato, Seiya Tokui, Eiichi Matsumoto, and Masashi Sugiyama. Learning Discrete Representations via Information Maximizing Self-Augmented Training. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1558–1567, International Convention Centre, Sydney, Australia, August 2017. PMLR.

Xu Ji, Joao F Henriques, and Andrea Vedaldi. Invariant information clustering for unsupervised image classification and segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9865–9874, 2019.

Andreas Krause, Pietro Perona, and Ryan Gomes. Discriminative Clustering by Regularized Information Maximization. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993, 2018. Publisher: IEEE.

Bharath K Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert RG Lanckriet. On integral probability metrics,ϕ-divergences and binary classification. *arXiv preprint arXiv:0901.2698*, 2009.

Méta-modèles – Analyse de sensibilité

Heterogeneous Treatment Effects Estimation: When Machine Learning meets multiple treatments regime

Naoufal Acharki*[1], Josselin Garnier[2], Antoine Bertoncello , and Lugo Ramiro

[1]Centre de Mathématiques Appliquées - Ecole Polytechnique – Ecole Polytechnique, Centre National de la Recherche Scientifique : UMR7641 – France
[2]École Polytechnique – CMAP, École Polytechnique – France

Résumé

Résumé. Dans beaucoup de domaines scientifiques et techniques, l'inférence de l'effet d'un traitement et l'exploration de son hétérogénéité sont déterminantes pour l'optimisation et la prise de décision. De nombreux méta-algorithmes ont été développés pour estimer la fonction d'effet moyen conditionnel du traitement (CATE) pour un traitement binaire, leur principal avantage est de ne pas restreindre l'estimation à une méthode d'apprentissage supervisée particulière. Dans ce travail, nous étudions le régime de traitement multiple sous le modèle causal de Rubin et nous nous focalisons sur l'estimation des effets hétérogènes de traitement. Nous généralisons les meta-algorithmes pour l'estimation de la fonction CATE pour chaque niveau possible de traitement. Nous évaluons la qualité de chaque méta-algorithme sur des données observationnelles en utilisant un jeu de données Semi-synthétique et nous soulignons en particulier les performances du X-learner.
Mots clés. Apprentissage Automatique; Inférence Causale; Traitement multiple; Effets hétérogènes.

Abstract. In many scientific and engineering domains, inferring the effect of treatment and exploring its heterogeneity is crucial for optimization and decision making. Several meta-algorithms have been developed to estimate the Conditional Average Treatment Effect (CATE) function in the binary setting, with the main advantage of not restraining the estimation to a specific supervised learning method. In this work, we investigate the multiple treatment regime under Rubin Causal Model and we focus on estimating heterogeneous treatment effects. We generalize Meta-learning algorithms to estimate the CATE for each treatment value. Using semi-synthetic simulation datasets, we assess the quality of each meta-learner in observational data and we highlight in particular the performances of the X-learner
Keywords. Machine Learning; Causal Inference; Multiple Treatments; Heterogeneous Effects.

*Intervenant

Risk bounds for Sobol'-based Global Sensitivity Analysis Indices when using metamodels.

Clément Bénesse [1]

[1] *Institut de Mathématiques de Toulouse; clement.benesse@math.univ-toulouse.fr*

Résumé. Les métamodèles sont de plus en plus utilisé comme un moyen d'obtenir à moindre coût une approximation d'algorithmes complexes et gourmands, dans le but d'en extraire le maximum d'informations sur les caractéristiques du système. Parmi celles-ci, les indices que l'on retrouve en Analyse de Sensibilité Globale (GSA) permettent de connaître l'influence des variables d'entrée sur la sortie d'une boîte noire. Nous nous intéressons ici aux déviations d'une classe de ces indices lors de l'utilisation d'un métamodèle afin de pouvoir les borner.

Mots-clés. Métamodèles, Analyse de Sensibilité Globale, Indices de Sobol', Indices de Cramér-von-Mises, Indices de Shapley, Risque d'approximation.

Abstract. Surrogate models or meta-models recently gained traction as a mean to obtain a low-resource approximation of expensive computer codes in order to obtain critical information. Among these characteristics of the system, Global Sensitivity Analysis (GSA) indices are often computed to know which variables are the main drivers behind the computer code output. It is therefore necessary to know if the indices of the surrogate model are close to the indices of the true model. We provide upper bounds and rates of convergence on the error for various GSA indices in a framework where inputs are non-independent.

Keywords. Metamodels, Global Sensitivity Analysis, Sobol' indices, Cramér-von-Mises indices, Shapley indices, Approximation risk.

1 Introduction

L'utilisation d'algorithmes est désormais la routine dans la recherche scientifique. Les progrès récents ont vu l'apparition de méthodes de plus en plus sophistiquées et offrant des résultats de plus en plus précis, au prix parfois de l'explicabilité ou encore d'un coût computationnel accru. L'Analyse de Sensibilité Globale est un domaine s'intéressant à la quantification de l'influence de différentes variables d'entrée (également appelées «inputs») sur la sortie de ces algorithmes (ou «output») lorsque les inputs sont supposés aléatoire et l'algorithme est une boîte noire. Cette quantification se fait par le biais d'un certain nombre d'indices ayant diverses propriétés. Néanmoins, il est parfois impossible de calculer ces indices directement sur l'algorithme et l'utilisateur peut n'avoir accès

qu'à un métamodèle, c'est-à-dire une approximation du système complet pour laquelle les éventuelles imprécisions sont compensées par une facilité d'accès. Étant donné cette situation, une question légitime est la suivante: «Étant donné la précision de mon métamodèle $\hat{f}$ par rapport à l'algorithme réel mais inaccessible f, à quel point les indices de GSA calculés sur $\hat{f}$ sont proches des indices de GSA calculés sur f?»

Nous fournissons une réponse par le biais d'une borne supérieure au risque associé à ces indices, borne dont le principal tenant est le risque quadratique d'approximation $\|f-\hat{f}\|_2$. Une réponse partielle avait été proposée par Panin dans le cas des indices de Sobol' lorsque les inputs sont indépendants. Notre réponse complète ce résultat en supprimant l'hypothèse d'indépendance et en l'étendant à une classe plus générale d'indices que nous appelons les «Sobol'-based GSA indices», incluant les indices de Cramér-von-Mises et les indices de Shapley, fréquemment utilisés dans la littérature.

2 Définitions

Soit $\mathbf{X} = (X_1, \cdots, X_p)$ la variable aléatoire continue représentant les inputs, de loi $\mathbb{P}_{\mathbf{X}}$. Soit f un algorithme de carré intégrable, c'est-à-dire une fonction de $\mathbb{L}^2(\mathbb{P}_{\mathbf{X}})$. La notation $\sim i$ désigne $\{1, \ldots, p\} \setminus \{i\}$.

Définition 2.1. *1. On définit l'indice de Sobol' associé à la variable d'entrée X_i comme*

$$Sob_i(f) = \frac{V_i}{V} = \frac{Var(\mathbb{E}[f(\mathbf{X})|X_i])}{Var(f(\mathbf{X}))}. \tag{1}$$

2. On définit l'indice de Cramér-von-Mises associé à la variable d'entrée X_i comme

$$CvM_i(f) = \int Sob_i(\mathbb{1}_{f(\mathbf{X}\leq t)}) \times \frac{Var(\mathbb{1}_{f(X)\leq t})}{\int Var(\mathbb{1}_{f(X)\leq t})dt} dt. \tag{2}$$

3. On définit l'indice de Shapley associé à la variable d'entrée X_i comme

$$Sh_i(f) = \frac{1}{p} \sum_{A \subset \sim i} \binom{|A|}{p-1}^{-1} \left(S^{clos}_{A\cup\{j\}} - S^{clos}_{A} \right). \tag{3}$$

Les indices définis ci-dessus quantifie la part d'influence sur la sortie $f(X)$ que l'on peut imputer à X_i, de manière seul (Sobol', CvM) ou jointe avec d'autres variables (Shapley). Nous recommandons au lecteur intéressé la monographie par Da Veiga et al. pour plus de détail sur les différences entre ces indices, ainsi que l'article par Bénesse et al. pour un exemple d'application au problème d'équité algorithmique. En réalité, les indices définis ci-dessus sont un exemple d'indices «étendus», introduit par Mara et Tarantola, prenant en compte différents type d'influence et permettant une analyse poussée de l'importance d'un input sur l'output mais, par souci de simplicité, pour chacune de ses classes, seulement un représentant est exhibé.

Table 1: Bornes supérieures pour les indices de GSA avec un métamodèle déterministe.

Indice de GSA GSA	Borne supérieure associée
Indices de Sobol' (étendus)	$\frac{\|f-\hat{f}\|_{2,\mathbf{X}}}{\mathrm{Var}^{1/2}(f)}$
Indices de Cramér-von-Mises (étendus)	$\|f-\hat{f}\|_{2,\mathbf{X}}^{1/2}$
Indices de Shapley	$2 \times \frac{\|f-\hat{f}\|_{2,\mathbf{X}}}{\mathrm{Var}^{1/2}(f)}$

Table 2: Bornes sur le risque encouru pour les différents indices de GSA

Indices de GSA	Borne supérieure associée
Indices de Sobol' (étendus)	$\frac{\mathbb{E}\|f-\hat{f}\|^2}{\mathrm{Var}(f)}$
Indices de Cramér-von-Mises (étendus)	$\mathbb{E}\|f-\hat{f}\|$
Indices de Shapley	$2 \times \frac{\mathbb{E}\|f-\hat{f}\|^2}{\mathrm{Var}(f)}$

3 Principaux résultats

Dans cette section, nous présentons les principaux résultats. Nous nous intéressons à la quantité $|GSA_i(f) - GSA_i(\hat{f})$ lorsque $GSA_i(.)$ est l'un des indices définis dans la section précédente. Nous récapitulons les bornes associés à chacune des familles dans la table 1 lorsque $\hat{f}$ est déterministe et 2 lorsque $\hat{f}$ est issu d'une stratégie d'optimisation basée sur des données (par exemple, minimisation de la perte quadratique).

Théorème 3.1. *Soit f et $\hat{f}$ deux fonctions de $\mathbb{L}^2(\mathbb{P}_{\mathbf{X}})$. Soit X_i une des variables d'entrée et soit $GSA_i(.)$ un des indices de GSA défini lors de la définition 2.1. Alors*

$$|GSA_i(f) - GSA_i(\hat{f})| \leq B_i(f), \tag{4}$$

$$\mathbb{E}(GSA(f) - GSA(\hat{f}))^2 \leq B_i'(f), \tag{5}$$

avec $B_i(f)$ explicitée dans la table 1 et $B_i'(f)$ dans la table 2.

Bibliographie

Da Veiga, S. et Gamboa, F. et Iooss, B. et Prieur, C.,*Basics and Trends in Sensitivity Analysis: Theory and Practice in R*

Panin, I. , *Risk of estimators for Sobol'sensitivity indices based on metamodels*

Bénesse, C. et Gamboa, F. et Loubes, J-M. et Boissin, T., *Fairness seen as Global Sensitivity Analysis*

Test d'indépendance basé sur les indices HSIC-ANOVA d'ordre total

Gabriel Sarazin[1], Amandine Marrel[1-2], Sébastien Da Veiga[3] & Vincent Chabridon[4]

1. *CEA, DES, IRESNE, DER, Cadarache, 13108, Saint-Paul-Lez-Durance.*
2. *Institut de Mathématiques de Toulouse, 118 Route de Narbonne, 31062, Toulouse.*
3. *Safran Tech, Modeling & Simulation, Rue des Jeunes Bois, Châteaufort, 78114, Magny-Les-Hameaux.*
4. *EDF R&D, 6 Quai Watier, 78401, Chatou.*

Résumé. L'apprentissage statistique dans le cas de données simulées par un code de calcul industriel, aussi appelé "métamodélisation", est une tâche dont la difficulté de mise en œuvre croît avec la dimension du problème et le manque de données d'apprentissage. Une analyse de sensibilité préliminaire peut venir en soutien de la construction du métamodèle pour éliminer les variables les moins pertinentes et trier les variables restantes par ordre d'influence sur la sortie. Pour mener une analyse de sensibilité, l'approche historique de Sobol' offre un cadre conceptuel confortable qui est articulé autour de la décomposition de la variance de la sortie. Toutefois, l'estimation précise des indices associés n'est plus possible si l'échantillon d'apprentissage est de petite taille. Pour contourner cette difficulté, il est désormais fréquent d'utiliser une mesure de sensibilité basée sur le critère d'indépendance de Hilbert-Schmidt (notée HSIC). Elle est appliquée à chaque couple entrée-sortie, et permet ainsi de définir la collection des indices HSIC. Leur interprétation est généralement moins intuitive que celle des indices de Sobol car leur construction repose sur la théorie des espaces de Hilbert à noyaux reproduisants. Face à ce constat, les indices HSIC-ANOVA ont été récemment introduits et permettent une séparation des effets principaux et des interactions, à l'instar de la décomposition de Hoeffding dans le formalisme des indices de Sobol'. Cette avancée a été obtenue au prix d'une hypothèse d'indépendance mutuelle des entrées et sous réserve de l'utilisation de noyaux spécifiques, comme les noyaux de Sobolev. Dans ce travail, on commence par montrer que tout noyau de Sobolev est caractéristique, c'est-à-dire que la nullité d'un indice HSIC-ANOVA est équivalente à une situation d'indépendance au sein du couple formé par l'entrée considérée et la sortie. Dans un second temps, il est montré qu'un test d'indépendance peut être construit pour l'indice HSIC-ANOVA d'ordre total en s'inspirant de ce qui est fait pour l'indice HSIC traditionnel. Enfin, une étude numérique révèle empiriquement que le nouveau test d'indépendance est au moins aussi puissant que celui basé sur l'indice HSIC traditionnel, ce qui offre des perspectives intéressantes pour améliorer le processus de sélection des variables à impliquer dans la construction d'un métamodèle.

Mots-clés. Analyse de sensibilité, criblage, critère d'indépendance de Hilbert-Schmidt, test d'indépendance, analyse de la variance.

E-mails : gabriel.sarazin@cea.fr, amandine.marrel@cea.fr, sebastien.da-veiga@safrangroup.com, vincent.chabridon@edf.fr

Abstract. Building a surrogate model for an industrial computationally-expensive simulation code is made difficult by the combined effect of the curse of dimensionality and the lack of input-output data. A preliminary sensitivity analysis may help discard non-influential inputs and rank the remaining inputs according to their impact on the output distribution. In order to perform sensitivity analysis, the historical approach proposed by Sobol' provides a convenient conceptual framework where the output variance is apportioned between input variables. However, the accurate estimation of the corresponding indices requires thousands of model evaluations, which is often unaffordable in an industrial context. To circumvent this pitfall, it has become quite common to resort to a sensitivity measure based on Hilbert-Schmidt independence criterion (denoted by HSIC). This measure is applied to all input-output pairs of variables and allows to define the so-called "HSIC indices". Their interpretation is much less intuitive than the one related to Sobol' indices since their foundations come from the theory of reproducing kernel Hilbert spaces. To ease interpretation, the HSIC-ANOVA indices have been recently introduced to allow for a strict separation of main effects and interactions, akin to what is proposed in Sobol' formalism with Hoeffding decomposition. This breakthrough was obtained after assuming mutual independence between inputs and provided that specific kernels, like Sobolev kernels, are used to compute HSIC-ANOVA indices. In this work, a first contribution consists in demonstrating that Sobolev kernels are characteristic. Because of this property, independence within input-output pairs of variables can be detected from the observed values of HSIC-ANOVA indices. Then, it is shown that a test of independence can be constructed for the total-order HSIC-ANOVA index after adapting existing methodologies in the HSIC-related literature. Finally, an extensive simulation study proves empirically that the newly-developed test of independence is, at least, as powerful as the older one based on the traditional HSIC index, which offers interesting prospects in order to improve the screening step performed before metamodeling.
Keywords. Sensitivity analysis, screening, Hilbert-Schmidt independence criterion, test of independence, analysis of variance.

1 Généralités sur l'analyse de sensibilité

1.1 Propagation d'incertitudes en milieu industriel

L'étude des performances d'un système complexe (comme une centrale nucléaire, un aéronef, une plate-forme pétrolière, un ouvrage d'art, etc.) requiert l'utilisation d'un code de simulation qui modélise les phénomènes physiques mis en jeu et reproduit les contraintes exercées sur le système. Ce code peut être vu comme une fonction déterministe $g : \mathbb{R}^d \longrightarrow \mathbb{R}$ de type "boîte noire" dont les entrées $\boldsymbol{x} := [x_1, \ldots, x_d]$ sont des variables physiques incertaines et dont la sortie y est une variable d'intérêt qui résume l'état du système (valeur maximale d'une température, d'une pression, d'une déformation, etc.) et guide une éventuelle prise de décision. Dans un cadre probabiliste, les entrées du

code sont décrites par un vecteur aléatoire $\boldsymbol{X}$ de loi conjointe $\mathbb{P}_{\boldsymbol{X}}$ et de lois marginales $\mathbb{P}_{X_1}, \ldots, \mathbb{P}_{X_d}$. La propagation des incertitudes à travers le code g conduit à l'obtention d'une loi de probabilité $\mathbb{P}_Y$ en sortie. Comme chaque appel au code de calcul est très coûteux, la connaissance du fonctionnement interne de g se limite à un échantillon de simulations $\boldsymbol{Z}_{\text{obs}} = \left\{(\boldsymbol{X}^{(k)}, Y^{(k)})\right\}_{k=1}^{n}$ appariant un petit nombre n d'entrées et de sorties de la boîte noire ($n \approx 10\,d$). L'apprentissage statistique vise à construire un modèle de substitution (ou métamodèle) $\hat{g}$ en s'appuyant sur la seule connaissance de l'échantillon $\boldsymbol{Z}_{\text{obs}}$. Lorsque la dimension augmente (plusieurs dizaines, voire une centaine d'entrées), les techniques usuelles de métamodélisation ne peuvent plus être appliquées car elles impliquent l'optimisation d'un trop grand nombre de paramètres et d'hyperparamètres du métamodèle $\hat{g}$. Face à ce problème, mener une analyse de sensibilité préliminaire permet d'écarter les entrées dont l'influence sur la sortie est négligeable (phase de **criblage**) et de disposer d'un classement des entrées restantes (phase de **hiérarchisation**) qui peut être exploité a posteriori si la construction du métamodèle obéit à une heuristique séquentielle. Cette approche est notamment mise en œuvre dans **Marrel *et al.* (2021)**.

1.2 Émergence et utilisation grandissante des indices HSIC

Pour analyser la sensibilité de la sortie Y à des entrées $X_1, \ldots, X_d$ supposées indépendantes, l'approche fondatrice de Sobol s'appuie sur la décomposition de Hoeffding (ou "ANOVA fonctionnelle") de g afin de répartir la variance $\mathbb{V}(Y)$ entre les différentes entrées contributrices. L'indice de Sobol' du premier ordre S_i quantifie la part de variance expliquée uniquement à partir de l'incertitude portée par X_i. L'indice de Sobol' d'ordre total S_i^T s'obtient en agglomérant l'effet principal et les interactions avec les autres entrées :

$$\forall\, 1 \leq i \leq d,\; S_i := \frac{\mathbb{V}\big(\mathbb{E}[Y \mid X_i]\big)}{\mathbb{V}(Y)} \text{ et } S_i^T := 1 - \frac{\mathbb{V}\big(\mathbb{E}[Y \mid \boldsymbol{X}_{-i}]\big)}{\mathbb{V}(Y)} \text{ avec } \boldsymbol{X}_{-i} := \big[X_k\big]_{\substack{1 \leq k \leq d \\ k \neq i}}.$$

La grande popularité de cette méthodologie ANOVA s'explique principalement par l'interprétation intuitive qui en découle, les indices de Sobol' s'apparentant à des pourcentages de $\mathbb{V}(Y)$. Toutefois, l'utilisation pratique de ces indices n'est pas toujours possible car leur définition mathématique à partir de variances conditionnelles implique des schémas d'estimation trop gourmands en termes de nombre d'appels à g. Pour remédier à cela, les indices HSIC (*Hilbert-Schmidt Independence Criterion*) de **Gretton *et al.* (2005)** ont été repris par **Da Veiga (2015)** pour l'analyse de sensiblité des simulateurs numériques. Ces indices sont d'une tout autre nature puisqu'ils puisent leurs fondements mathématiques dans la théorie des espaces de Hilbert à noyaux reproduisants. L'indice $\text{HSIC}(X_i, Y)$ peut s'écrire comme une mesure de dissimilarité entre la loi conjointe $\mathbb{P}_{X_iY}$ décrivant la nature de la dépendance stochastique au sein de la paire (X_i, Y) et la loi $\mathbb{P}_{X_i} \otimes \mathbb{P}_Y$ obtenue en cas d'indépendance parfaite. Pour définir cette mesure de dissimilarité, il faut commencer par affecter des noyaux $K_1, \ldots, K_d$ et K_Y aux entrées et à la sortie. On rappelle que pour tout vecteur aléatoire $\boldsymbol{Z}$ à valeurs dans $\mathcal{Z}$ et de loi $\mathbb{P}_{\boldsymbol{Z}}$, tout noyau $K_{\boldsymbol{Z}} : \mathcal{Z} \times \mathcal{Z} \longmapsto \mathbb{R}$

est en relation directe avec un espace fonctionnel $\mathcal{H}_{\boldsymbol{Z}}$ appelé RKHS (*Reproducing Kernel Hilbert Space*) qui est susceptible d'offrir une certaine représentation de la loi $\mathbb{P}_{\boldsymbol{Z}}$. L'indice $\mathrm{HSIC}(X_i, Y)$ rend alors compte de la distance entre les représentants des lois $\mathbb{P}_{X_iY}$ et $\mathbb{P}_{X_i} \otimes \mathbb{P}_Y$ dans le RKHS tensorisé $\mathcal{H}_{X_i} \otimes \mathcal{H}_Y$. Il est montré dans **Gretton *et al.* (2007)** que la quantité $\mathrm{HSIC}(X_i, Y)$ fait uniquement intervenir les espérances de variables aléatoires définies à partir de deux copies indépendantes (X_i, Y) et (X_i', Y') de la loi $\mathbb{P}_{X_iY}$:

$$\begin{aligned}\mathrm{HSIC}(X_i, Y) \;=\; & \mathbb{E}_{\substack{X_iY \\ X_i'Y'}}\big[K_i(X_i, X_i')\, K_Y(Y, Y')\big] + \mathbb{E}_{X_iX_i'}\big[K_i(X_i, X_i')\big]\, \mathbb{E}_{YY'}\big[K_Y(Y, Y')\big] \\ & -2\, \mathbb{E}_{X_iY}\Big[\mathbb{E}_{X_i'}\big[K_i(X_i, X_i')\big]\mathbb{E}_{Y'}\big[K_Y(Y, Y')\big]\Big] \,.\end{aligned}$$

Un tel indice peut donc être estimé à l'aide de la V-statistique suivante :

$$\widehat{\mathrm{HSIC}}_i(\boldsymbol{Z}_{\mathrm{obs}}) := \frac{1}{n^2}\mathrm{Tr}\big(\boldsymbol{L}_i \boldsymbol{H} \boldsymbol{L}_Y \boldsymbol{H}\big) \quad \text{avec} \quad \boldsymbol{L}_i := \Big[K_i\big(X_i^{(s)}, X_i^{(t)}\big)\Big]_{1\leq s,t\leq n} \,,$$

$$\boldsymbol{L}_Y := \big[K_Y\big(Y^{(s)}, Y^{(t)}\big)\big]_{1\leq s,t\leq n} \qquad \text{et} \qquad \boldsymbol{H} := [\delta_{st} - 1/n]_{1\leq s,t\leq n} \,.$$

En pratique, on obtient une précision suffisante dès lors que $n \geq 100$. L'indice $\mathrm{HSIC}(X_i, Y)$ correspond en fait à la norme de Hilbert-Schmidt d'un hyper-opérateur de covariance qui amalgame tout le catalogue des transformations constitutives (ou *feature functions*) des RKHS $\mathcal{H}_{X_i}$ et $\mathcal{H}_Y$. Ainsi, lorsque les noyaux $K_1, ..., K_d$ et K_Y sont caractéristiques, le spectre des motifs de dépendance détectables devient suffisamment large pour que la nullité de $\mathrm{HSIC}(X_i, Y)$ traduise l'indépendance parfaite de X_i et Y.

Comme l'estimation de $\mathrm{HSIC}(X_i, Y)$ à partir de $\boldsymbol{Z}_{\mathrm{obs}}$ ne donne jamais zéro, et ce quelle que soit la nature de l'estimateur employé, on a besoin d'une procédure statistique pour tester l'hypothèse (H_0) : $\mathrm{HSIC}(X_i, Y) = 0$. La loi asymptotique de l'estimateur $\widehat{\mathrm{HSIC}}_i(\boldsymbol{Z}_{\mathrm{obs}})$ sous (H_0) étant connue, il est possible de s'appuyer sur cette statistique pour construire un test d'indépendance dont la conclusion doit aiguiller le choix de sélectionner ou d'évincer X_i lors du criblage des variables d'entrée. Toutefois, comme il est peu raisonnable en pratique de se considérer en régime asymptotique, un test alternatif décrit dans **De Lozzo et Marrel (2016)** permet de tester (H_0) quel que soit n en effectuant une séquence de permutations des observations $Y^{(k)}$ de la sortie.

Bien que les indices HSIC apportent une solution à de nombreux écueils (entrées dépendantes, entrées fonctionnelles, faible budget d'observations), ils restent difficiles à appréhender pour les non-spécialistes car ils ne s'inscrivent pas dans un paradigme de type ANOVA. Pour améliorer leur interprétabilité, **Da Veiga (2021)** a démontré qu'une décomposition de la quantité $\mathrm{HSIC}(\boldsymbol{X}, Y)$ peut être mise en place, à condition que les entrées soient indépendantes et moyennant l'utilisation de noyaux spécifiques en entrée :

$$\mathrm{HSIC}(\boldsymbol{X}, Y) = \sum_{\boldsymbol{u}\subseteq\{1,\dots,d\}} \mathrm{HSIC}_{\boldsymbol{u}} = \sum_{\boldsymbol{u}\subseteq\{1,\dots,d\}} \sum_{\boldsymbol{v}\subseteq\boldsymbol{u}} (-1)^{|\boldsymbol{u}|-|\boldsymbol{v}|}\, \mathrm{HSIC}(\boldsymbol{X}_{\boldsymbol{v}}, Y) \,.$$

Lorsque les variables X_i sont toutes de loi uniforme sur $[0\,,1]$, les noyaux de Sobolev vérifient toutes les conditions requises. Ces noyaux sont définis de la manière suivante :

$$\forall x, x' \in [0\,,1],\ K^r_{\mathrm{Sob}}(x,x') = 1 + k^r_{\mathrm{Sob}}(x,x') = 1 + \sum_{k=1}^{r} \frac{B_k(x)B_k(x')}{(k!)^2} + \frac{(-1)^{r+1}}{(2r)!} B_{2r}\big(|x-x'|\big)$$

avec $(B_k)_{k\geq 1}$ les polynômes de Bernoulli et r un paramètre de lissage rendant compte de la régularité des fonctions dans le RKHS sous-jacent. Sur le modèle des indices de Sobol', on peut alors définir des indices HSIC-ANOVA du premier ordre et d'ordre total :

$$\forall\, 1 \leq i \leq d,\ S_i^{\mathrm{HSIC}} := \frac{\mathrm{HSIC}(X_i, Y)}{\mathrm{HSIC}(\boldsymbol{X}, Y)} \text{ et } T_i^{\mathrm{HSIC}} := 1 - \frac{\mathrm{HSIC}(\boldsymbol{X}_{-i}, Y)}{\mathrm{HSIC}(\boldsymbol{X}, Y)} = \frac{\tau_i^{\mathrm{HSIC}}}{\mathrm{HSIC}(\boldsymbol{X}, Y)}\ .$$

Dans ce contexte, l'objectif de ce travail est d'examiner le bienfondé théorique et la faisabilité algorithmique d'un test d'indépendance basé sur la nullité de l'indice τ_i^{HSIC}.

2 Description de la contribution

On commence par démontrer que les noyaux de Sobolev sont caractéristiques. Il en découle un résultat intéressant qui marque la singularité de l'approche HSIC-ANOVA :

$$X_i \perp Y \iff S_i^{\mathrm{HSIC}} = 0 \iff T_i^{\mathrm{HSIC}} = 0 \iff \tau_i^{\mathrm{HSIC}} = 0\ .$$

La première équivalence, bien que très naturelle du point de vue des indices HSIC, est fausse si on la transpose au cas des indices de Sobol'. Comme les indices HSIC-ANOVA basés sur les noyaux de Sobolev sont capables de caractériser l'indépendance au sein d'une paire de variables, on propose de les utiliser pour tester $(H_0) : X_i \perp Y$. Pour cela, on considère l'estimateur V-statistique du numérateur τ_i^{HSIC} de l'indice T_i^{HSIC} :

$$\widehat{\tau}_i(\boldsymbol{Z}_{\mathrm{obs}}) := \frac{1}{n^2}\mathrm{Tr}\left[\Big(\big(\bigodot_{j\neq i} \boldsymbol{L}_j\big) \odot \boldsymbol{l}_i\Big)\boldsymbol{H}\boldsymbol{L}_Y\boldsymbol{H}\right] \quad \text{avec} \quad \begin{cases} \boldsymbol{L}_j := \Big[K^r_{\mathrm{Sob}}\big(X_j^{(s)}, X_j^{(t)}\big)\Big]_{1\leq s,t\leq n}\ , \\ \boldsymbol{l}_i := \Big[k^r_{\mathrm{Sob}}\big(X_i^{(s)}, X_i^{(t)}\big)\Big]_{1\leq s,t\leq n}\ , \end{cases}$$

où $\odot$ est le produit de Hadamard. Pour simuler la loi de $\widehat{\tau}_i(\boldsymbol{Z}_{\mathrm{obs}})$ sous (H_0), on construit une procédure de test articulée autour d'un schéma de permutation. Il s'agit plus précisément d'effectuer une séquence $\big(\sigma_j\big)_{1\leq j\leq B}$ de permutations aléatoires des observations de X_i au sein de $\boldsymbol{Z}_{\mathrm{obs}}$ puis de recalculer la statistique de test :

$$\widehat{\tau}_i^{[\sigma_j]} := \widehat{\tau}_i\left(\boldsymbol{Z}_{\mathrm{obs}}^{[\sigma_j]}\right) \quad \text{avec} \quad \boldsymbol{Z}_{\mathrm{obs}}^{[\sigma_j]} := \left\{\big(X_i^{(\sigma_j(k))}, \boldsymbol{X}_{-i}^{(k)}, Y^{(k)}\big)\right\}_{k=1}^{n}\ .$$

En reprenant les éléments de preuve de **Albert *et al.* (2022)**, et en notant $\hat{q}_{1-\alpha}$ le quantile empirique d'ordre $1-\alpha$ de la séquence $\big(\widehat{\tau}_i^{[\sigma_j]}\big)_{1\leq j\leq B+1}$ où $\sigma_{B+1} = \mathrm{Id}$ conduit à $\widehat{\tau}_i(\boldsymbol{Z}_{\mathrm{obs}})$, il

est facile de montrer que $\mathbb{P}_{H_0}\big(\hat{q}_{1-\alpha} \leq \alpha\big)$ et donc que le test est de bon niveau. En termes d'implémentation, la complexité de la procédure de test peut être limitée en adoptant la stratégie de **El Amri et Marrel (2021)** visant à minimiser le nombre B de permutations au regard de l'atteinte d'un objectif (stabilisation du criblage ou de la hiérarchisation). De surcroît, une astuce algébrique permet d'écrire que $n^2\,\widehat{\tau}_i(\boldsymbol{Z}_{\text{obs}}) = \text{Tr}(\boldsymbol{AW})$ avec $\boldsymbol{A}$ et $\boldsymbol{W}$ des matrices carrées symétriques. On montre qu'il est possible de calculer analytiquement les deux premiers moments de la variable $\text{Tr}(\boldsymbol{AW})$ lorsque $\boldsymbol{W}$ subit des permutations aléatoires affectant simultanément ses lignes et ses colonnes et que la loi uniforme discrète modélise l'aléa sur l'ensemble $\mathbb{S}_n$ des permutations. La p-valeur du test peut ensuite être estimée en supposant que la loi de $n^2\,\widehat{\tau}_i(\boldsymbol{Z}_{\text{obs}})$, sous l'effet des permutations, est une loi Gamma de paramètres calculés à l'aide de la méthode des moments exacts. On peut ainsi se départir de toute permutation au moment de tester (H_0). Une étude par simulation des performances du test basé sur $\widehat{\tau}_i(\boldsymbol{Z}_{\text{obs}})$ est réalisée sur différents cas analytiques, dont certains ont été spécialement conçus et/ou sélectionnés pour leur propension à fortement activer les indices HSIC-ANOVA d'ordre total. Le test proposé s'avère au moins aussi puissant (quand il ne l'est pas beaucoup plus) que le test initial basé sur l'indice HSIC traditionnel avec utilisation de noyaux gaussiens.

Bibliographie

Albert, M., Laurent, B., Marrel, A. et Meynaoui, A. (2022), Adaptive test of independence based on HSIC measures, à paraître dans *The Annals of Statistics*.

Da Veiga, S. (2015), Global sensitivity analysis with dependence measures. *Journal of Statistical Computation and Simulation*, 85(7), 1283-1305.

Da Veiga, S. (2021), Kernel-based ANOVA decomposition and Shapley effects - Application to global sensitivity analysis, arXiv preprint arXiv:2101.05487.

De Lozzo, M. et Marrel, A. (2016), New improvements in the use of dependence measures for sensitivity analysis and screening, *Journal of Statistical Computation and Simulation*, 86(15), 3038-3058.

El Amri, M. R. et Marrel, A. (2021), Optimized HSIC-based tests for sensitivity analysis - Application to thermalhydraulic simulation of accidental scenario on nuclear reactor, *Quality and Reliability Engineering International 2021*, 1-18.

Gretton, A., Bousquet, O., Smola, A. et Schölkopf, B. (2005), Measuring statistical dependence with Hilbert-Schmidt norms. *International conference on algorithmic learning theory*, 63-77, Springer, Berlin, Heidelberg.

Gretton, A., Fukumizu, K., Teo, C., Song, L., Schölkopf, B. et Smola, A. (2007), A kernel statistical test of independence, *Advances in neural information processing systems*, 20.

Marrel, A., Iooss, B. et Chabridon, V. (2021), The ICSCREAM Methodology: Identification of Penalizing Configurations in Computer Experiments Using Screening and Metamodel - Applications in Thermal-Hydraulics, *Nuclear Science and Engineering*, 1-21.

Remerciements

Ce travail de recherche est mené dans le cadre du projet **SAMOURAI** (***S**imulation **A**nalytics and **M**eta-model-based solutions for **O**ptimization, **U**ncertainty and **R**eliability **A**nalys**I**s*) financé par l'Agence Nationale de la Recherche (ANR-20-CE46-0013). Les auteurs sont très reconnaissants de ce soutien financier.

Sélection de variables lors de la généralisation d'un effet causal

Bénédicte Colnet [1,2], Julie Josse [3], Erwan Scornet[3], & Gaël Varoquaux[2]

[1] *Équipe projet SoDa, INRIA Saclay, Palaiseau. (email: benedicte.colnet@inria.fr)*
[2] *Centre de Mathémathiques Appliquées, Institut Polytechnique de Paris, Palaiseau.*
[3] *Inria Sophia-Antipolis, Montpellier.*

Résumé. Les essais contrôlés randomisés (ECR) sont considérés comme l'étalon-or pour conclure à l'effet causal d'une intervention donnée. Pour autant, ils présentent des limites quand la population éligible à l'essai est significativement différente de la population cible. Disposer d'un échantillon de la population cible d'intérêt permet de généraliser l'effet causal. Ce processus nécessite des covariables dans les deux jeux de données, en particulier toute variable qui serait un modulateur de l'effet du traitement ou dont la distribution change d'une population à l'autre. Dans ce travail, nous proposons d'étudier l'impact des covariables manquantes ou partiellement manquantes sur l'estimation, en quantifiant le biais sous une hypothèse semi-linéaire. Notre travail complète également les preuves de consistance pour les trois estimateurs usuels basés sur la repondération (IPSW), la modélisation de la réponse (G-formula), ou la combinaison des deux dans des approches dites doublement robustes (AIPSW). Ce résultat fournit une analyse de sensibilité pour les praticiens. Inversement, nous avons également étudié l'impact de l'ajout d'un trop grand nombre de covariables dans l'estimation, en particulier sur la variance asymptotique.
Mots-clés. Effet causal moyen, généralisation, analyse en sensibilité, validité externe, validité interne, consistance, variance asymptotique.

Abstract. Randomized Controlled Trials (RCTs) are considered as the gold standard to conclude on the causal effect of a given intervention, but they may lack of external validity when the population eligible to the RCT is substantially different from the target population. Having at hand a sample of the target population of interest allows to generalize the causal effect. This process needs covariates in both sets to capture all treatment effect modifiers that are shifted between the two sets. In this work, we propose to investigate the impact of missing or partially missing covariate on the estimation, quantifying the bias under a semi-linear assumption while completing consistency proofs for three tandard estimators relying on weighting (IPSW), outcome modeling (G-formula) or combining the two in doubly robust approaches (AIPSW). This result allows to provide a sensitivity analysis for practitionners. Conversely, we also investigated the impact of adding too many covariates in the estimation, deriving explicit asymptotic variance result.
Keywords. Average treatment effect (ATE), generalization, sensitivity analysis, external validity, internal validity, consistency, variance asymptotique.

In this document, we present a glimpse of the theoretical results obtained when covariates are missing, focusing on the key theorem being the asymptotic bias. We then only detail the semi-synthetic simulation performed, to illustrate the sensitivity analysis than can be build from this bias. The complete work contains consistency proofs for three estimators, a detail of the sensitivity analysis methodology, an application in critical care, and explicit variance results focusing on the impact of adding too many covariates.

1 Theoretical results when missing a covariate

1.1 Context and notations

Notations are grounded on the potential outcome framework [6]. We model each observations i in the RCT or observational population as described by a random tuple $(X_i, Y_i(0), Y_i(1), A_i, S_i)$ drawn from a distribution $(X, Y(0), Y(1), A, S) \in \mathcal{X} \times \mathcal{Y}^2 \times \{0,1\}^2$, such that the observations are iid. For each observations, X_i is a p-dimensional vector of covariates, A_i denotes the binary treatment assignment (with $A_i = 0$ if no treatment and $A_i = 1$ if treated), $Y_i(a)$ is the continuous outcome had the subject been given treatment a (for $a \in \{0,1\}$), and S_i is a binary indicator for RCT eligibility (i.e., meet the RCT inclusion and exclusion criteria) and willingness to participate if being invited to the trial ($S_i = 1$ if eligible and $S_i = 0$ if not). Assuming *consistency of potential outcomes*, and no interaction between treated and non-treated subject (SUTVA assumption), we denote by $Y_i = A_i Y_i(1) + (1 - A_i) Y_i(0)$ the outcome realized under treatment assignment A_i.
We define the conditional average treatment effect (CATE):

$$\forall x \in \mathcal{X}\,, \quad \tau(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x]\,,$$

and the population average treatment effect (ATE):

$$\tau = \mathbb{E}[Y(1) - Y(0)] = \mathbb{E}[\tau(X)]\,.$$

We model the patients belonging to an RCT sample of size n and in an observational data sample of size m by $n + m$ independent random tuples: $\{X_i, Y_i(0), Y_i(1), A_i, S_i\}_{i=1}^{n+m}$, where the RCT samples are identically distributed according to $\mathcal{P}(X, Y(0), Y(1), A, S \mid S = 1)$, and the observational data samples are identically distributed according to $\mathcal{P}(X, Y(0), Y(1), A, S)$. We also denote $\mathcal{R} = \{1, \ldots, n\}$ the index set of units observed in the RCT study, and $\mathcal{O} = \{n+1, \ldots, n+m\}$ the index set of units observed in the observational study.
For each RCT sample $i \in \mathcal{R}$, we observe $(X_i, A_i, Y_i, S_i = 1)$, while for observational data $i \in \mathcal{O}$, we consider the setting where *we only observe the covariates* X_i, which is a common case in practice.
Typical data set is presented on Table 1. Because the samples in the RCT and observational data do not follow the same covariate distribution, the ATE τ is different from the

RCT (or sample[1]) average treatment effect τ_1 which can be expressed as:

$$\tau \neq \tau_1 = \mathbb{E}[Y(1) - Y(0) \mid S = 1]\,.$$

This difference is the core of the *lack of external validity* introduced in the beginning of the work, but formalized with a mathematic expression.
For the rest of the work we denote $\mu_a(x)$ the conditional mean outcome under treatment $a \in \{0, 1\}$ in the observational data: $\mu_a(x) = \mathbb{E}(a) \mid X = x]$. $\mu_{(a)}(\cdot)$ are also called responses surfaces and can be estimated by fitting them on the controls and treated individuals in the RCT respectively to obtain $\hat{\mu}_0(\cdot)$ and $\hat{\mu}_1(\cdot)$. We denote respectively by $e_1(x)$ the propensity score in the RCT population $e_1(x) = \mathbb{P}(A = 1 \mid X = x, S = 1)$. Note that, the function $e_1(x)$ is imposed by the trial characteristics and usually is a constant e_1, for example $e_1 = 0.5$ (other cases include stratified RCT trials).

		Covariates				
	Set	X_1	X_2	X_3	A	Y
1	$\mathcal{R}$	1.1	20	5.4	1	10.1
	$\mathcal{R}$	-6	45	8.3	0	8.4
n	$\mathcal{R}$	0	15	6.2	1	14.5
$n+1$	$\mathcal{O}$		...		...	...
	$\mathcal{O}$	-2	52	NA	NA	NA
	$\mathcal{O}$	-1	35	NA	NA	NA
$n+m$	$\mathcal{O}$	-2	22	NA	NA	NA

		Covariates				
	Set	X_1	X_2	X_3	A	Y
1	$\mathcal{R}$	1.1	20	NA	1	10.1
	$\mathcal{R}$	-6	45	NA	0	8.4
n	$\mathcal{R}$	0	15	NA	1	14.5
$n+1$	$\mathcal{O}$		...		...	...
	$\mathcal{O}$	-2	52	3.4	NA	NA
	$\mathcal{O}$	-1	35	3.1	NA	NA
$n+m$	$\mathcal{O}$	-2	22	5.7	NA	NA

Figure 1: **Typical structure of the data considered**, where a covariate - quantitative or categorical - would be available in the RCT, but not in the observational data set (left) or the reverse situation (right). In this specific example $obs = \{1, 2\}$ ($mis = \{3\}$), corresponds to common (resp. different) covariates in the two datasets.

1.2 Assumptions

We admit there exist[2] $\delta \in \mathbb{R}^p$, and $\sigma \in \mathbb{R}^+$ such that:

$$Y = g(X) + A\langle X, \delta\rangle + \varepsilon, \qquad \text{where } \varepsilon \sim \mathcal{N}\left(0, \sigma^2\right), \tag{1}$$

where $\tau(X) := \langle X, \delta\rangle$ and $\varepsilon_0 = \varepsilon_1 = \varepsilon$. Under this model, the Average Treatment Effect (ATE) takes the following form:

$$\tau = \int \mathbb{E}\left[Y(1) - Y(0) \mid X = x\right] d\mathbb{P}(x) = \int \delta^T x d\mathbb{P}(x) = \delta^T \mathbb{E}[X]$$

[1]Usually τ_1 is also called the Sample Average Treatment Effect (SATE), when τ is named the Population Average Treatment Effect (PATE) [9, 8, 4, 3].

[2]In the main presentation we will detail why this assumption is not so strong.

Only variables being both treatment effect modifier ($\delta_j \neq 0$) and subject to a distributional change between the RCT and the target population are necessary to generalize the ATE. If such a key covariate is missing, it leads to a biased estimation of the target population ATE. We want to express the bias of such an omitted variable on the transported ATE. But first, we have to specify a context in which a certain permanence of the relationship between X_{obs} and X_{mis} in the two data sets holds. Therefore, we introduce the *Transportability of covariates relationship* assumption.

Assumption 1 (Transportability of covariates relationship)**.** *The distribution of X is Gaussian, that is, $X \sim \mathcal{N}(\mu, \Sigma)$, and transportability of Σ is true, that is, $X \mid S = 1 \sim \mathcal{N}(\mu_{RCT}, \Sigma)$.*

The plausibility of assumption 1 can be partially-assessed through a statistical test on $\Sigma_{obs,obs}$ for example Box's M test [2], supported with vizualizations [5]

1.3 Key result

In the detailed work, the asymptotic bias is given for three main estimators. Here, to remain general we denote an estimator relying on the set of observed covariate $\hat{\tau}_{n,m,obs}$.

Theorem 1. *Assume that Model* (1) *holds, along with identifiability Assumptions, and Assumption 1. Let B be the following quantity:*

$$B = -\sum_{j \in mis} \delta_j \left(\mathbb{E}[X_j] - \mathbb{E}[X_j \mid S = 1] - \Sigma_{j,obs}\Sigma^{-1}_{obs,obs}(\mathbb{E}[X_{obs}] - \mathbb{E}[X_{obs} \mid S = 1]) \right),$$

where $\Sigma_{obs,obs}$ is the sub matrix of Σ corresponding to observed index rows and columns, and $\Sigma_{j,obs}$ is the row j with column corresponding to observed index of Σ. Then, the two following statements hold.
Granting usual consistency assumption on the estimator on the set of observed covariate $\hat{\tau}_{n,m,obs}$, the asymptotic bias is equal to B, that is

$$\lim_{n,m \to \infty} \mathbb{E}[\hat{\tau}_{n,m,obs}] - \tau = B.$$

2 Sensitivity analysis on a semi-synthetic simulation

The above theoretical bias can be used to translate expert judgments about the strength of the missing covariates, which corresponds to sensitivity analysis.
To illustrate the previous methods on data for which the data generation process in not under our control, we use the data from a randomized controlled trial, the Tennessee Student/Teacher Achievement Ratio (STAR) study. This RCT is a pioneering randomized study from the domain of education [1], started in 1985, and designed to estimate the

effects of smaller classes in primary school, on the results. This experiment showed a strong payoff to smaller classes. In addition, the effect has been shown to be heterogeneous, where class size has a larger effect for minority students and those on subsidized lunch. For our purposes, we focus on the same subgroup of children, same treatment (small versus regular classes), and same outcome (average of all grades at the end) as in [7]. 4 509 students are concerned by the treatment randomization, with treatment assignment at first grade only. On the whole data, we estimated an average treatment effect of 12.80 additional points on the grades (95% CI [10.41-15.2]) with the difference-in-means estimator (**??**). We consider this estimate as the ground truth τ as it is the global RCT. Then, we generate a random sample of 500 children to serve as the observational study. From the rest of the data, we sample a biased RCT according to a logistic regression that defines probability for each class to be selected in the RCT, and using only the variable `g1surban` informing on the neighborhood of the school. The final selection is performed using a Bernoulli procedure, which leads to 563 children in the RCT. The resulting RCT is such that τ_1 is 4.85 (95% CI [-2.07-11.78]) which is underestimated. This is due to the fact that that the selection is performed toward children that benefit less from the class size reduction according to previous studies. When generalizing the ATE with the G-formula on the full set of covariates, estimating the nuisance components with a linear model (resp. random forest), and estimating the confidence intervals with a stratified bootstrap (500 repetitions), the target population ATE is recovered with an estimate of 13.05 (95% CI [5.13-20.88]) (resp. 13.02 (95% CI [4.40-20.04])). When not including the covariate on which the selection is performed (`g1surban`) leads to a biased generalized ATE of 5.87 (95% CI [-2.34-13.06]) (resp. 6.41 (95% CI [-3.02-12.77])). These results are represented on Figure 2.

References

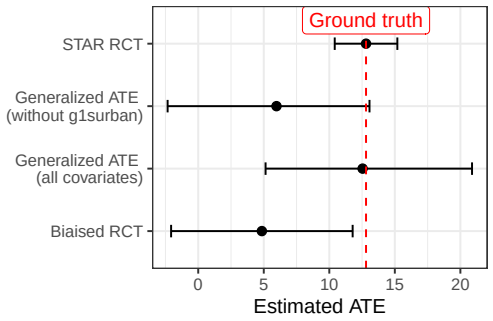
[1] Joshua D. Angrist and Jörn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, Dec. 2008. ISBN: 0691120358.

[2] G. E. P. Box. "A general distribution theory for a class of likelihood criteria". In: *Biometrika* 36.3-4 (Dec. 1949), pp. 317–346. ISSN: 0006-3444. DOI: 10.1093/biomet/36.3-4.317. eprint: https://academic.oup.com/biomet/article-pdf/36/3-4/317/530828/36-3-4-317.pdf. URL: https://doi.org/10.1093/biomet/36.3-4.317.

[3] Irina Degtiar and Sherri Rose. *A Review of Generalizability and Transportability*. 2021. arXiv: 2102.11904 [stat.ME].

[4] Naoki Egami and Erin Hartman. *Covariate Selection for Generalizing Experimental Results: Application to Large-Scale Development Program in Uganda*. 2021. arXiv: 1909.02669 [stat.ME].

Figure 2: **Simulated STAR data**: True target population ATE estimation using all the STAR's RCT data is represented, along with the ATE estimate of a biased simulated RCT (difference-in-means), and the generalized estimate of the ATE is called `Generalized ATE`, and is estimated with all covariates or without `g1surban` with the G-formula using a linear model for the nuisance components. In the case of the G-formula, the confidence intervals are estimated with a stratified bootstrap (500 repetitions). The true population ATE is the red dashed line. Similar results are obtained when nuisance components are estimated with random forest.



[5] Michael Friendly and Matthew Sigal. "Visualizing Tests for Equality of Covariance Matrices". In: *The American Statistician* 74.2 (2020), pp. 144–155. DOI: 10.1080/00031305.2018.1497537. eprint: https://doi.org/10.1080/00031305.2018.1497537. URL: https://doi.org/10.1080/00031305.2018.1497537.

[6] Guido W Imbens and Donald B Rubin. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge UK: Cambridge University Press, 2015.

[7] Nathan Kallus, Aahlad Manas Puli, and Uri Shalit. "Removing hidden confounding by experimental grounding". In: *Advances in neural information processing systems*. 2018, pp. 10888–10897.

[8] Luke W. Miratrix et al. *Worth Weighting? How to Think About and Use Weights in Survey Experiments*. 2017. arXiv: 1703.06808 [stat.ME].

[9] Elizabeth A Stuart et al. "The use of propensity scores to assess the generalizability of results from randomized trials". In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 174 (2011), pp. 369–386.

Calage conditionnel bayésien d'un modèle numérique

Oumar Baldé [1], Guillaume Damblin [1], Loïc Giraldi [2], Antoine Bouloré[2] et Amandine Marrel[2,3]

[1] *CEA, DES, Saclay,* 91191*, Gif-sur-Yvette*
[2] *CEA, DES, IRESNE, DEC, Cadarache,* 13108 *St Paul Lez Durance*
[3] *Institut de Mathématiques de Toulouse,* 118 *Route de Narbonne,* 31062*, Toulouse*

Résumé. L'étude de certains phénomènes physiques complexes requiert l'utilisation de modèles numériques. La précision et la robustesse de ces derniers peuvent être significativement affectées du fait des approximations simplifiées des réalités physiques. C'est pourquoi, la quantification et la réduction des erreurs et incertitudes des modèles sont des enjeux cruciaux pour les applications industrielles (nucléaires, etc.). Une source importante d'incertitude au sein d'un modèle est celle reliée à ses paramètres de modélisation $\theta \in \mathbb{R}^p\,(p \geq 1)$. L'estimation de ces paramètres (calage de modèle) peut significativement améliorer la précision du modèle. Notre travail porte plus précisément sur le calage bayésien conditionnel. Il s'agit de réaliser le calage d'un vecteur de paramètres d'entrée $\theta(\lambda) \in \mathbb{R}^p\,(p \geq 1)$ fonction d'une covariable incertaine ($\lambda \in \mathbb{R}^D$ avec $D \geq 1$), à partir des données expérimentales disponibles sous l'hypothèse que ces données observées ne dépendent pas de la valeur de λ. Pour cela, nous proposons un formalisme bayésien hiérarchique du problème, en supposant que le modèle est linéaire en $\theta(\lambda)$. L'approche est hiérarchique au sens où chaque composante de $\theta(\lambda)$ est supposée être la trajectoire indépendante d'un processus gaussien *a priori* dont les hyperparamètres sont estimés par maximisation de la vraisemblance marginale des données observées. À court terme, cette approche sera appliquée au calage des paramètres d'un modèle de gaz de fission en fonction de la conductivité ($\lambda \in \mathbb{R}$) du combustible nucléaire.

Abstract. The study of certain complex physical phenomena requires the use of numerical models. The accuracy and robustness of these models can be significantly affected by simplified approximations of the physical realities. This is why the quantification and reduction of model errors and uncertainties are crucial issues for many industrial applications (nuclears, etc.). An important source of uncertainty within a model is that related to its modeling parameters $\theta \in \mathbb{R}^p\,(p \geq 1)$. The estimation of these parameters (model calibration) can significantly improve the accuracy of the model. Our work focuses here conditionnal Bayesian calibration. This aims to calibrate the vector of input parameters $\theta(\lambda) \in \mathbb{R}^p\,(p \geq 1)$ as a function of an uncertain covariate ($\lambda \in \mathbb{R}^D$ with $D \geq 1$), from the available experimental data under the hypothesis these observed data do not depend on the value of λ. For this, we have proposed a hierarchical Bayesian formalism of the problem, assuming that the model is linear in $\theta(\lambda)$. This approach is hierarchical in the sense that each component of $\theta(\lambda)$ is assumed to be the independent trajectory of a Gaussian process prior whose hyperparameters are estimated by maximizing the marginal likelihood of the observed data. In the sort term, this approach will be applied to the calibration of the parameters of a fission gas model as a function of fuel's conductivity ($\lambda \in \mathbb{R}$).

Mots-clés. Calage bayésien conditionnel, approche hiérarchique, processus gaussien.
Keywords. Conditionnal Bayesian calibration, hierarchical approach, Gaussian process.

E-mails : oumar.balde@cea.fr guillaume.damblin@cea.fr loic.giraldi@cea.fr antoine.boulore@cea.fr amandine.marrel@cea.fr

1 Introduction

Les modèles numériques utilisés pour modéliser les systèmes physiques mis en jeu dans un contexte industriel sont entachés d'incertitudes. Par conséquent, avant toute utilisation d'un modèle, il est fondamental de quantifier toutes les sources d'erreurs et incertitudes pouvant affecter sa capacité à prédire le système physique étudié. Autrement dit, il doit subir une procédure de vérification et validation. La vérification a pour but de quantifier toutes les erreurs numériques. Ensuite, en confrontant les simulations numériques avec des données expérimentales, la validation doit déterminer si le modèle numérique est capable de prédire le système physique avec précision. Le calage de modèle, étape nécessaire à la validation, est un processus d'ajustement des paramètres incertains en fonction des données expérimentales disponibles. Dans ce contexte, notre travail porte sur le calage bayésien conditionnel. Il s'agit d'estimer un vecteur de paramètres $\theta(\lambda) \in \mathbb{R}^p$ fonction d'une covariable incertaine $\lambda \in \mathbb{R}^D$, d'un modèle $y_{\theta(\lambda)}(x)$ sous l'hypothèse de compensation physique selon laquelle les données expérimentales ne dépendent pas de la valeur de λ.

2 Calage bayésien conditionnel

Soit un système physique d'intérêt $r(x) \in \mathbb{R}$. Soit un modèle numérique $y_{\theta(\lambda)}(x)$ linéaire en $\theta(\lambda)$ vérifiant l'hypothèse de compensation et étant suffisamment représentatif de $r(x)$. Par conséquent, le vecteur des observations $z \in \mathbb{R}^n$ du système physique est relié au vecteur des simulations $y_{\theta(\lambda)} = g_\lambda \theta(\lambda)$ en fonction des n configurations $(x_i)_{i=1}^n$ par l'équation :

$$\begin{aligned} z &= g_\lambda \theta(\lambda) + \epsilon, \\ \epsilon &\sim \mathcal{N}(0, \Sigma_\epsilon), \end{aligned} \tag{1}$$

où $z = (z_1, \cdots, z_n)^t$, $g_\lambda = \left(g_\lambda^t(x_1), \cdots, g_\lambda^t(x_n)\right)^t$ et $\Sigma_\epsilon = \mathrm{diag}(\sigma_{\epsilon_1}^2, \cdots, \sigma_{\epsilon_n}^2)$ connue.

Nous disposons de m valeurs λ_j $(1 \leq j \leq m)$ qui sont associées aux paramètres

$$\Theta_m := (\theta(\lambda_1), \cdots, \theta(\lambda_m))^t \in \mathbb{R}^{p \times m}.$$

On ne disposera que d'une seule réalisation z pour estimer Θ_m. Posons donc $\mathbf{z} = z\mathbf{1}_m^t \in \mathbb{R}^{n \times m}$ la matrice des m copies de z et $D_m := (\lambda_1, \cdots, \lambda_m)^t$.

2.1 Proposition d'une approche jointe hiérarchique

L'objectif est double : estimer la loi *a posteriori* de Θ_m (i.e conditionnellement aux observations z) et calculer la distribution de probabilité *a posteriori* de $\theta(\lambda^\star)$ pour de nouvelles réalisations $\lambda^\star$. Cette distribution est appelée « prédictive ». Pour répondre au premier objectif, nous proposons dans ce qui suit une approche hiérarchique jointe.

Cette approche jointe est hiérarchique au sens où chaque composante de $\theta(\lambda) \in \mathbb{R}^p$ est supposée

être une trajectoire indépendante *a priori* d'un processus gaussien qui dépend lui-même de paramètres inconnus appelés hyperparamètres :

$$\theta_k(.) \sim \mathcal{GP}\big(m_{\beta_k}(.), \sigma_k^2 K_{\psi_k}(.,.)\big),\ 1 \leq k \leq p.$$

Pour estimer la loi *a posteriori* de $\Theta_m \in \mathbb{R}^{p\times m}$, on considère la distribution gaussienne multivariée sur sa forme vectorisée $\vec{\Theta}_m \in \mathbb{R}^{pm}$. Ainsi, la loi *a priori* sur Θ_m est la suivante :

$$\pi(\Theta_m|\phi) \propto |\mathbf{K}_\phi|^{\frac{-1}{2}} \exp -\frac{1}{2}(\vec{\Theta}_m - \vec{M}_\phi)^t \mathbf{K}_\phi^{-1}(\vec{\Theta}_m - \vec{M}_\phi)$$

où

$$\begin{aligned}
&\vec{M}_\phi = \big(m_{\beta_1}(D_m)^t, \cdots, m_{\beta_p}(D_m)^t\big)^t \in \mathbb{R}^{pm},\ \phi := (\beta_k,\ \sigma_k^2,\ \psi_k)_{k=1}^p \\
&(\mathbf{K}_\phi)_{jj'} = \mathrm{cov}(\theta(\lambda_j), \theta(\lambda_{j'})),\ 1 \leq j, j' \leq m, \\
&\mathrm{cov}(\theta(\lambda_j), \theta(\lambda_{j'})) = \mathrm{diag}\Big\{\sigma_l^2 K_{\psi_l}(\lambda_j,\ \lambda_{j'})\Big\}_{l=1}^p \in \mathbb{R}^{p\times p}.
\end{aligned}$$

En combinant la vraisemblance $\mathcal{L}(\mathbf{z}|\Theta_m)$ et la loi *a priori* $\pi(\Theta_m|\phi)$ grâce à la formule de Bayes, nous obtenons la loi *a posteriori* $\pi(\Theta_m|\mathbf{z}, \phi)$:

$$\pi(\Theta_m|\mathbf{z}, \phi) \propto \mathcal{L}(\mathbf{z}|\Theta_m)\pi(\Theta_m|\phi).$$

On montre facilement que cette loi est gaussienne de moyenne $\vec{\mu}_\phi$ et de matrice de covariance $\mathbf{\Sigma}_\phi$ données ci-dessous :

$$\boxed{\vec{\mu}_\phi = \mathbf{\Sigma}_\phi(\mathbf{K}_\phi^{-1}\vec{M}_\phi + G^t\Sigma_\epsilon^{-1}z)} \tag{2}$$

$$\boxed{\mathbf{\Sigma}_\phi = (\mathbf{\Delta}^{-1} + \mathbf{K}_\phi^{-1})^{-1}} \tag{3}$$

où

$$G = (g_{\lambda_1}, \cdots, g_{\lambda_m}) \in \mathbb{R}^{n\times pm},\ \mathbf{\Delta}^{-1} = \mathrm{Diag}(g_{\lambda_1}^t\Sigma_\epsilon^{-1}g_{\lambda_1}, \cdots, g_{\lambda_m}^t\Sigma_\epsilon^{-1}g_{\lambda_m}) \in \mathbb{R}^{pm\times pm}.$$

Les hyperparamètres ϕ sont estimés par maximisation de la vraisemblance marginale (approche appelée empirical Bayes) :

$$\hat{\phi} = \underset{\phi}{\mathrm{argmax}} \int \mathcal{L}(\mathbf{z}|\Theta_m)\pi(\Theta_m|\phi)d\Theta_m. \tag{4}$$

Maintenant, nous pouvons calculer la densité prédictive $\pi^{pred}(\theta(\lambda^\star)|\mathbf{z}, \phi)$ pour de nouvelles réalisations $\lambda^\star \in \mathbb{R}^{D\times k}$. Elle s'obtient en intégrant la distribution conditionnelle gaussienne $\pi(\theta(\lambda^\star)|\Theta_m, \phi)$ par rapport à la mesure de probabilité *a posteriori* :

$$\pi^{pred}(\theta(\lambda^\star)|\mathbf{z}, \phi) = \int \pi(\theta(\lambda^\star)|\Theta_m, \phi)\pi(\Theta_m|\mathbf{z}, \phi)d\Theta_m \sim \mathcal{N}_{pk}(\mu_\phi^{pred}(\lambda^\star),\ \Sigma_\phi^{pred}(\lambda^\star, \lambda^{\star\prime})). \tag{5}$$

Dans un exemple analytique, nous pouvons nous intéresser à la performance de la loi prédictive en la comparant à une loi dite loi cible pour de nouvelles réalisations $\lambda^\star \in \mathbb{R}^{D\times k}$. Cette loi, comme son nom l'indique, est la loi que nous aimerions reconstruire basée sur les $g_{\lambda^\star}$ qui seraient connus. Cette loi cible sera calculée en utilisant les deux ingrédients suivants : la loi *a priori* la plus objective sur $\theta(\lambda^\star) \in \mathbb{R}^{p\times k}$ qui sera l'*a priori* de Jeffreys et la vraisemblance conditionnelle $\mathcal{L}(z\mathbf{1}_k^t|\theta(\lambda^\star))$. La loi *a posteriori* $\pi(\theta(\lambda^\star)|z\mathbf{1}_k^t)$ obtenue par la formule de Bayes est gaussienne.

2.2 Exemple analytique

Pour comparer la loi prédictive à la loi cible, nous considérons le modèle analytique suivant :

$$g_\lambda(x) = \frac{r(x)}{\theta(\lambda)} \text{ et } \theta(\lambda) = \frac{\lambda \sin(\lambda)}{10} + 2.$$

Les mesures $z = (z_1, \cdots, z_n)^t \in \mathbb{R}^n$ sont simulées quelle que soit la réalisation de $\lambda \sim \mathcal{U}[0,\, 10]$ de la manière suivante :

$$z_i = g_\lambda(x_i)\theta(\lambda) + \epsilon_i,\ i = 1, \cdots, n,\ \epsilon_i \overset{iid}{\sim} \mathcal{N}(0, 4).$$

Le polynôme suivant :

$$r(x) = 2x^2 + 3x + 1,\ \forall\, x \in [-4,\, 4]$$

joue le rôle du système physique. Pour modéliser la loi *a priori* de Θ_m, on choisit un processus gaussien avec :

- une fonction de covariance de type Matérn 5/2,
- un polynôme de degré un pour la fonction moyenne : $m_\beta(\lambda) = \lambda\beta \in \mathbb{R}$.

À noter que le plan d'expériences D_m de taille m sur la covariable λ est constitué par échantillonnage de type hypercube latin. Une fois les hyperparamètres ϕ estimés (cf. Équation (4)), la distribution prédictive pour une nouvelle réalisation de $\lambda_0^\star$ est calculée via l'Équation (5) et comparée à la loi cible qui est la loi *a posteriori* associée à $\lambda_0^\star$.

Deux exemples pour des échantillons de $n = 50$ observations sont données par les Figures 1 et 2 respectivement pour des tailles $m = 10$ et $m = 15$ de plan d'expériences sur λ. Sur chaque figure, sont représentées sur le graphe de gauche la fonction théorique $\theta(\lambda)$ ainsi que les moyennes de la loi prédite et de la loi *a posteriori* cible obtenues, respectivement notées $\bar{\theta}_{pred}(\lambda)$ et $\bar{\theta}_{cible}(\lambda)$. Les intervalles de crédibilité à 95% associés à ces deux lois sont également représentés. Sur les graphes de droite, les deux distributions de probabilité obtenues pour un λ particulier (au point $\lambda_0^\star = 3.48$) sont comparées :

- la loi prédictive $\pi^{pred}(\theta(\lambda_0^\star)|\mathbf{z}, \hat{\phi})$ en orange,
- la loi cible $\pi^{cible}(\theta(\lambda_0^\star)|z) = \pi^{cible}(\theta(\lambda_0^\star)|z\mathbf{1}_1^t)$ en verte.

Nous constatons tout d'abord que les moyennes *a posteriori* de ces deux lois gaussiennes sont proches de $\theta(\lambda_0^\star) = 1.8853$. On obtient par exemple pour $m = 10$:

$$\bar{\theta}_{cible}(\lambda_0^\star) = 1.8793,\ \bar{\theta}_{pred}(\lambda_0^\star) = 1.8727.$$

Toujours pour $m = 10$, on note par ailleurs que l'intervalle de crédibilité à 95% de la prédictive est beaucoup plus large que celui de la cible. Cette variance beaucoup plus élevée pourrait s'expliquer par la petite taille du plan d'expériences D_m (ici $m = 10$). En effet, en augmentant la taille du plan (à $m = 15$), nous constatons que la variance prédictive est nettement réduite (Figure 2 , graphe de gauche). On constate un intervalle de crédibilité à 95% est légèrement plus étroit que celui de la cible pour $\lambda_0^\star = 3.48$. Ceci s'explique par le choix de l'*a priori* gaussien et par l'estimation des hyperparamètres par empirical Bayes. Cette estimation néglige l'incertitude *a priori* sur les hyperparamètres et ceci semble induire une incertitude *a posteriori* plus resserrée autour de la moyenne *a posteriori* $\bar{\theta}_{pred}(\lambda_0^\star)$.

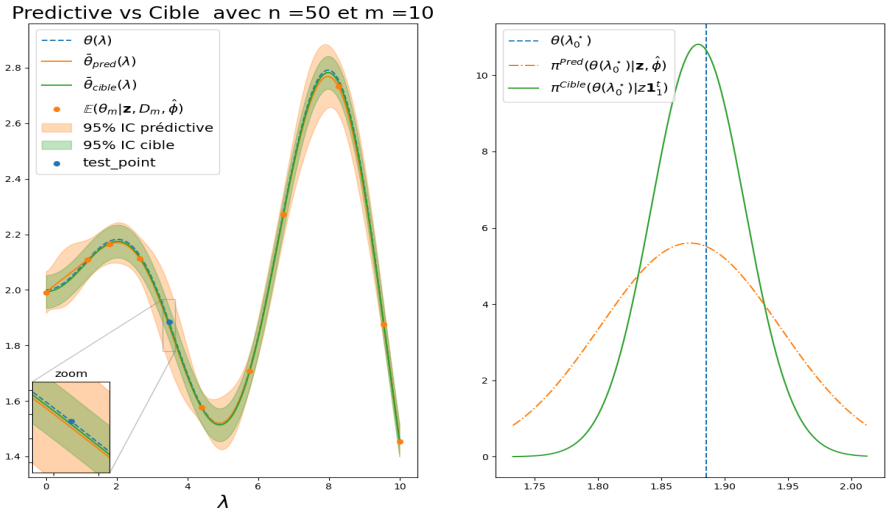


FIGURE 1 – Fonction théorique $\theta(\lambda)$, moyenne prédite notée $\bar{\theta}_{pred}(\lambda)$, moyenne *a posteriori* cible notée $\bar{\theta}_{cible}(\lambda)$ et intervalles de crédibilité à 95% associés, obtenus sur un échantillon de $n = 50$ observations et un plan d'expériences de taille $m = 10$ sur λ.

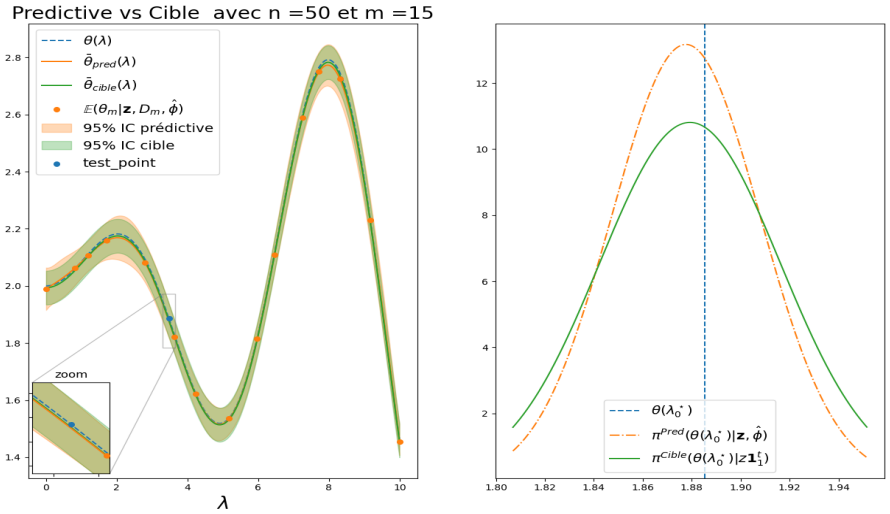


FIGURE 2 – Fonction théorique $\theta(\lambda)$, moyenne prédite notée $\bar{\theta}_{pred}(\lambda)$, moyenne *a posteriori* cible notée $\bar{\theta}_{cible}(\lambda)$ et intervalles de crédibilité à 95% associés, obtenus sur un échantillons de $n = 50$ et un plan d'expériences de taille $m = 15$ sur λ.

Nous avons testé l'approche jointe hiérarchique en $1D$ sur un exemple analytique et nous avons obtenu des résultants convaincants. Cette approche basée sur les processus gaussiens permet d'effectuer un calage joint des paramètres de modèle $\theta(\lambda_j)$ correspondant aux différents échantillons λ_j d'apprentissage. À partir de ce calage, nous déduisons une loi prédictive pour de nouvelles réalisations $\lambda^\star$. De cette loi prédictive, nous obtenons une fonction moyenne prédite $\bar{\theta}_{pred}$ qui prédit la fonction θ en ces nouvelles réalisations $\lambda^\star$ avec des intervalles de crédibilité associés. Notons que cette prédictive est obtenue sous l'hypothèse que les hyperparamètres du processus gaussien sont connus et fixés. Cependant, en pratique, les hyperparamètres sont inconnus et doivent être estimés. Pour ce faire, nous avons utilisé l'approche empirical Bayes. L'estimation par empirical Bayes néglige l'incertitude *a priori* sur les hyperparamètres. Ceci peut induire une incertitude *a posteriori* parfois trop étroite sur θ. Pour une meilleure prise en charge de l'incertitude *a priori* sur les hyperparamètres, nous pouvons soit quantifier cette incertitude par une loi *a priori* (loi hyperprior) soit combiner l'approche empirical Bayes à une approche bootstrap. Cette dernière permettra de réintroduire l'incertitude *a priori* sur les hyperparamètres négligée durant la phase d'estimation.

Bibliographie

Baratt, S. (2018), A matrix gaussian distribution, arXiv preprint : 1804.11010.

Campbell, K. (2006), Statistical calibration of computer simulations, Reliability Engineering & System Safety, $91(10-11) : 1358-1363$.

Damblin, G. (2015), Contributions statistiques au calage et à la validation des codes de calcul, PhD thesis, Paris, AgroPArisTech.

De Rocquigny, E., Devictor, N. et Tarantola, S. (2008), Uncertainty in industrial practice : a guide to quantitative uncertainty management, John Wiley & Sons.

Kennedy, M. et O'Hagan, A. (2001), Bayesian calibration, Journal of the Royal Statistical Society : Series B (Statistical Methodology), 63(3) :425–464.

Laird, M. N. et Louis, T. (1987), Empirical Bayes confidence intervals based on bootstrap samples, Journal of the American Statistical Association, 82(399) :739–750, 1987.

McKay, M. D., Bechman, R. J. et Conover, W. J. (1979), A comparison of three methods for selecting values of input variables in the analysis of output from computer code, Technometrics, 21 :239-245.

Santner, T. J., Williams, B. J. et Notz, W. I (2018), Physical experiments and computer experiments, In The design and analysis of computer experiments, Springer.

Surjanovic S. et Bingham, D. (2013), Virtual library of simulation experiments : Test functions and datasets, from `http://www.sfu.ca/~ssurjano`.

ML & Regret

Multitask Online Mirror Descent

Pierre Laforgue [1], Nicolò Cesa-Bianchi [2], Andrea Paudice [3] & Massimiliano Pontil [4]

[1] *University of Milan, pierre.laforgue@unimi.it*
[2] *University of Milan, nicolo.cesa-bianchi@unimi.it*
[3] *University of Milan and Istituto Italiano di Tecnologia, andrea.paudice@unimi.it*
[4] *Istituto Italiano di Tecnologia and University College London, massimiliano.pontil@iit.it*

Résumé. Nous introduisons et analysons MT-OMD, une extension multitâche d'*Online Mirror Descent* (OMD) basée sur l'échange d'informations entre les différentes tâches. Nous prouvons que le regret de MT-OMD est de l'ordre de $\sqrt{1+\sigma^2(N-1)}\sqrt{T}$, où σ^2 est la variance des tâches vis-à-vis de la géométrie induite par le régulariseur, N est le nombre de tâches, et T est l'horizon. Dès que les tâches sont similaires, c'est-à-dire quand $\sigma^2 \leq 1$, notre méthode présente un regret inférieur à $\sqrt{NT}$, obtenu en appliquant OMD indépendamment sur chaque tâche. Nous prouvons enfin que l'amélioration permise par MT-OMD est optimale, et appuyons nos résultats théoriques par une analyse empirique.

Mots-clés. Optimisation convexe en ligne, apprentissage multitâche

Abstract. We introduce and analyze MT-OMD, a multitask generalization of Online Mirror Descent (OMD) which operates by sharing updates between tasks. We prove that the regret of MT-OMD is of order $\sqrt{1+\sigma^2(N-1)}\sqrt{T}$, where σ^2 is the task variance according to the geometry induced by the regularizer, N is the number of tasks, and T is the time horizon. Whenever tasks are similar, that is $\sigma^2 \leq 1$, our method improves upon the $\sqrt{NT}$ bound obtained by running independent OMDs on each task. We further provide a matching lower bound, and present experiments supporting our theoretical findings.

Keywords. Online convex optimization, multitask learning

1 Introduction

In multitask learning (Caruana, 1997), one faces a set of tasks to solve, and tries to leverage their similarities to learn faster. Task similarity is often formalized in terms of Euclidean distances among the best performing models for each task. However, in online convex optimization, and Online Mirror Descent (OMD) in particular, it is well known that using different geometries to measure distances in the model space can bring substantial advantages — see, e.g., Orabona (2019). For instance, when the model space is the probability simplex in $\mathbb{R}^d$, running OMD with the KL divergence (corresponding to an entropic regularizer) allows one to learn at a rate depending only logarithmically on d. It is thus natural to investigate to what extent measuring task similarities using

geometries that are possibly non-Euclidean could improve the analysis of online multitask learning. From an application perspective, typical online multitask scenarios include federated learning applications for mobile users (e.g., personalized recommendation or health monitoring) or for smart homes (e.g., energy consumption prediction), mobile sensor networks for environmental monitoring. In this work, we introduce MT-OMD, a multitask generalization of OMD which applies to any strongly convex regularizer. We present a regret analysis establishing that MT-OMD outperforms OMD (run independently on each task) whenever tasks are similar according to the geometry induced by the regularizer.

2 Multitask Online Learning

Problem formulation and reminders on OMD. We consider an online convex optimization setting with a set of $N \in \mathbb{N}$ agents, each learning a possibly different task on a common convex decision set $V \subset \mathbb{R}^d$. At each time step $t = 1, 2, \ldots$ some agent $i_t \leq N$ makes a prediction $x_t \in V$ for its task, incurs loss $\ell_t(x_t)$, and observes a subgradient of ℓ_t at x_t, where ℓ_t is a convex loss function. We say that i_t is the active agent at time t. Both the sequence $i_1, i_2, \ldots$ of active agents and the sequence $\ell_1, \ell_2, \ldots$ of convex losses are chosen adversarially and hidden from the agents. Our goal is to minimize the *multitask regret*, which is defined as the sum of the individual regrets:

$$R_T = \sum_{i=1}^{N} \left(\sum_{t:\, i_t = i} \ell_t(x_t) - \inf_{u \in V} \sum_{t:\, i_t = i} \ell_t(u) \right) = \sum_{t=1}^{T} \ell_t(x_t) - \sum_{i=1}^{N} \inf_{u \in V} \sum_{t:\, i_t = i} \ell_t(u)\,. \tag{1}$$

A natural idea to minimize (1) is to run N independent OMDs, one for each agent. Recall that OMD refers to a family of algorithms, typically used to minimize a regret of the form $\sum_t \ell_t(x_t) - \inf_{u \in V} \sum_t \ell_t(u)$, for any sequence of proper convex loss functions ℓ_t. An instance of OMD is parameterized by a λ-strongly convex regularizer $\psi \colon \mathbb{R}^d \to \mathbb{R}$, and has the update rule

$$x_{t+1} = \operatorname*{argmin}_{x \in V} \quad \langle \eta_t g_t, x \rangle + B_\psi(x, x_t)\,,$$

where $g_t \in \mathbb{R}^d$ is a subgradient of ℓ_t at point x_t, $B_\psi(x, y) = \psi(x) - \psi(y) - \langle \nabla\psi(y), x - y \rangle$ denotes the Bregman divergence associated to ψ, and $\eta_t > 0$ is a tunable learning rate. Standard results allow to bound the regret achieved by the sequence of iterates produced by OMD. For a fixed η and any initial point $x_1 \in V$, we have (Orabona, 2019) that for all $u \in V$

$$\sum_{t=1}^{T} \ell_t(x_t) - \ell_t(u) \leq \frac{B_\psi(u, x_1)}{\eta} + \frac{\eta}{2\lambda} \sum_{t=1}^{T} \|g_t\|_\star^2\,,$$

with $\|\cdot\|_\star$ the dual norm of the norm with respect to which ψ is strongly convex. The choice of the regularizer ψ shapes the above bound through the quantities $B_\psi(u, x_1)$ and $\|g_t\|_\star$. When $\psi = \frac{1}{2}\|\cdot\|_2^2$, we have $B_\psi(x, y) = \frac{1}{2}\|x - y\|_2^2$, $\|\cdot\|_\star = \|\cdot\|_2$, $\lambda = 1$, and the

algorithm is called Online Gradient Descent (OGD). However, depending on the problem, a different choice of the regularizer might better captures the underlying geometry. A well-known example is Exponentiated Gradient (EG), an instance of OMD in which V is the probability simplex in $\mathbb{R}^d$, such that $V = \Delta \coloneqq \{x \in \mathbb{R}_+^d \colon \sum_j x_j = 1\}$. EG uses the negative entropy regularizer $x \mapsto \sum_j x_j \ln(x_j)$, and assuming that $\|g_t\|_\infty \leq L_g$, one achieves bounds of order $\mathcal{O}(L_g\sqrt{T \ln d})$, while OGD yields bounds of order $\mathcal{O}(L_g\sqrt{dT})$. We emphasize that our cooperative extension adapts to several types of regularizers, and can therefore exploit these improvements with respect to the dependence on d, see Proposition 5. Let C be a generic constant such that $C\sqrt{T}$ bounds the regret incurred by the chosen OMD (e.g., $C = L_g\sqrt{\ln d}$, or $C = L_g\sqrt{d}$ above), and $T_i = \sum_{t=1}^T \mathbb{I}\{i_t = i\}$ be the number of times agent i is active. Then, by Jensen's inequality the multitask regret of N independent OMDs satisfies: $R_T \leq \sum_{i=1}^N C\sqrt{T_i} \leq C\sqrt{NT}$.

A multitask extension. We first focus on OGD. For $i \leq N$ and $t \leq T$, let $x_{i,t} \in \mathbb{R}^d$ denote the prediction maintained by agent i at time step t. Note that the independent OGDs updates can be rewritten for all $i \leq N$ and t such that $i_t = i$:

$$x_{i,t+1} = \Pi_{V,\|\cdot\|_2}\big(x_{i,t} - \eta_t g_t\big)\,, \tag{2}$$

where $\Pi_{V,\|\cdot\|}$ denotes the projection operator onto the convex set V according to the norm $\|\cdot\|$, that is $\Pi_{V,\|\cdot\|}(x) = \operatorname{argmin}_{y\in V} \|x-y\|$. Our analysis relies on *compound representations*, that we explain next. We use bold notation to refer to compound vectors, such that for $u_1, \dots, u_N \in \mathbb{R}^d$, the compound vector is $\boldsymbol{u} = [u_1, \dots, u_N] \in \mathbb{R}^{Nd}$. For $i \leq N$, we use $\boldsymbol{u}^{(i)}$ to refer to the i^{th} block of $\boldsymbol{u}$, such that $\boldsymbol{u}^{(i)} = u_i$ in the above example. So $\boldsymbol{x}_t$ is the compound vector of the $(x_{i,t})_{i=1}^N$, such that $x_t = \boldsymbol{x}_t^{(i_t)}$, and the multitask regret rewrites as $R_T(\boldsymbol{u}) = \sum_{t=1}^T \ell_t\big(\boldsymbol{x}_t^{(i_t)}\big) - \ell_t\big(\boldsymbol{u}^{(i_t)}\big)$. For any set $V \subset \mathbb{R}^d$, let $\boldsymbol{V} = V^{\otimes N} \subset \mathbb{R}^{Nd}$ denote the compound set such that $\boldsymbol{u} \in \boldsymbol{V}$ is equivalent to $\boldsymbol{u}^{(i)} \in V$ for all $i \leq N$. Equipped with this notation, the independent OGD updates (2) rewrite as:

$$\boldsymbol{x}_{t+1} = \Pi_{\boldsymbol{V},\|\cdot\|_2}\big(\boldsymbol{x}_t - \eta_t \boldsymbol{g}_t\big)\,,$$

with $\boldsymbol{g}_t \in \mathbb{R}^{Nd}$ such that $\boldsymbol{g}_t^{(i)} = g_t$ for $i = i_t$, and $0_{\mathbb{R}^d}$ otherwise. In other words, only the active agent has a non-zero gradient and therefore makes an update. Our goal is to incorporate communication into this independent update. To that end, we consider the general idea of *sharing updates* by considering (sub) gradients of the form $\boldsymbol{A}^{-1}\boldsymbol{g}_t$, where $\boldsymbol{A}^{-1} \in \mathbb{R}^{Nd\times Nd}$ is a shortcut notation for $A^{-1} \otimes I_d$ and $A \in \mathbb{R}^{N\times N}$ is any symmetric positive definite interaction matrix. A simple computation shows that $(\boldsymbol{A}^{-1}\boldsymbol{g}_t)^{(i)} = A_{ii_t}^{-1} g_t$. Thus, every agent i makes an update proportional to $A_{ii_t}^{-1}$ at each time step t. In other words, the active agent (the only one to suffer a loss) shares its update with the other agents. Results in Section 3 are proved by designing a matrix A^{-1} (or equivalently A) such that $A_{ii_t}^{-1}$ captures the similarity between tasks i and i_t. Intuitively, the active agent

i_t should share its update (gradient) with another agent i to the extent their respective tasks are similar. Overall, denoting by $\|u\|_M = \sqrt{u^\top M u}$ the Mahalanobis norm of u, the MT-OGD update writes:

$$\widetilde{\boldsymbol{x}}_{t+1} = \operatorname*{argmin}_{\boldsymbol{x}\in\mathbb{R}^{Nd}} \ \langle \eta_t \boldsymbol{g}_t, \boldsymbol{x}\rangle + \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{x}_t\|_{\boldsymbol{A}}^2, \quad \text{and} \quad \boldsymbol{x}_{t+1} = \operatorname*{argmin}_{\boldsymbol{x}\in\boldsymbol{V}} \ \frac{1}{2}\|\boldsymbol{x}-\widetilde{\boldsymbol{x}}_{t+1}\|_{\boldsymbol{A}}^2 . \tag{3}$$

Equation (3) reveals that the MT-OGD update rule is an OMD update with regularizer $\boldsymbol{x} \mapsto \frac{1}{2}\|\boldsymbol{x}\|_{\boldsymbol{A}}^2 = \boldsymbol{\psi}_{\mathrm{Euc}}\big(\boldsymbol{A}^{1/2}\boldsymbol{x}\big)$. This provides a natural path for extending our multitask approach to any regularizer. Given a base regularizer $\psi\colon \mathbb{R}^d \to \mathbb{R}$, the *compound regularizer* $\boldsymbol{\psi}$ is given by $\boldsymbol{\psi}\colon \boldsymbol{x}\in\mathbb{R}^{Nd} \mapsto \sum_{i=1}^N \psi\big(\boldsymbol{x}^{(i)}\big)$. Using regularizer $\boldsymbol{\psi}\big(\boldsymbol{A}^{1/2}\cdot\big)$, whose associate divergence is $B_{\boldsymbol{\psi}}\big(\boldsymbol{A}^{1/2}\boldsymbol{x}, \boldsymbol{A}^{1/2}\boldsymbol{x}'\big)$, the MT-OMD update thus reads:

$$\boldsymbol{x}_{t+1} = \operatorname*{argmin}_{\boldsymbol{x}\in\boldsymbol{V}} \ \langle \eta_t \boldsymbol{g}_t, \boldsymbol{x}\rangle + B_{\boldsymbol{\psi}}\big(\boldsymbol{A}^{1/2}\boldsymbol{x}, \boldsymbol{A}^{1/2}\boldsymbol{x}_t\big) .$$

3 Regret Analysis

Throughout the section, $V \subset \mathbb{R}^d$ is a convex set of comparators, and $(\ell_t)_{t=1}^T$ is a sequence of proper convex loss functions chosen by the adversary.

Theorem 1. *Let $\psi\colon \mathbb{R}^d \to \overline{\mathbb{R}}$ be λ-strongly convex with respect to norm $\|\cdot\|$ on V, let $A \in \mathbb{R}^{N\times N}$ be symmetric positive definite, and set $\boldsymbol{x}_1 \in \boldsymbol{V}$. Then, MT-OMD with $\eta_t \coloneqq \eta$ produces a sequence $(\boldsymbol{x}_t)_{t=1}^T$ such that for all $\boldsymbol{u}\in\boldsymbol{V}$, $R_T(\boldsymbol{u})$ is bounded by*

$$\frac{B_{\boldsymbol{\psi}}\big(\boldsymbol{A}^{1/2}\boldsymbol{u}, \boldsymbol{A}^{1/2}\boldsymbol{x}_1\big)}{\eta} \ + \ \max_{i\le N} A_{ii}^{-1} \ \frac{\eta}{2\lambda}\sum_{t=1}^T \|g_t\|_\star^2 . \tag{4}$$

We now design matrices A that make (4) smaller than $\sqrt{NT}$. Let $L = I_N - \mathbb{1}\mathbb{1}^\top/N$. We consider matrices of the form $A(b) = I_N + b\,L$, where $b \ge 0$ quantifies the magnitude of the communication. We next show that MT-OMD run with such matrices achieves an important improvement for comparators satisfying the following variance condition. This multitask acceleration is then shown to be optimal.

Definition 1. *Let $\|\cdot\|\colon \mathbb{R}^d \to \mathbb{R}$ be any norm, and $\bar{\boldsymbol{u}} = (1/N)\sum_{i=1}^N \boldsymbol{u}^{(i)}$, for any $\boldsymbol{u}\in\mathbb{R}^{Nd}$. We define the variance of $\boldsymbol{u}$ with respect to $\|\cdot\|$ as $\mathrm{Var}_{\|\cdot\|}(\boldsymbol{u}) = \frac{1}{N-1}\sum_{i=1}^N \big\|\boldsymbol{u}^{(i)} - \bar{\boldsymbol{u}}\big\|^2$. Let $D = \sup_{u\in V}\|u\|$, and $\sigma > 0$. The comparators with variance smaller than $\sigma^2 D^2$ are denoted by $\boldsymbol{V}_{\|\cdot\|,\sigma} = \big\{\boldsymbol{u}\in\boldsymbol{V}\colon \mathrm{Var}_{\|\cdot\|}(\boldsymbol{u}) \le \sigma^2 D^2\big\}$.*

Proposition 2. *Let $\psi = \frac{1}{2}\|\cdot\|_2^2$, $D = \sup_{x\in V}\|x\|_2$, and $\sigma \le 1$. Assume that $\|\partial\ell_t(x)\|_2 \le L_g$ for all $t \le T$ and any $x \in V$. Set $b = N$, $\boldsymbol{x}_1 = \boldsymbol{0}$, and $\eta = D\sqrt{N(N+1)(1+(N-1)\sigma^2)}/L_g\sqrt{2T}$. Then, MT-OGD produces a sequence $(\boldsymbol{x}_t)_{t=1}^T$ such that for all $\boldsymbol{u}\in\boldsymbol{V}_{\|\cdot\|_2,\sigma}$:*

$$R_T(\boldsymbol{u}) \le D L_g \sqrt{1+\sigma^2(N-1)}\sqrt{2T}. \tag{5}$$

Proposition 3. *Under the conditions of Proposition 2, the regret of any algorithm satisfies*

$$\sup_{\boldsymbol{u}\in\boldsymbol{V}_{\|\cdot\|_2,\sigma}} R_T(\boldsymbol{u}) \geq \frac{1}{4}\left(DL_g\sqrt{1+\sigma^2(N-1)}\sqrt{2T}\right).$$

In the next proposition, we show that multitask acceleration can be achieved with arbitrary norms. A crucial application is the use of the p-norm on the probability simplex.

Proposition 4. *Let $\|\cdot\|\colon \mathbb{R}^d \to \mathbb{R}$ be any norm, $\psi = \frac{1}{2}\|\cdot\|^2$, $D = \sup_{x\in V}\|x\|$, and $\sigma \leq 1$. Assume that $\|\partial\ell_t(x)\|_\star \leq L_g$ for all $t \leq T$, $x \in V$. Set $b = N$, $\boldsymbol{x}_1 = \mathbf{0}$ and $\eta = D\sqrt{N(N+1)(1+(N-1)\sigma^2)}/L_g\sqrt{2T}$. Then, MT-OMD produces a sequence $(\boldsymbol{x}_t)_{t=1}^T$ such that for all $\boldsymbol{u} \in \boldsymbol{V}_{\|\cdot\|,\sigma}$ we have $R_T(\boldsymbol{u}) \leq DL_g\sqrt{1+\sigma^2(N-1)}\sqrt{8T}$. In particular, for $d \geq 3$ and $V = \Delta$, choosing $\|\cdot\| = \|\cdot\|_p$, for $p = 2\ln d/(2\ln d - 1)$, and assuming that $\|\partial\ell_t(x)\|_\infty \leq L_g$, it holds for all $\boldsymbol{u} \in \boldsymbol{\Delta}_{\|\cdot\|_p,\sigma}$:*

$$R_T(\boldsymbol{u}) \leq L_g\sqrt{1+\sigma^2(N-1)}\sqrt{16e\ T\ln d}.$$

In comparison, under the same assumptions, (5) would write: $L_g\sqrt{1+\sigma^2(N-1)}\sqrt{2Td}$.

The next proposition is dedicated to regularizers on the simplex that have value $+\infty$ outside of it (as opposed to p-norms). The following variance definition is then necessary.

Definition 2. *For any $\boldsymbol{u} \in \mathbb{R}^d$ and $j \leq d$, let $\boldsymbol{u}_j^{\max} = \max_{i\leq N}\boldsymbol{u}_j^{(i)}$, and $\boldsymbol{u}_j^{\min} = \min_{i\leq N}\boldsymbol{u}_j^{(i)}$. Then, with the convention $0/0 = 0$ we define $\mathrm{Var}_\Delta(\boldsymbol{u}) = \max_{j\leq d}\left(1 - \boldsymbol{u}_j^{\min}/\boldsymbol{u}_j^{\max}\right)^2$, and for any $\sigma \leq 1$, $\boldsymbol{\Delta}_\sigma = \left\{\boldsymbol{u} \in \boldsymbol{\Delta}\colon \mathrm{Var}_\Delta(\boldsymbol{u}) \leq \sigma^2\right\}$.*

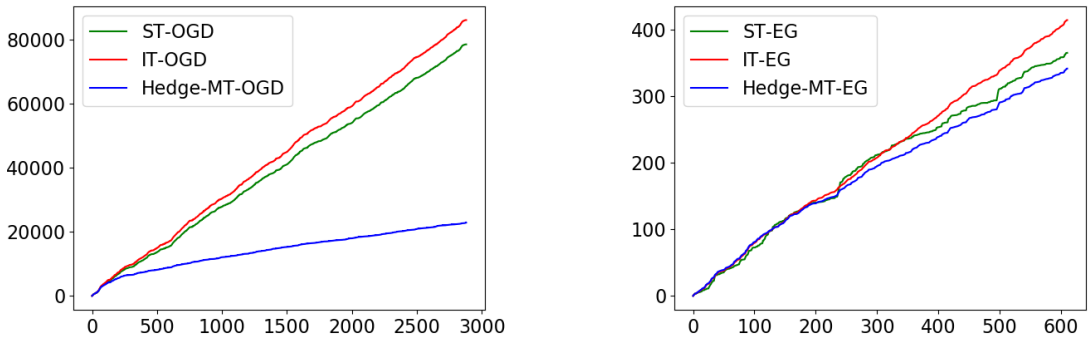
Proposition 5. *Let $\psi : \Delta \to \mathbb{R}$ be λ-strongly convex w.r.t. norm $\|\cdot\|$, and such that there exist $x^* \in \Delta$ and $C < +\infty$ such that for all $x \in \Delta, B_\psi(x, x^*) \leq C$. Note that the negative entropy satisfies this assumption with $x^* = \mathbb{1}/d$, $C = \ln d$, $\lambda = 1$, and $\|\cdot\| = \|\cdot\|_1$. Let $\sigma \leq 1$, and assume that $\|\partial\ell_t(x)\|_\star \leq L_g$ for all $t \leq T$ and $x \in \Delta$. Set $b = (1-\sigma^2)/\sigma^2$, $\boldsymbol{x}_1 = [x^*, \ldots, x^*]$, and $\eta = N\sqrt{2\lambda(1+b)C}/L_g\sqrt{(b+N)T}$. Then, MT-OMD produces a sequence $(\boldsymbol{x}_t)_{t=1}^T$ such that for all $\boldsymbol{u} \in \boldsymbol{\Delta}_\sigma$:*

$$R_T(\boldsymbol{u}) \leq L_g\sqrt{1+\sigma^2(N-1)}\sqrt{2CT/\lambda}\,.$$

Finally, we present an Hedge-based extension of MT-OMD, denoted Hedge-MT-OMD, that does not require any prior information on σ^2. It consists then in using Hedge—see, e.g., (Orabona, 2019, Section 6.8)—over a set of experts, each running an instance of MT-OMD with a different value of σ^2 chosen on a uniform grid of the interval $[0, 1]$.

Theorem 6. *Let $D = \sup_{x\in V}\|x\|_2$, and assume that $\|\partial\ell_t(x)\|_2 \leq L_g$ for all $t \leq T$, $x \in V$. Then, for all $\boldsymbol{u} \in \boldsymbol{V}$ the regret of Hedge-MT-OGD is bounded by*

$$DL_g\left(2 + \sqrt{\log N} + \sqrt{\min\left\{\mathrm{Var}_{\|\cdot\|_2}(\boldsymbol{u}), 1\right\}\cdot N}\right)\sqrt{2T}\,.$$



(a) Lenk (OGD, square loss)

(b) EMNIST (EG, logistic loss)

Figure 1: Comparison between multitask (MT), independent-task (IT), and single-task (ST) OGD and EG on the *Lenk* and *EMNIST* datasets. We plot the cumulative losses against time. *Lenk* is known to work well in multi-task settings, and Hedge-MT-OGD performs significantly better than both baselines. *EMNIST* has a variance significantly higher but Hedge-MT-EG is still outperforming the baselines, though by a small margin.

4 Experiments

We now empirically compare the performance of Hedge-MT-OGD/EG against two natural alternatives: an independent-task approach (IT-OGD/EG) where the agents do not communicate, and a single-task approach (ST-OGD/EG) where a single model is learned and shared by all agents. Note that both IT and ST approaches are special cases of MT-OMD, obtained respectively with $b = 0$ (i.e., $\sigma^2 \geq 1$), or $b = +\infty$ (i.e., $\sigma^2 = 0$).

Online Gradient Descent. We use the *Lenk* dataset, that consists of 2880 computer ratings, made by 180 individuals (the tasks) on the basis of 14 binary features. We ran Hedge-MT-OGD with $A = (1 + N)I_N - \mathbb{1}\mathbb{1}^\top$ and the square loss. We used the theoretical values for η, while σ^2 is learned in a set of 5 experts uniformly spaced over $[0, 1]$.

Exponentiated Gradient. We consider *EMNIST*, that consists of 62 classes (images of digits, small and capital letters). We created 61 binary classification tasks by considering the 0 digit class against each other class. We used the linear logistic regression and ran Hedge-MT-EG with $A(b) = (1 + b)I_N - b\mathbb{1}\mathbb{1}^\top/N$. We used the theoretical values for η and b, while σ^2 is learned in a set of 5 experts uniformly spaced over $[0, 1]$.

References

Caruana, R. (1997). Multitask learning. *Machine Learning*, 28(1):41–75.

Orabona, F. (2019). A modern introduction to online learning. *Preprint, arXiv:1912.13213*.

Identification de Meilleur Bras à Budget Fixé

Antoine Barrier [1,2] & Aurélien Garivier [1] & Gilles Stoltz [2]

[1] *Univ. Lyon, ENS de Lyon, UMPA UMR 5669, 46 allée d'Italie, 69364 Lyon cedex 07*
[2] *Université Paris-Saclay, CNRS, Laboratoire de mathématiques d'Orsay, 91405, Orsay, France*

antoine.barrier@ens-lyon.fr | aurelien.garivier@ens-lyon.fr | gilles.stoltz@universite-paris-saclay.fr

Résumé. Le problème d'identification de meilleur bras consiste à trouver, parmi un ensemble de distributions, celle de plus grande espérance en observant de manière séquentielle des réalisations indépendantes de ces lois. Les applications (essais cliniques, systèmes de recommandation, ...) en font un cadre d'étude dense pour lequel plusieurs objectifs sont considérés. Dans le cadre du budget fixé, le nombre d'observations est limité et on s'intéresse aux propriétés d'algorithmes qui minimisent la probabilité d'identifier une mauvaise distribution (probabilité d'erreur). Dans ce travail on fait apparaître une notion de complexité à base de théorie de l'information, complétant les travaux préalables qui ne faisaient intervenir que les écarts entre les moyennes des distributions.

Mots-clés. apprentissage séquentiel, identification de meilleur bras, budget fixé

Abstract. The problem of best arm identification consist to find the distribution of highest expectation among a set of distributions by sequentially observing independent realizations of those laws. Numerous applications (clinical trials, recommander system, ...) make it an important framework for which several objectives can be considered. In the fixed budget setting, the number of observations is limited and we look for algorithms that minimize the probability of misidentifying the best distribution. In this work we make appear a notion of complexity based on information theory, specifying previous works that were only based on gaps between the means of the distributions.

Keywords. sequential learning, best arm identification, fixed budget

1 Introduction

Un *bandit* est un ensemble de distributions de probabilités, appelées *bras*, avec lequel un algorithme interagit de manière séquentielle en choisissant à chaque pas de temps, en s'appuyant sur les observations passées, un des bras et en recevant une réalisation de la distribution associée. Dans ce papier on suppose que les distributions des bras appartiennent à un modèle exponentiel (on pourra penser à des gaussiennes de variance fixée ou à des Bernoulli) et on note $d(\mu, \mu')$ la divergence de Kullback-Leibler entre les lois

de moyennes μ et μ' de ce modèle. Les distributions étant caractérisées par leur moyenne, un bandit est identifié par la donnée de son *vecteur d'espérances* $\boldsymbol{\mu} = (\mu_a)_{1\leq a\leq K} \in \mathbb{R}^K$ où $K \geq 2$ est le nombre de bras, on note alors $a^*(\boldsymbol{\mu}) = \text{argmax}_{1\leq a\leq K} \mu_a$ l'ensemble de ses *bras optimaux* et $\Delta_a = \mu_{a^*(\boldsymbol{\mu})} - \mu_a$ le *gap* du bras $a \in \{1, \dots, K\}$.

Dans ce papier on s'intéresse à l'identification du meilleur bras (BAI pour Best Arm Identification) à budget fixé. Une stratégie collecte un nombre d'observations $T \in \mathbb{N}^*$ et retourne un bras I_T qu'elle estime être de plus haute espérance. On peut penser par exemple à un essai clinique, où l'on teste plusieurs médicaments sur un panel constitué d'un nombre fini de patients. Pour être performante, la stratégie doit minimiser la probabilité d'erreur $p_T(\boldsymbol{\mu}) = \mathbb{P}_{\boldsymbol{\mu}}(I_T \notin a^*(\boldsymbol{\mu}))$ commise en estimant le meilleur bras d'un bandit $\boldsymbol{\mu}$. On cherche alors à trouver des bornes inférieures, assurant que toute stratégie raisonnable vérifie $p_T(\boldsymbol{\mu}) \gtrsim \exp(-\frac{T}{C(\boldsymbol{\mu})})$ pour une constante $C(\boldsymbol{\mu})$ la plus grande possible, appelée *complexité* ; autrement dit on cherche à déterminer

$$C(\boldsymbol{\mu}) = \inf_{\text{stratégies "raisonnables"}} \left(\limsup_{T\to+\infty} -\frac{1}{T}\log p_T(\boldsymbol{\mu})\right)^{-1}.$$

Parallèlement, on cherche des stratégies garantissant une borne supérieure proche de $C(\boldsymbol{\mu})$. Si le cadre du BAI à budget fixé intéresse pour ces nombreuses applications, c'est un problème assez difficile qui n'a été que partiellement traité jusqu'à présent (Audibert et al., 2010; Karnin et al., 2013; Kaufmann et al., 2016; Carpentier and Locatelli, 2016). $C(\boldsymbol{\mu})$ n'a pas encore été identifiée pour des modèles simples (gaussiennes, Bernoulli) : les bornes inférieure et supérieure ne correspondent pas et sont basées uniquement sur les gaps. Il n'y a également pas encore d'étude systématique faisant apparaître des complexités plus fines pour des modèles exponentiels.

En comparaison, le BAI à *confiance* fixée (un algorithme ne peut pas prendre de décision tant qu'il n'est pas certain de se tromper avec probabilité contrôlée) est beaucoup mieux compris à ce jour, la complexité ayant été identifiée et des algorithmes optimaux proposés (voir notamment Garivier and Kaufmann, 2016).

Dans la Section 2, nous présentons des résultats de borne inférieure, puis nous analysons l'algorithme de rejet successif dans la Section 3 afin d'obtenir une borne supérieure, dont nous discutons le lien avec la borne inférieure dans la Section 4.

2 Borne inférieure

Dans un premier temps on cherche une borne inférieure à $p_T(\boldsymbol{\mu})$ ou plus précisément à $C(\boldsymbol{\mu})$ pour un bandit $\boldsymbol{\mu}$ ayant un unique bras optimal. Cela nécessite de s'entendre sur la notion de stratégie "raisonnable" définissant $C(\boldsymbol{\mu})$.

Tout d'abord, une stratégie doit pouvoir identifier le meilleur bras avec forte probabilité lorsque le budget est conséquent, et ce quelque soit le problème de bandit. On considère donc des stratégies *identifiant asymptotiquement le meilleur bras*, c'est-à-dire telles que $p_T(\boldsymbol{\mu}) = o_{T\to+\infty}(1)$ pour tout bandit $\boldsymbol{\mu}$.

De plus, le comportement d'un algorithme sous un bandit $\boldsymbol{\mu}$ doit être le même que celui sous $\sigma(\boldsymbol{\mu})$ un bandit pour lequel les bras de $\boldsymbol{\mu}$ ont été permutés (autrement dit la stratégie ne doit dépendre que des récompenses et pas des indices des bras), sinon on observe de mauvaises performances sur certains bandits spécifiques. Une telle stratégie est dite *symétrique*.

Enfin, on peut s'attendre à ce qu'une stratégie performante soit *monotone*, c'est-à-dire que meilleur est l'espérance d'un bras, plus ce bras est tiré (en espérance). C'est l'hypothèse que l'on fera dans un premier temps. Cependant, il n'a pas été démontré à ce jour que les meilleures stratégies sont monotones, donc on relâchera cette hypothèse dans un second temps pour obtenir une seconde borne plus générale.

Ces 3 propriétés ne sont pas restrictives : les stratégies proposées à ce jour (par exemple la stratégie uniforme qui tire chaque bras T/K fois et recommande le bras de meilleure moyenne empirique) vérifient les 3 critères.

2.1 Pour un algorithme monotone

Si l'on ne considère que des stratégies monotones, on a le résultat suivant :

Proposition 1. *Soit $\boldsymbol{\mu}$ un problème de bandit ayant un unique meilleur bras et dont on suppose les bras ordonnés $\mu_1 > \mu_2 \geq \mu_3 \geq \cdots \geq \mu_K$. Alors pour toute stratégie identifiant le meilleur bras, symétrique et monotone on a :*

$$p_T(\boldsymbol{\mu}) \geq \exp\left(-\frac{T}{\frac{1}{2}\max\limits_{2\leq k\leq K}\frac{k}{d(\mu_k,\mu_1)}}(1+o(1))\right) \quad \textit{soit} \quad C(\boldsymbol{\mu}) \geq \frac{1}{2}\max_{2\leq k\leq K}\frac{k}{d(\mu_k,\mu_1)}\,. \tag{1}$$

Cette borne généralise celle de Audibert et al. (2010) qui ont montré que pour des Bernoulli de moyennes dans $[p, 1-p]$:

$$C(\boldsymbol{\mu}) \geq \frac{p(1-p)}{6}\max_{2\leq k\leq K}\frac{k}{\Delta_{(k)}^2}$$

où les gaps sont triés dans l'ordre $\Delta_{(2)} \leq \cdots \leq \Delta_{(K)}$. Contrairement à notre borne, celle-ci dépend des gaps et non de d, ce qui la rend moins informative et donne une borne plus faible (on peut montrer que $d(\mu_k,\mu_1) \leq \frac{\Delta_k^2}{2p(1-p)}$). Notons que la preuve est valable pour des stratégies quelconques (pas nécessairement monotones), mais le raisonnement utilisé, spécifique aux Bernoulli, est moins limpide que celui proposé ci-dessous.

Schéma de preuve. On notera $N_a(T)$ le nombre total d'observations du bras $a \in \{1,\ldots,K\}$. On choisit $L \in \operatorname{argmax}_{2\leq k\leq K}\frac{k}{d(\mu_k,\mu_1)}$ et on considère le bandit $\boldsymbol{\lambda} = (\mu_L, \mu_2, \ldots, \mu_K)$ (seul le premier bras est modifié par rapport à $\boldsymbol{\mu}$).

En utilisant une technique de changement de loi et en notant kl la divergence de Kullback-Leibler entre lois de Bernoulli on obtient :

$$\begin{aligned}
p_T(\boldsymbol{\mu}) &= \mathbb{P}_{\boldsymbol{\mu}}(I_T \neq 1) \\
&\geq \exp\left(-\frac{\mathrm{kl}(\mathbb{P}_{\boldsymbol{\lambda}}(I_T \neq 1), \mathbb{P}_{\boldsymbol{\mu}}(I_T \neq 1)) + \log 2}{\mathbb{P}_{\boldsymbol{\lambda}}(I_T \neq 1)}\right) \qquad \text{(inégalité kl)} \\
&\geq \exp\left(-\frac{\sum_{b\in[K]} \mathbb{E}_{\boldsymbol{\lambda}}[N_b(T)] d(\mu_b, \lambda_b) + \log 2}{\mathbb{P}_{\boldsymbol{\lambda}}(I_T \neq 1)}\right) \qquad \text{(in. de data-compressing)} \\
&= \exp\left(-\frac{\mathbb{E}_{\boldsymbol{\lambda}}[N_1(T)] d(\mu_L, \mu_1) + \log 2}{\mathbb{P}_{\boldsymbol{\lambda}}(I_T \neq 1)}\right).
\end{aligned}$$

Le bras 1 étant le L-ième meilleur bras sous le bandit $\boldsymbol{\lambda}$, on a que $\mathbb{E}_{\boldsymbol{\lambda}}[N_1(T)] \leq T/L$ par monotonicité et symétrie et donc

$$\mathbb{E}_{\boldsymbol{\lambda}}[N_1(T)] d(\mu_L, \mu_1) \leq \frac{T}{L} d(\mu_L, \mu_1) = \frac{T}{\max_{2\leq k\leq K} \frac{k}{d(\mu_k, \mu_1)}}.$$

Par ailleurs, comme la stratégie identifie asymptotiquement le meilleur bras, on a $\liminf_{T\to+\infty} \mathbb{P}_{\boldsymbol{\lambda}}(I_T \neq 1) \geq \frac{1}{2}$ (si $\mu_L < \mu_2$, cela tend vers 1, sinon tous les bras sont les mêmes sous $\boldsymbol{\lambda}$ et alors la probabilité tend par symétrie vers $\frac{L-1}{L} \geq \frac{1}{2}$).

En combinant ces deux observations on obtient la borne souhaitée. □

2.2 Vers une borne inférieure pour toute stratégie

En généralisant la preuve de Kaufmann et al. (2016, Theorem 16), on peut également montrer le résultat suivant valable pour des stratégies non nécessairement monotones.

Théorème 2. *Soit $\boldsymbol{\mu}$ un problème de bandits d'unique meilleur bras 1. Alors il existe un problème de bandit $\boldsymbol{\lambda}$ d'unique meilleur bras $b \neq 1$ tel que $\sum_{k\neq b} \frac{1}{d(\lambda_k, \lambda_b)} \leq \sum_{k\neq 1} \frac{1}{d(\mu_k, \mu_1)}$ et pour toute stratégie identifiant asymptotiquement le meilleur bras et symétrique, on a*

$$\max\big(p_T(\boldsymbol{\mu}), p_T(\boldsymbol{\lambda})\big) \geq \exp\Big(-\frac{T}{\sum_{k\neq 1} \frac{1}{d(\mu_k, \mu_1)} \times r(\mu_b, \mu_1)}\Big)$$

où $r(\mu_b, \mu_1) = \frac{d(\mu_b, \mu_1)}{d(\mu_b, \overline{\mu}_b)}$ avec $\overline{\mu}_b$ le symétrique de μ_b par rapport à μ_1.

Moralement ce résultat signifie la borne sur la complexité

$$C(\boldsymbol{\mu}) \geq \sum_{k\neq 1} \frac{1}{d(\mu_k, \mu_1)} \times \min_{2\leq b\leq K} r(\mu_b, \mu_1). \tag{2}$$

Cette borne fait apparaître la quantité $\sum_{k\neq 1} \frac{1}{d(\mu_k, \mu_1)}$, qui est légèrement meilleure que le $\max_{2\leq k\leq K} \frac{k}{d(\mu_k, \mu_1)}$ de la borne (1).

Cependant le terme supplémentaire avec les rapports de divergence rend la borne moins claire. On se pose alors la question de minorer les $(r(\mu_b, \mu_1))_{2\leq b\leq K}$. Avec des gaussiennes de variance fixée ce rapport vaut toujours 1/4, il s'agit du résultat de Kaufmann et al. (2016). Avec des Bernoulli on peut minorer ce rapport par un terme de variance $p(1-p)$. Pour un modèle exponentiel quelconque, il reste à comprendre s'il est possible d'avoir une minoration explicite, l'espoir étant que le rapport soit proche de 1/4 (ce qui est le cas lorsque les gaps sont petits).

3 Borne supérieure : l'algorithme de Rejet Successif

L'algorithme de rejet successif (Audibert et al., 2010) consiste à éliminer les bras un par un lors de $K-1$ phases durant lesquelles les bras restants sont tirés de manière uniforme et le bras de moins bonne moyenne empirique est éliminé. On notera n_r le nombre d'observations d'un bras encore actif à la fin des r premières phases.

On définit $D(\mu, \mu')$ l'information de Chernoff entre les distributions de moyennes $\mu \neq \mu'$ du modèle exponentiel, c'est-à-dire $D(\mu, \mu') = d(\theta, \mu)$ où θ est tel que $d(\theta, \mu) = d(\theta, \mu')$.

On a alors la borne supérieure suivante :

Théorème 3. *Soit $\boldsymbol{\mu}$ un problème de bandits d'unique meilleur bras* 1. *Alors l'algorithme de rejet successif avec le choix $n_r = \left\lfloor \frac{T-K}{\overline{\log}(K)(K+1-r)} \right\rfloor$ pour $r \in \{1, \dots, K-1\}$ satisfait :*

$$p_T(\boldsymbol{\mu}) \leq 2KT \exp\Big(-\frac{T-K}{\frac{\overline{\log}(K)}{2} \times \max\limits_{2\leq k\leq K} \frac{k}{D(\mu_k,\mu_1)}}\Big) \quad \textit{soit} \quad C(\boldsymbol{\mu}) \leq \frac{\overline{\log}(K)}{2} \max_{2\leq k\leq K} \frac{k}{D(\mu_k, \mu_1)} \tag{3}$$

où $\overline{\log}(K) = \frac{1}{2} + \sum_{k=2}^{K} \frac{1}{k} \sim_{K\to+\infty} \log(K)$.

Audibert et al. (2010, Theorem 2) ont montré que

$$C(\boldsymbol{\mu}) \leq \overline{\log}(K) \max_{2\leq k\leq K} \frac{k}{\Delta_{(k)}^2}$$

pour des variables sous-gaussiennes de paramètre 1/2. Une nouvelle fois la borne proposée généralise celle-ci, puisque l'on peut vérifier dans ce cas que $D(\mu_k, \mu_1) = \frac{\Delta_k^2}{2}$. Il reste encore à faire parler la borne (3) pour des modèles quelconques. Le facteur $\overline{\log}(K)$, qui n'est pas satisfaisant, apparaît dans ces deux bornes, mais une manière plus astucieuse de choisir les $(n_r)_{1\leq r\leq K-1}$ (peut-être de manière adaptative) pourrait permettre de le faire disparaître, voire même de faire apparaître une borne de forme plus proche de la borne inférieure (2).

4 Discussion sur les bornes obtenues

Dans cette section on discute des liens entre les bornes proposées dans ce document.

Comparaison des bornes inférieures et supérieures Entre les bornes (2) et (3), il y a le facteur $\overline{\log}(K)$ (de l'ordre de $\log(K)$) qui n'est pas satisfaisant car ne permet pas de faire correspondre les deux bornes (même pour des modèles simples comme des gaussiennes). Pour la comparaison entre les deux maximums définis dans les bornes, on peut espérer que $d(\mu_k, \mu_1) \simeq 4D(\mu_1, \mu_k)$ (c'est le cas par exemple des gaussiennes et est une bonne approximation lorsque le gap est faible) si bien que dans ce cas les deux maximums sont très proches à un facteur multiplicatif près.

Conclusion Dans l'objectif de mieux comprendre la complexité du problème de BAI à budget fixé avec un modèle exponentiel, nous avons obtenu des généralisations des bornes inférieures existantes ainsi que de la borne supérieure de l'algorithme de rejet successif. Il reste à faire parler davantage ces nouvelles bornes selon le modèle exponentiel.

Bibliographie

Audibert, J.-Y., Bubeck, S., and Munoz, R. (2010). Best Arm Identification in Multi-Armed Bandits. In *COLT - 23th Conference on Learning Theory - 2010*, page 13 p., Haifa, Israel.

Carpentier, A. and Locatelli, A. (2016). Tight (lower) bounds for the fixed budget best arm identification bandit problem. In Feldman, V., Rakhlin, A., and Shamir, O., editors, *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 590–604, Columbia University, New York, New York, USA. PMLR.

Garivier, A. and Kaufmann, E. (2016). Optimal best arm identification with fixed confidence. In Feldman, V., Rakhlin, A., and Shamir, O., editors, *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 998–1027, Columbia University, New York, New York, USA. PMLR.

Karnin, Z., Koren, T., and Somekh, O. (2013). Almost optimal exploration in multi-armed bandits. In Dasgupta, S. and McAllester, D., editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1238–1246, Atlanta, Georgia, USA. PMLR.

Kaufmann, E., Cappé, O., and Garivier, A. (2016). On the complexity of best-arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17(1):1–42.

On Best Policy Identification with Instance-specific sample complexity

Aymen Al Marjani

UMPA, ENS Lyon
Lyon, France
aymen.al_marjani@ens-lyon.fr

Résumé. Le problème de l'identification de la meilleure politique (BPI) a récemment suscité l'intérêt de la communauté de l'apprentissage par renforcement (RL). Formellement, un agent de RL interagit avec un processus de décision Markovien (PDM) et vise à apprendre une politique ε-optimale avec une grande confiance en utilisant le moins d'observations possible. Dans le cas particulier d'un PDM avec un facteur d'actualisation $\gamma = 0$ (ou de manière équivalente lorsque l'horizon $H = 1$), le problème se réduit à l'identification du meilleur bras (BAI) dans un bandit multi-armé. Inspirés par le succès des algorithmes du type Track-and-Stop dans ce cas particulier, nous proposons des algorithmes pour BPI avec des bornes spécifiques à l'instance sur leur complexité d'échantillon. Notamment, nos algorithmes sont optimaux pour l'instance dès lors qu'un oracle résolvant un certain problème d'optimisation est disponible.

Mots-clés. Complexité d'échantillon, Processus de décision Markovien.

Abstract. The problem of Best Policy Identification (BPI) has recently attracted interest in the Reinforcement Learning (RL) community. Formally, a RL agent interacts with a Markov Decision Process (MDP) and aims to learn an ε-optimal policy with high confidence using as few observations as possible. In the special case of a MDP with discount factor $\gamma = 0$ (or equivalently when the horizon $H = 1$), the problem reduces to Best Arm Identification (BAI) in a multi-armed bandit. Inspired by the success of Track-and-Stop style algorithms in this special case, we propose algorithms for BPI with instance-specific bounds on their sample complexity. Notably, our algorithms are instance-optimal whenever an oracle solving a certain optimization problem is available.

Keywords. Sample Complexity, Markov Decision Processes.

1 Introduction

The BPI framework was introduced by [Fie94] under the name of *PAC-RL*. In BPI the algorithm explores the MDP until it has gathered enough samples to return an ε-optimal policy with probability at least $1 - \delta$. Crucially, the algorithm halts at a *random time step*, determined by a *stopping rule* which guarantees that the probability of returning a

wrong answer is less that δ. The optimality of BPI algorithms is measured through their *sample complexity*, defined as their expected stopping time. Best policy identification in MDPs has been mostly investigated under the lens of minimax-optimality. The minimax framework may be overly pessimistic by accounting for the worst possible MDP, whereas an algorithm with instance-specific guarantees can adapt its exploration procedure to the hardness of the MDP instance that it faces. Recently, a few works made the first attempts towards understanding the instance-specific sample complexity of reinforcement learning, whether under the assumption of access to a generative model [ZKB19, MP21], or for episodic MDPs [WSJ21].

Minimax Best-Policy Identification. BPI in the online setting has been investigated recently by a number of works with minimax sample complexity bounds. In the case of episodic MDPs [KMDD+21], [MDJ+20] proposed algorithms that identify an ε-optimal policy at the initial state w.h.p. In contrast, in the case of infinite-horizon MDPs one is rather interested in finding a good policy at *every* state. Recent works provide convergence analysis for Q-learning [LWC+20] or policy gradient algorithms [AHKS20], [FYAY21], [ZCA21]. Their results typically state that if the algorithm is fed with the appropriate hyperparameters, it can return an ε-optimal policy w.h.p. after collecting a polynomial number of samples. In practice, a pure exploration agent needs a stopping rule to determine when to halt the learning process. In particular, the question of how to tune the number of iterations of these algorithms without prior knowledge of the ground truth instance remains open.

2 Setting and notation

2.1 Infinite Horizon MDPs

We consider infinite horizon Markov Decision Processes (MDP) with discount and finite state and action spaces $\mathcal{S}$ and $\mathcal{A}$. Such a MDP $\mathcal{M}$ is defined through its transition and reward distributions $p_{\mathcal{M}}(\cdot|s,a)$ and $q_{\mathcal{M}}(\cdot|s,a)$ for all (s,a). For simplicity, $q_{\mathcal{M}}(\cdot|s,a)$ will denote the density of the reward distribution w.r.t. a positive measure λ with support included in $[0,1]$. Specifically, $p_{\mathcal{M}}(s'|s,a)$ denotes the probability of transitioning to state s' after playing action a at state s while $R(s,a)$ is the random immediate reward that is collected. Finally, $\gamma \in [0,1)$ is a discounting factor. An optimal policy $\pi^\star$ is a mapping $\mathcal{S} \to \mathcal{A}$ that maximizes the long-term discounted reward:

$$\pi^\star \in \arg\max_{\pi:\ \mathcal{S}\to\mathcal{A}} V^\pi_{\mathcal{M}}, \quad V^\pi_{\mathcal{M}} = \mathbb{E}_{\mathcal{M},\pi}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, \pi(s_t))\right],$$

where $\mathbb{E}_{\mathcal{M},\pi}$ denotes the expectation w.r.t the randomness in the rewards and the trajectory when the policy π is played. Classically, the value of the optimization problem

above is denoted by $V^\star_{\mathcal{M}}$ and referred to as the optimal value function of $\mathcal{M}$. For simplicity, **we will assume that the optimal policy is unique.**

2.2 Best Policy Identification

When $\mathcal{M}$ is unknown, we wish to devise a learning algorithm $\mathbb{A}$ that identifies $\pi^\star$ as quickly as possible with some given level of confidence.

Definition 1. *An algorithm $\mathbb{A}$ is δ-Probably Approximately Correct (δ-PAC) over some set $\mathbb{M}$ of MDPs, if for any $\mathcal{M} \in \mathbb{M}$, it returns (in finite time) the optimal policy with probability at least $1-\delta$. The performance of a δ-PAC algorithm $\mathbb{A}$ is measured through its sample complexity $\mathbb{E}_{\mathcal{M},\mathbb{A}}[\tau_\delta]$, where τ_δ is the stopping time at which we stop collecting observations.*

We distinguish two possible modes of interaction with the MDP:
Generative setting. We have access to a *generative model*, aka a simulator of the MDP. Thus the agent can query a sample (i.e. observe the reward and next-state) from any state-action pair (s,a) independently of its current state.
Online setting. The agent can only collect observations from a single controlled trajectory by choosing actions and following the resulting transitions.
For BPI to be relevant in the second case, **we will assume that MDP is communicating**, i.e. we can reach any state s starting from any other state s' using some policy.

3 Main Results

In this section, we present our lower and upper bounds on the number of samples needed by any δ-PAC algorithm to find the optimal policy. Before we state our first result, let us introduce some notation. Σ denotes the simplex of $\mathbb{R}^{SA}$ and $\mathrm{KL}_{\mathcal{M}|\mathcal{M}'}(s,a) := KL(p_{\mathcal{M}}(\cdot|s,a)\|p_{\mathcal{M}'}(\cdot|s,a)) + KL(q_{\mathcal{M}}(\cdot|s,a)\|q_{\mathcal{M}'}(\cdot|s,a))$ denotes the local Kullback-Leibler divergence between MDPs $\mathcal{M}$ and $\mathcal{M}'$ at (s,a). Finally, $\mathrm{Alt}(\mathcal{M}) := \{\mathcal{M}' \text{ MDP} : \pi^\star \text{ is not optimal in } \mathcal{M}'\}$ is the set of alternative MDPs.

Proposition 1. *([MP21, AGP21]) The expected sample complexity $\mathbb{E}_{\mathcal{M},\mathbb{A}}[\tau_\delta]$ of any δ-PAC algorithm $\mathbb{A}$ satisfies $\liminf_{\delta\to 0} \frac{\mathbb{E}_{\mathcal{M},\mathbb{A}}[\tau]}{\log(1/\delta)} \geq T^\star(\mathcal{M})$ where*

$$T^\star(\mathcal{M})^{-1} = \sup_{\omega\in\Omega(\mathcal{M})} \inf_{\mathcal{M}'\in\mathrm{Alt}(\mathcal{M})} \sum_{s,a} \omega_{sa} KL_{\mathcal{M}|\mathcal{M}'}(s,a), \tag{1}$$

and $\Omega(\mathcal{M}) := \Sigma$ in the generative setting and $\Omega(\mathcal{M}) := \left\{\omega \in \Sigma : \ \forall s \in \mathcal{S},\ \sum_a \omega_{sa} = \sum_{s',a'} p_{\mathcal{M}}(s|s',a')\omega_{s'a'}\right\}$ in the online setting.

Note the additional constraints appearing in the definition of $\Omega(\mathcal{M})$ in the online setting, which express that the learning problem is harder. The optimization problem in (1) is difficult to solve, notably due to the non-convex structure of the set of alternative MDPs. When an optimization oracle solving this problem is available, we designed algorithms that match the lower bound above asymptotically. Otherwise, the upper bounds on the sample complexity of our algorithms still feature instance-specific quantities.

Problem-dependent quantities. The sub-optimality gap of action a in state s is defined as $\Delta_{sa} = V^\star_{\mathcal{M}}(s) - Q^\star_{\mathcal{M}}(s,a)$. Let $\Delta_{\min} = \min_{s,a\neq\pi^\star(s)} \Delta_{sa}$ be the minimum gap and $\mathrm{sp}(V^\star_{\mathcal{M}}) = \max_{s,s'} |V^\star_{\mathcal{M}}(s) - V^\star_{\mathcal{M}}(s')|$ be the span of $V^\star_{\mathcal{M}}$. Finally, we introduce the variance of the next-state value function in (s,a) as $\mathrm{Var}_{p(s,a)}[V^\star_{\mathcal{M}}] = \mathrm{Var}_{s'\sim p_{\mathcal{M}}(.|s,a)}[V^\star_{\mathcal{M}}(s')]$.

Theorem 2. *([MP21, AGP21]) There exists δ-PAC algorithms both in the generative and online settings such that their sample complexity satisfies:*
If an oracle solving (1) is available, $\limsup_{\delta\to 0} \frac{\mathbb{E}_{\mathcal{M}}[\tau_\delta]}{\log(1/\delta)} \leq 2T^\star(\mathcal{M})$.
Otherwise:

$$\limsup_{\delta\to 0} \frac{\mathbb{E}_{\mathcal{M}}[\tau_\delta]}{\log(1/\delta)} \leq 2 \inf_{\omega\in\Omega(\mathcal{M})} \max_{(s,a):a\neq\pi^\star(s)} \frac{H_{sa}}{\omega_{sa}} + \frac{H^\star}{S \min_s \omega_{s,\pi^\star(s)}},$$

where

$$H_{sa} = \frac{2}{\Delta_{sa}^2} + \max\left(\frac{16\mathrm{Var}_{p(s,a)}[V^\star_{\mathcal{M}}]}{\Delta_{sa}^2}, \frac{6\,\mathrm{sp}(V^\star_{\mathcal{M}})^{4/3}}{\Delta_{sa}^{4/3}}\right),$$
$$H^\star = \frac{2S}{[\Delta_{\min}(1-\gamma)]^2} + \mathcal{O}\left(\frac{S}{\Delta_{\min}^2(1-\gamma)^3}\right).$$

In our presentation, we will explain how to derive the lower bounds as well as the algorithmic ideas behind the design of BPI algorithms with such instance-specific guarantees.

References

[AGP21] Aymen Al Marjani, Aurélien Garivier, and Alexandre Proutiere. Navigating to the Best Policy in Markov Decision Processes. *arXiv e-prints*, page arXiv:2106.02847, June 2021.

[AHKS20] Alekh Agarwal, Mikael Henaff, Sham Kakade, and Wen Sun. Pc-pg: Policy cover directed exploration for provable policy gradient learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 13399–13412. Curran Associates, Inc., 2020.

[Fie94] C. Fiechter. Efficient reinforcement learning. In *COLT '94*, 1994.

[FYAY21] Fei Feng, W. Yin, Alekh Agarwal, and Lin F. Yang. Provably correct optimization and exploration with non-linear policies. *ArXiv*, abs/2103.11559, 2021.

[KMDD+21] Emilie Kaufmann, Pierre Ménard, Omar Darwiche Domingues, Anders Jonsson, Edouard Leurent, and Michal Valko. Adaptive reward-free exploration. In Vitaly Feldman, Katrina Ligett, and Sivan Sabato, editors, *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, volume 132 of *Proceedings of Machine Learning Research*, pages 865–891. PMLR, 16–19 Mar 2021.

[LWC+20] Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, and Yuxin Chen. Sample complexity of asynchronous q-learning: Sharper analysis and variance reduction, 2020.

[MDJ+20] Pierre M'enard, O. D. Domingues, Anders Jonsson, E. Kaufmann, Edouard Leurent, and Michal Valko. Fast active learning for pure exploration in reinforcement learning. *ArXiv*, abs/2007.13442, 2020.

[MP21] Aymen Al Marjani and Alexandre Proutiere. Adaptive sampling for best policy identification in markov decision processes. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 7459–7468. PMLR, 18–24 Jul 2021.

[WSJ21] Andrew Wagenmaker, Max Simchowitz, and Kevin Jamieson. Beyond No Regret: Instance-Dependent PAC Reinforcement Learning. *arXiv e-prints*, page arXiv:2108.02717, August 2021.

[ZCA21] Andrea Zanette, Ching-An Cheng, and Alekh Agarwal. Cautiously optimistic policy optimization and exploration with linear function approximation. *ArXiv*, abs/2103.12923, 2021.

[ZKB19] Andrea Zanette, Mykel J Kochenderfer, and Emma Brunskill. Almost horizon-free structure-aware best policy identification with a generative model. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 5625–5634. Curran Associates, Inc., 2019.

Bayesian Actor for Deep Reinforcement Learning

Léo Grill [1,2] & Yousri Slaoui [1] & David Nortershauser [2] & Stéphane Le Masson [2]

[1] *Laboratoire de Mathématiques et Applications, Université de Poitiers, 11 Boulevard Marie et Pierre Curie, Bâtiment H3 - TSA 61125, 86073 POITIERS CEDEX 9.*
leo.grill@math.univ-poitiers.fr, Yousri.Slaoui@math.univ-poitiers.fr
[2] *Orange, Orange Labs, 2 av Pierre Marzin, 22300 Lannion, France*
david.nortershauser@orange.com, stephane.lemasson@orange.com

Résumé. L'apprentissage renforcé s'est beaucoup développé, et trouve de nombreuses applications. En parallèle les méthodes bayésiennes sont de plus en plus étudiées, notamment pour l'apprentissage profond. Ces méthodes permettent de palier à des problèmes tels que le surapprentissage et la transparence des modèles. Dans ce papier nous étudions comment un acteur bayésien peut s'appliquer aux méthodes d'acteur-critique de l'apprentissage par renforcement profond. Une étude numérique est présentée, elle montre l'intérêt de l'utilisation de ces méthodes dans l'apprentissage renforcé profond.

Mots-clés. Apprentissage renforcé, Apprentissage profond, Apprentissage bayésien, Méthode acteur-critique, Recherche de politique.

Abstract. While reinforcement learning finds applications in various domains, the Bayesian methods have special attention especially in deep learning. They can be used to deal with issues such as overfitting or explainable modeling. In this paper, the usage of a Bayesian actor in deep actor-critic methods is developed. The results of simulation are presented to show the interest of using these methods for deep-reinforcement learning.

Keywords. Reinforcement Learning, Deep Learning, Bayesian Learning, Actor Critic Methods, Policy Search.

1 Introduction

Dynamic optimization problems are encountered in various industrial domains. Some physical control applications may benefit from the usage of artificial intelligence and especially deep reinforcement learning. Some recent research [1] has shown impressive results. Each problem can have its own issues, which makes them hard to solve with traditional methods. Reinforcement learning may also face some difficulties to converge with issues such as overoptimism bias. This problem is amplified by traditional learning issues like overfitting [8]. Bayesian methods introduce new ways of computing the weights of neural networks. It tackles the overfitting problem [5] while modeling the uncertainty. It has also interesting properties for sampling from the model. Using Bayesian neural

networks in deep reinforcement learning is promising. Some methods have already been considered especially for the function approximation [2]. On one hand, using distributions over the weights of an actor is a way of considering an ensemble method. On the other hand, it introduces additional knowledge by using prior and posterior distributions. The ensemble aspect is a way of smoothing the learning because the distribution on the actor considers a set of solutions instead of a unique one. Reinforcement learning can benefit from the Bayesian methods developed for supervised deep learning.

2 Preliminaries

2.1 Markov Decision Process and Reinforcement Learning

Actor critic algorithms are used for optimization. They model an actor which defines a policy and a critic which evaluates this policy. As we focus on deep reinforcement learning, those models are based on neural architectures. The problems solved by reinforcement learning are often described by Markov Decision Processes. They consider a state space S, an action space A, a Markovian model with a transition function $T : S \times A \to S$, a policy set $\Pi \subset A^S$. A Reward function $R : S \times A \to \mathbb{R}$ and a discount value γ . We consider $\forall t \in \mathbb{N}, \forall \pi \in \Pi, s_t \in S, a_t \in A, r_t \in \mathbb{R}$ the following system :

$$\begin{cases} s_{t+1} = T(s_t, a_t), \\ a_t = \pi\left(s_t\right), \\ r_t = R\left(s_t, a_t\right). \end{cases}$$

$(s_t, a_t, r_t)^{\pi}_{t\in\mathbb{N}}$ is the set of trajectories conditioned by π. $(s_t)^{\pi}_{t\in\mathbb{N}}$ is the state trajectory, $(a_t)^{\pi}_{t\in\mathbb{N}}$ is the action trajectory while $(r_t)^{\pi}_{t\in\mathbb{N}}$ is the reward trajectory measuring how well the policy is doing. The Value function, for a suited discount factor $0 < \gamma < 1$ is given for a policy $\pi \in \Pi$ by $V^{\pi} : s_{t_0} \to \mathbb{E}\left[\sum_{t>t_0} \gamma^t r^{\pi}_t\right]$ which is the global reward conditioned by the policy π. The point is to choose the policy to maximize the value function, which is the expectation of the cumulative reward under a policy, starting from the state s_{t_0}. Actor-critic algorithms estimate the long term expected reward with a critic function, called Q function $Q^{\pi} : (s_{t_0}, a_{t_0}) \to \mathbb{E}\left[\sum_{t>t_0} \gamma^{t-t_0} R\left(s^{\pi}_t, a^{\pi}_t\right)\right]$. The goal is to estimate the best action for the environment at every time step. The gradient methods are used to update the parameters of the policy and Q function simultaneously. The Bellman operator provides a way of estimating the Q function at every time step with its recursive formula. For continuous states and actions space, it is well known that both policy and Q-value estimator can be encoded with neural networks.

2.2 Bayesian Deep Learning

Using the bayesian paradigm to learn the neural network parameters is a way of dealing with issues such as overfitting or uncertainty of the model. Moreover, the bayesian paradigm brings some interesting methods such as variational inference, and the ability to sample the parameters which can be advantageous in Reinforcement Learning. When using stochastic neural networks, the reparameterization trick is used to not break the backpropagation. The parameters are still deterministic, the stochasticity comes from supplementary random variables usually drawn from standard normal distributions. For the sake of simplicity, we consider an independent gaussian distribution for each parameter of the network. We denote the learning parameters for the layers' set $L = [\![1; h]\!]$, with h the number of layers: $\Theta = (W^l_\mu, W^l_\rho, B^l_\mu, B^l_\rho)_{l\in L}$. They encode distributions. The ρ parameters can be considered as log-variance of the parameters. This trick is used to avoid getting negative standard deviation during the training. For a layer l, $W^l_\mu = (\mu^l_i)_{i\in[\![1;m\times n]\!]}$ and $W^l_\rho = (\rho^l_i)_{i\in[\![1;m\times n]\!]}$ with n the number of neurons in the layer l and m the number of neurons of the layer $l-1$ or the size of the input if it is the first layer. The same goes for the bias except the dimension is n and not $m \times n$. The random variables are $\epsilon_\theta = (\epsilon^l_W, \epsilon^l_B)_{l\in L}$. When a set of parameters is drawn from the distribution Θ, it is actually the random variables ϵ_θ which are drawn from their distributions. We have $\epsilon^l_W = (\epsilon^{W,l}_i)_{i\in[\![1;m\times n]\!]}$ with $\epsilon^{W,l}_i$ following standard normal law for all $i \in [\![1; m \times n]\!]$. The same goes for the bias except the dimension is n instead of $m \times n$.

When drawing from the random variable, a set of deterministic parameters encoding a neural network is obtained : $\theta = (W^l_\mu, W^l_\rho, \epsilon^l_W, B^l_\mu, B^l_\rho, \epsilon^l_B)_{l\in L}$.

We have the following bayesian neural network's layer:

$$\psi^l_\theta(x) \quad = \quad \phi_l\left(\left(W^l_\mu + \exp(\frac{1}{2}.W^l_\rho) * \epsilon^l_W\right).x \,+\, B^l_\mu + \exp(\frac{1}{2}.B^l_\rho) * \epsilon^l_B\right),$$

Where ϕ_l is the activation function, which is hyperbolic tangent in our experimentation. Moreover, we indicate by $\bigcirc$ the composition operator $\bigcirc_{i\in[\![1;n]\!]} f_i(x) = f_1 \circ f_2 \circ ... \circ f_n(x)$. The whole neural network can be considered as $\psi_\theta(x) = \bigcirc_{l\in L} \psi^l_\theta(x)$.

Stochastic neural networks can be trained with stochastic gradient descent (SGD) algorithms. It is shown in [7] that SGD and its variants such as ADAM [6] can be seen as a variational Expectation-Maximization algorithm, which aims to maximize the joint probability. To learn the parameters of the model the log joint probability given by $\log p(\theta, x) = \log(p(x, \theta)) + \log(p(\theta))$ is maximized. The usual following loss function $\mathcal{L}$ depending on the data $D = (D^k_y, D^k_x)_{0<k<S}$ and the parameters θ is considered:

$$\mathcal{L}(\theta, D) = \frac{1}{S}\sum_{k\in S}\left(-\log(p(D^k_y|D^k_x, \theta)) - \frac{1}{N}\log(p(\theta))\right).$$

With S the number of batches and N the number of elements by batch. These methods can be adapted to reinforcement lerning and are easy to implement.

3 Bayesian Actor-Critic

The Bayesian paradigm has a serious consideration in deep learning developments. With the complexity of continuous action and state spaces, the convergence and error control are necessary to deal with overoptimism bias [3]. Bayesian Q-learning has been developed to stabilize the learning. A lot of work has been made to enforce the Q-estimate, but actors did not get the same attention. Improving actors' performance can be beneficial for the exploration and for the critic's robustness. Considering bayesian actors is close to ensemble methods. During the learning, a set of actors is considered based on the distribution of parameters. The neural network architecture defines a space of policies which is induced by all parameters possibilities for this architecture. To train a bayesian neural network as an actor in a reinforcement learning context, the usual formula given by the stochastic gradient descent algorithm must be adapted. The data is generated through the interaction of the agent with the environment. There is no likelihood for the actor as we are not in a supervised context Instead, it is expected to choose actions that maximize the long-term reward. We assume the distribution given in the Soft Actor Critic (SAC) algorithm [4] fulfills this objective. Intuitively, the probability of getting the action which maximizes the long-term reward is maximized. Like in [4], we get rid of normalization term which does not participate in the gradient. A parameter α is added to scale the prior part in the loss. So the loss of the actor is:

$$J_\pi(\Theta) = \mathbb{E}_{s_t \sim Env, \theta \sim \Theta}[-Q(\psi_\theta(s_t), s_t) - \alpha \log(p(\theta))]$$

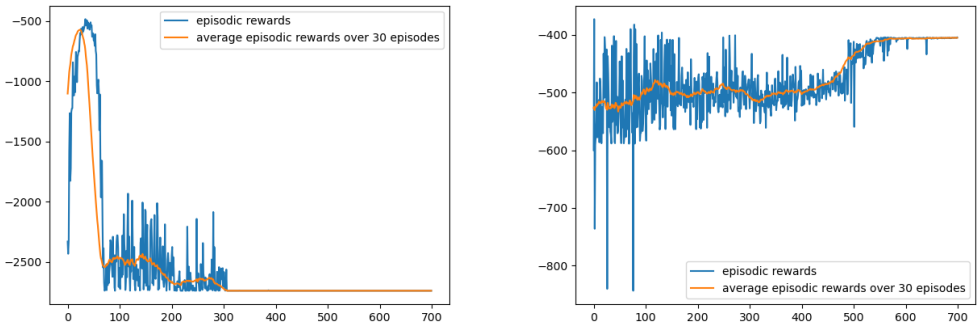
The prior can be determined at the beginning or can be learned as well. In our work, we choose to recycle the posterior as the prior during the learning. Thus, the prior can be seen as a regularization, which prevents the actor from changing too much from an iteration to the next.

Algorithm 1 is made to exploit a time coherent exploration scheme. The noise comes from the certainty of the action network. But the randomness is applied between each episode or roll-out. With the bayesian actor, it is possible to make coherent exploration schemes. For the parameters updates, using ADAM optimizer is a good choice because it is a stochastic gradient descent algorithm and has good performance in various cases.

Algorithm 1 Bayesian Deep Policy Gradient

procedure LEARNING(Θ, ψ_Q)
    Env Initialization
    Θ Initialization
    ψ_Q Initialization
    D = {}
    for number of steps **do**
        for number of samples **do**
            Draw $\theta \sim \Theta$
            s = Env.reset()
            for number of iterations **do**
                $a \leftarrow \psi_\theta(s)$
                $r, s' \leftarrow Env.step(a)$
                append s, a, s', r to D
                $s \leftarrow s'$
        for number of samples **do**
            Draw $\theta \sim \Theta$
            draw from D to update Θ with stochastic gradient descent algorithm
            draw from D to update ψ_Q with backpropagation algorithm

4 Results and Discussion



(a) Deterministic Actor-Critic

(b) Bayesian Actor-Critic

Figure 1: The reward per episode for the Bayesian Actor-Critic model is converging to high rewards policy while the Deterministic model struggles to obtain good results.

The model was trained on a physical use case. The environment models a room with a heated load and ventilation. The variables describing the environment are the inside temperature and the outside temperature. The action is the ventilation usage. The reward is based on the capacity to respect some temperature condition while minimizing the ventilation usage. The goal is to create a policy that is able to control the air conditioning system based on experience and data. With the results of our experiments, we show that the Bayesian actor solves this task with more stability and efficiency than classical deterministic actor-critic models. Using Bayesian Neural Network as actor in

actor-critic algorithms can benefit from a coherent exploration, which is useful in specific environments. It also brings knowledge about uncertainty of the network. The probability associated to each parameter and to the network can be used to watch the convergence of the actor. It highlights some regularization for the network. Some knowledge can be inserted for the exploration thanks to the prior. The posterior and the network architecture can be adjusted to the specificity of the optimal policies. However, the Bayesian Neural Network needs more computations for the training and evaluation as most of the framework is based on sampling methods. The parallelization and computation power keep increasing which makes those algorithms practicable. Bayesian learning should be used to solve problems with unstable environment and complex learning. It is close to ensemble methods with a single Q-value estimator. This method shows empirically a positive impact on exploration, initialization and critic robustness.

References

[1] Degrave and al. Magnetic control of tokamak plasmas through deep reinforcement learning. *Nature*, 602(7897):414–419, February 2022.

[2] Carlo D'Eramo, Andrea Cini, Alessandro Nuara, Matteo Pirotta, Cesare Alippi, Jan Peters, and Marcello Restelli. Gaussian Approximation for Bias Reduction in Q-Learning. *Journal of Machine Learning Research*, 22(277):1–51, 2021.

[3] Scott Fujimoto, Herke van Hoof, and David Meger. Addressing Function Approximation Error in Actor-Critic Methods. *arXiv:1802.09477 [cs, stat]*, October 2018. arXiv: 1802.09477.

[4] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor. *arXiv:1801.01290 [cs, stat]*, August 2018. arXiv: 1801.01290.

[5] Laurent Valentin Jospin, Wray Buntine, Farid Boussaid, Hamid Laga, and Mohammed Bennamoun. Hands-on Bayesian Neural Networks – a Tutorial for Deep Learning Users. *arXiv:2007.06823 [cs, stat]*, September 2021. arXiv: 2007.06823.

[6] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]*, January 2017. arXiv: 1412.6980.

[7] Stephan Mandt, Matthew D. Hoffman, and David M. Blei. Stochastic Gradient Descent as Approximate Bayesian Inference. *arXiv:1704.04289 [cs, stat]*, January 2018. arXiv: 1704.04289.

[8] Chiyuan Zhang, Oriol Vinyals, Remi Munos, and Samy Bengio. A Study on Overfitting in Deep Reinforcement Learning. *arXiv:1804.06893 [cs, stat]*, April 2018. arXiv: 1804.06893.

Annulé

Apprentissage robuste pour la résolution d'EDPs Applications aux marchés de l'énergie

Carl REMLINGER [1,2] & Joseph MIKAEL [2]

[1] *Université Gustave Eiffel,* [2] *EDF Lab*

Résumé. Un modèle résolvant une famille d'équations aux dérivées partielles (EDPs) avec un seul apprentissage est proposé. Pour cela, nous considérons des réseaux d'opérateurs profonds afin d'approcher avec précision des opérateurs continus non linéaires. Nous voulons résoudre des EDPs lorsque l'environnement n'est pas stationnaire ou pour plusieurs conditions initiales à la fois. Notre modèle apprend la solution générale associée à chaque fonction paramètre simultanément. Mais, en fin de compte, nous voulons généraliser la résolution avec des modèles sous-jacents ou des conditions qui n'étaient pas présents lors de l'entraînement. Nous confirmons l'efficacité de la méthode avec plusieurs problèmes de gestion des risques. Dans ce cas, re-calibrer un modèle de facteurs de risque ou ré-entraîner un modèle pour la résolution chaque fois que les conditions de marché changent est coûteux et insatisfaisant. Nous évaluons notre DeepOHedger pour la couverture d'options, incluant des modèles de volatilité locale et des options spread impliquées dans les marchés de l'énergie.

Mots-clés. équation aux dérivées partielles, réseaux de neurones pour opérateurs, couverture de risques, marchés de l'énergie

Abstract. A model solving a family of partial differential equations (PDEs) with a single learning is proposed. For this purpose, we consider deep operator networks to accurately approximate continuous nonlinear operators. We want to solve PDEs when the environment is not stationary or for several initial conditions at the same time. Our model learns the general solution associated with each parameter function simultaneously. But, ultimately, we want to generalize the solution with underlying models or initial conditions that were not present during the training. We confirm the effectiveness of the method with several risk management problems. In this case, re-calibrating a risk factor model or re-training a model for resolution every time market conditions change is costly and unsatisfactory. We evaluate our DeepOHedger on option pricing tasks, including local volatility models and option spread involved in energy markets.

Keywords. partial differential equation, deep operator network, risk hedging, energy market

1 Introduction

Le contrôle stochastique reposant sur des simulations Monte Carlo nécessite des hypothèses de modélisation. En pratique, il existe souvent une incertitude sur le modèle

et l'environnement n'est pas stationnaire. Les marchés de l'énergie sont en constante évolution et les modèles peuvent donc rapidement être invalidés. Notamment, les prix de l'électricité dépendent fortement de phénomènes physiques (s'il fait froid, le chauffage consomme plus et donc les prix montent; s'il y a beaucoup de vent, les éoliennes produisent en quantité et les prix chutent).

Les changements de régime imposés par des conditions météorologiques inhabituelles (vagues de grands froids, sécheresses) nécessitent l'utilisation de modèles plus flexibles et plus robustes pour la gestion des risques dans le secteur de l'énergie. En outre, l'utilisation de modèles mal calibrés peut conduire à prendre des décisions sous-optimales. Le calibrage des modèles stochastiques doit être effectué avec prudence et, idéalement, en continu. Cependant, en raison du coût de calcul associé et du besoin de stabilité, ce n'est pas toujours le cas. La conception de stratégies de couverture non neutres vis-à-vis du risque reste donc particulièrement difficile.

Une méthode numérique populaire pour la couverture du risque repose sur la résolution d'EDP, mais elle est difficile en haute dimension et manque de flexibilité. Pour répondre à ce problème, de nouvelles méthodes ont été développées en utilisant des réseaux neuronaux profonds pour approcher la solution (voir Germain et al. (2021) pour revue complète). Cependant, la plupart du temps, ces méthodes sont entraînées pour un modèle de facteur de risque donné, ce qui peut conduire à ajouter des erreurs d'approximation lorsque le modèle de risque n'est pas précis.

Des propositions récentes se concentrent à résoudre des problèmes d'EDP paramétriques en utilisant des réseaux de neurones. En physique, Khoo et al. (2021) résout des EDPs déterministes à partir de quantités physiques initialement choisies aléatoirement. Glau et al. (2020) présente une méthode paramétrique générale capable de résoudre une famille d'EDP avec un seul réseau de neurones. Pour approcher la solution, le réseau est conditionné avec des valeurs constantes décrivant le modèle de facteur de risque ou l'option. La fonction de perte repose sur la formulation des moindres carrés d'une EDP comme dans Deep Galerkin Sirignano et al. (2018). Ce modèle permet des applications larges et est testé sur la tarification d'options multi-actifs.

L'approche est intéressante et les résultats prometteurs, mais ces méthodes pourraient aller plus loin et exploiter davantage le potentiel de l'apprentissage profond. Nous proposons d'apprendre des opérateurs pour un ensemble d'EDPs, au lieu de les conditionner la résolution par des paramètres à valeur réelle. Les solutions sont approchées par un seul réseau de neurones dans une approche non supervisée. Afin de garantir de capturer correctement la correspondance entre les fonctions et la sortie de l'état à un instant donné, nous considérons des réseaux d'opérateurs profonds (DeepONets) Lu, Jin, and Karniadakis (2019). L'application numérique se concentre sur la couverture du risque. Le Deep Operator Hedger (DeepOHedger) introduit est un solveur universel pour un ensemble de modèles de facteurs de risque ou d'objectifs donnés.

Contributions

- Nous introduisons un modèle basé sur l'approximation d'opérateurs continus capable de résoudre des EDPs non linéaires.
- Le modèle résout une famille d'EDPs à l'aide d'un seul entraînement.
- Une application numérique pour la couverture des risques utilisant des réseaux neuronaux pour opérateurs est proposée. Notre DeepOHedger rivalise avec les méthodes classiques d'apprentissage automatique sur divers modèles sous-jacents.

2 Cadre général du contrôle stochastique

Soit T un horizon fini. Soit $X = (X_t)_{t\in[0,T]}$ une dynamique de diffusion contrôlée en temps continu dont les réalisations appartiennent à $\mathbb{R}^d$

$$dX_t = \mu(t, X_t, \alpha_t)dt + \sigma(t, X_t, \alpha_t)dW_t, \tag{1}$$

où W est un mouvement Brownien standard de dimension d sur un espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$ équipé d'une filtration $(\mathcal{F}_t)_{t\in[0,T]}$ représentant les informations disponibles à l'instant t, et X_0 une variable aléatoire $\mathcal{F}_0$-mesurable d'une valeur de $\mathbb{R}^d$. La fonction $\mu : [0,T] \times \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ décrit le drift et $\sigma : [0,T] \times \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{M}^d$ la volatilité (dans l'ensemble des matrice semi-définies positives). Les deux fonctions (déterministes) satisfont les conditions habituelles de Lipschitz Ikeda et al. (2014) assurant l'existence et l'unicité de la solution de (1). Le contrôle $\alpha = (\alpha_t)_{t\in[0,T]}$ est un processus $\mathcal{F}_t$-adapté sur l'ensemble des contrôles admissibles $\mathcal{A}$. Nous cherchons une stratégie optimale, qui minimise le coût J du processus α :

$$J(\alpha) = \mathbb{E}\left[\int_0^T f(t, X_t, \alpha_t)dt + g(X_T))\right],$$

où f est une fonction génératrice non linéaire définie sur $[0,T] \times \mathbb{R}^d \times \mathbb{R}^d$, et g est une fonction terminale définie sur $\mathbb{R}^d$, appelée *gain*. Dans ce cas, l'équation de Bellman de programmation dynamique conduit à une EDP de la forme

$$\begin{aligned} \partial_t u + \mu D_x u + \frac{1}{2}\mathrm{Tr}(\sigma\sigma^T D_x^2 u) = f \quad &\text{sur } [0,T) \times \mathbb{R}^d \\ u(T,.) = g \quad &\text{sur } \mathbb{R}^d \end{aligned} \tag{2}$$

Dans le cas de la couverture de risque, $f = 0$.

3 Résolution robuste par opérateur

Nous proposons de résoudre une famille d'EDPs à l'aide d'un seul et même modèle et d'un unique entraînement. Nous utilisons un modèle qui repose sur l'apprentissage

d'opérateurs, c'est-à-dire qui associent un ensemble de fonctions paramètres à un autre ensemble de fonctions. Les fonctions paramètres peuvent décrire le modèle sous-jacent X (en particulier sa fonction de volatilité $\sigma(.,.)$), des caractéristiques d'une option ou des contraintes. Afin de mieux généraliser l'apprentissage, nous proposons d'apprendre à la fois la solution de l'EDP et sa fonction paramètre associée.

À cette fin, nous considérons une architecture de réseau de neurones spécifique, les réseaux d'opérateurs profonds appelés DeepONets Lu, Jin, Pang, et al. (2021). L'architecture du modèle consiste en deux sous-réseaux, un qui apprend une fonction d'entrée approchée sur un nombre fixé de points et un autre pour représenter l'état du processus à un instant donné.

3.1 Apprendre avec une seule fonction

La fonction de paramètre est désignée par $u \in \mathcal{U}$, et $s \in \mathcal{S}$ est la solution inconnue correspondante de l'EDP. Le problème (2) est résolu pour un u donné par les méthodes traditionnelles Germain et al. (2021). La question est de savoir si l'on peut trouver une solution pour tout $u \in \mathcal{U}$. Plus précisément, nous supposons que, pour tout $u \in \mathcal{U}$, il existe une solution unique $s = s(u) \in \mathcal{S}$ (sous réserve de conditions initiales et limites appropriées). Nous pouvons alors définir l'opérateur solution $G : \mathcal{U} \to \mathcal{S}$ comme suit

$$G(u) = s(u)$$

Le DeepONet $h_\theta = (h_0^\theta, h_1^\theta, h_2^\theta)$ de paramètre θ est composé de trois éléments $h_0^\theta \in \mathbb{R}$, $h_1^\theta : \mathbb{R}^q \to \mathbb{R}^K$, et $h_2^\theta : \mathbb{R}^{d+1} \to \mathbb{R}^K$, pour $K \in \mathbb{N}^*$. Le modèle h_θ prend en entrée de manière indépendante une approximation de la fonction u et un vecteur $[t, X_t]$ à valeur dans $\mathbb{R}^{d+1}$ afin de renvoyer les contrôles correspondants

$$\alpha_\theta(u)(t, X_t) = \sum_{k=1}^{K} h_{1,k}^\theta(u(x_1), \ldots, u(x_q)).h_{2,k}^\theta(t, X_t) + h_0^\theta. \tag{3}$$

L'équation (3) se résume donc à un produit scalaire. L'approximation de la fonction u est déterminante pour l'identification de la dynamique sous-jacente ou de l'option. Ainsi, les capteurs $\{x_1, \ldots, x_q\}$ doivent être choisis avec précaution. Nous considérons une grille régulière de $q \in \mathbb{N}$ points de $\mathbb{R}^d$.

3.2 Généralisation avec une famille de fonctions

La méthode introduite nécessite de fournir une approximation de la fonction u. En donnant une fonction trompeuse, nous pourrions ajouter une erreur de modèle à l'erreur d'approximation. Comme notre premier objectif est de pouvoir résoudre de manière robuste, nous entraînons notre modèle non pas sur une fonction u mais sur une famille de fonctions. Le réseau d'opérateurs apprend différentes fonctions u, ainsi nous tirons

réellement avantage de l'utilisation des réseaux de neurones, en nous attaquant à un problème de haute dimension et à des fonctions éventuellement difficiles à estimer. Dans l'application numérique, nous proposons d'apprendre un ensemble de fonctions de volatilité (linéaires ou non) et des estimations basées sur des données. Lorsque l'on est incertain sur la structure du marché, on peut fournir une approximation des fonctions de volatilité, notre modèle doit être capable de résoudre correctement l'EDP, que le réseau ait été entraîné avec la fonction ou non.

4 Application à la couverture des risques

Nous confirmons l'efficacité de notre méthode sur un problème de couverture de risques d'options. Notre modèle, dénommé DeepOHedger, est comparé à d'autres approches reposant sur l'apprentissage profond appliquées à divers modèles de risques.

Une approche globale Une stratégie de couverture peut être réduite à un problème de contrôle stochastique, en achetant ou en vendant une certaine quantité α d'un actif à chaque date. En discrétisant sur une grille temporelle régulière $\mathcal{T} = \{t_0 = 0, \ldots, t_N = T\}$ pour $N \in \mathbb{N}$, on défini un portefeuille auto-financé Π_{t_i} de valeur terminale comme :

$$\Pi_{T,\theta;u} = p_\theta + \sum_{i=0}^{N-1} \alpha_\theta(t_i, X_{t_i}; u)(X_{t_{i+1}} - X_{t_i}).$$

On se donne la possibilité de rééquilibrer le portefeuille chaque jour. Il n'y a pas de contrôle à la dernière date. L'approche globale de résolution de (2) conduit au problème d'optimisation suivant :

$$\min_\theta \mathbb{E}[(\Pi_{T,\theta;u} - g(X_T; u))^2].$$

Nous évaluons notre DeepOHedger sur une tâche de tarification d'options classique, incluant les modèles Black-Scholes et de volatilité locale CEV.

La figure 1 illustre les erreurs de DeepOHedger en fonction de chaque modèle de prix. DeepOHedger généralise bien pour des paramètres inconnus, comme l'indiquent les valeurs à $\sigma \in \{.25, .45, .65\}$. L'erreur de réplication augmente lorsque la volatilité est élevée, ce qui est un comportement attendu : il est plus difficile de répliquer une option lorsque le marché est instable.

Nous proposons aussi de couvrir une option spread. Un fournisseur d'électricité peut être amené à acheter du gaz pour produire de l'électricité. Une centrale électrique peut alors être modélisée comme une option spread, où l'électricité est vendue à un prix donné X_t^e auquel il faut retrancher le prix du gaz X_t^g et les taxes sur le CO2 X_t^{co2}, en fonction des taux de conversion β, γ. Le gain d'une telle option est alors:

$$g(X_T^e, X_T^g, X_T^{co2}) = (X_T^e - \beta X_T^g - \gamma X_T^{co2} - K)_+$$

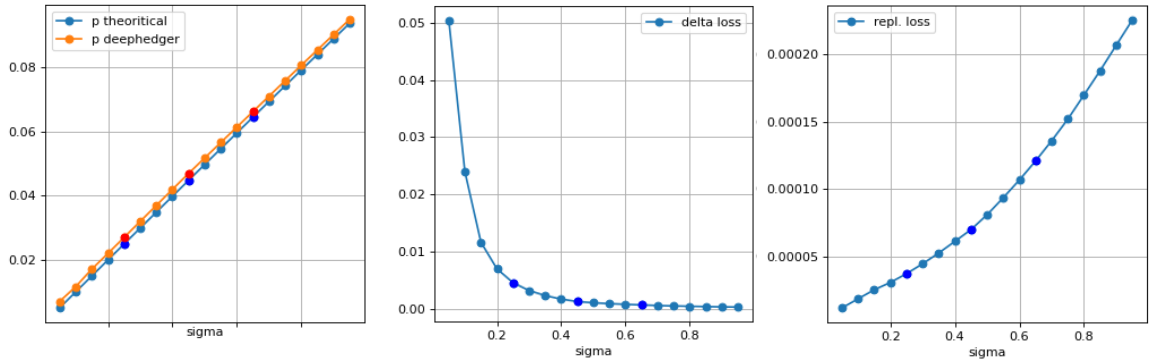


Figure 1: Prime, erreurs sur les contrôles et erreur de réplication selon les valeur de la volatilité de modèles Black-Scholes.

où K indique le strike et $\beta, \gamma \in \mathbb{R}$ les taux de conversion.

Afin de construire une stratégie de couverture robuste, notre modèle est entraîné avec plusieurs modèles et pour différentes corrélations. Notre approche assure de meilleures performances que les autres approches (en particulier celle conditionnée avec des valeurs paramètres au lieu d'apprendre des opérateurs), et est notamment capable de mieux généraliser pour des paramètres inconnus d'un modèle d'aléas donné.

Bibliographie

Germain, M. et al. (2021). Neural networks-based algorithms for stochastic control and PDEs in finance. *arXiv preprint arXiv:2101.08068*.

Glau, K. et al. (2020). The deep parametric PDE method: application to option pricing. *arXiv preprint arXiv:2012.06211*.

Ikeda, N. et al. (2014). *Stochastic differential equations and diffusion processes*. Elsevier.

Khoo, Y. et al. (2021). Solving parametric PDE problems with artificial neural networks. *European Journal of Applied Mathematics* 32.3, pp. 421–435.

Lu, L., P. Jin, and G. E. Karniadakis (2019). Deeponet: Learning nonlinear operators for identifying differential equations based on the universal approximation theorem of operators. *arXiv preprint arXiv:1910.03193*.

Lu, L., P. Jin, G. Pang, et al. (2021). Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators. *Nature Machine Intelligence* 3.3, pp. 218–229.

Sirignano, J. et al. (2018). DGM: A deep learning algorithm for solving partial differential equations. *Journal of computational physics* 375, pp. 1339–1364.

Statistique mathématique

Estimation confidentielle de quantiles en présence d'atomes

Clément Lalanne [1], Clément Gastaud [2], Nicolas Grislain [3], Aurélien Garivier [4] & Rémi Gribonval [5]

[1] *LIP, École Normale Supérieure de Lyon, Lyon, France, clement.lalanne@ens-lyon.fr*
[2] *Sarus Technologies, Paris, France, cg@sarus.tech*
[3] *Sarus Technologies, Paris, France, ng@sarus.tech*
[4] *UMPA, École Normale Supérieure de Lyon, Lyon, France, aurelien.garivier@ens-lyon.fr*
[5] *LIP, École Normale Supérieure de Lyon & INRIA, Lyon, France, remi.gribonval@inria.fr*

Résumé. Nous abordons l'estimation sous confidentialité différentielle de plusieurs quantiles (MQ) d'un ensemble, un élément clé de l'analyse moderne des données. Nous appliquons le mécanisme récent de sensibilité inverse (SI) non lissé à ce problème spécifique et établissons que la méthode qui en résulte est étroitement liée à l'état de l'art actuel, l'algorithme JointExp, partageant en particulier la même complexité algorithmique et une efficacité similaire. Cependant, nous démontrons à la fois théoriquement et empiriquement que JointExp (non lissé) souffre d'un manque important de performance dans le cas de distributions à pics, avec un impact potentiellement catastrophique en présence d'atomes. Alors que sa version lissée permettrait de tirer parti des garanties de performance de IS, elle reste un défi ouvert à mettre en œuvre. Pour résoudre ce problème, nous proposons une méthode simple et numériquement efficace appelée Heuristically Smoothed JointExp (HSJointExp), qui est dotée de garanties de performance pour une large classe de distributions et permet d'obtenir des résultats significativement meilleurs sur des ensembles de données problématiques.

Mots-clés. Confidentialité différentielle, Quantiles, Sensibilité Inverse

Abstract. We address the differentially private estimation of multiple quantiles (MQ) of a dataset, a key building block in modern data analysis. We apply the recent non-smoothed Inverse Sensitivity (IS) mechanism to this specific problem and establish that the resulting method is closely related to the current state-of-the-art, the JointExp algorithm, sharing in particular the same computational complexity and a similar efficiency. However, we demonstrate both theoretically and empirically that (non-smoothed) JointExp suffers from an important lack of performance in the case of peaked distributions, with a potentially catastrophic impact in the presence of atoms. While its smoothed version would allow to leverage the performance guarantees of IS, it remains an open challenge to implement. As a proxy to fix the problem we propose a simple and numerically efficient method called Heuristically Smoothed JointExp (HSJointExp), which is endowed with

performance guarantees for a broad class of distributions and achieves results that are orders of magnitude better on problematic datasets.

Keywords. Differential privacy, Quantiles, Inverse Sensitivity

1 Confidentialité différentielle et estimation de quantiles

Alors que de plus en plus de données sont collectées sur les individus et que les techniques des sciences des données deviennent de plus en plus puissantes, les menaces liées à la vie privée se sont multipliées et de sérieuses préoccupations ont émergé. Dans ce contexte, la *confidentialité différentielle* (Dwork et Roth (2014)) est devenue le standard en matière de protection de la vie privée. En introduisant de l'aléa à un niveau calibré à la *sensibilité* d'une requête, elle permet d'inférer des statistiques globales sur un ensemble de données tout en limitant l'influence de chaque échantillon et en garantissant que la présence ou l'absence d'un individu ne peut pas être déduite du résultat. Au cours de la dernière décennie, la recherche a apporté des primitives confidentielles de base et des théorèmes de composition. Ils ont ouvert la voie à de nombreuses applications dans l'analyse de données, des statistiques de base aux algorithmes avancés d'intelligence artificielle. Notamment, la confidentialité différentielle est maintenant utilisée en production par le bureau du recensement des États-Unis, Google, Apple et Microsoft, entre autres.

Nous considérons un ensemble de données $\mathbf{X} = (X_1, \dots, X_n) \in \mathfrak{X}^n$, où $\mathfrak{X}$ est l'ensemble dans lequel vivent les informations de chaque utilisateur et où $n \geq 1$ est le nombre de données collectées. La *distance de Hamming* $d(\mathbf{X}, \mathbf{Y})$ entre deux ensembles de données $\mathbf{X}, \mathbf{Y} \in \mathfrak{X}^n$ est défini comme le nombre minimal de changements (c.a.d., de substitutions de composantes) nécessaires pour transformer $\mathbf{X}$ en une permutation de $\mathbf{Y}$. Donc, $d(\mathbf{X}, \mathbf{Y}) = 0$ si et seulement si il existe une permutation σ de $\{1, \dots, n\}$ telle que $\forall i \in \{1, \dots, n\}, X_i = Y_{\sigma(i)}$. Nous disons que $\mathbf{X}$ et $\mathbf{Y}$ sont *voisins* (noté $\mathbf{X} \sim \mathbf{Y}$) si $d(\mathbf{X}, \mathbf{Y}) = 1$, c'est à dire si il existe deux permutations σ_1, σ_2 telles que $\forall i \in \{1, \dots, n-1\}, X_{\sigma_1(i)} = Y_{\sigma_2(i)}$. Notons que $d(\mathbf{X}, \mathbf{Y})$ est la longueur minimale d'un chemin liant $\mathbf{X}$ à $\mathbf{Y}$ en utilisant $\sim$. Pour un budget de confidentialité $\epsilon > 0$, un algorithme aléatoire $\mathcal{A} : \mathfrak{X}^n \to O$ est dit *ϵ-confidentiel différentiellement* (ϵ-DP) si pour toute paire d'ensembles de données $\mathbf{X}, \mathbf{Y} \in \mathfrak{X}^n$ et tout ensemble mesurable $S \subseteq O$,

$$\mathbf{X} \sim \mathbf{Y} \Rightarrow \mathbb{P}\big(\mathcal{A}(\mathbf{X}) \in S\big) \leq e^{\epsilon} \times \mathbb{P}\big(\mathcal{A}(\mathbf{Y}) \in S\big) .$$

Cette définition permet de borner la puissance de tout test visant à distinguer $\mathbf{X}$ de $\mathbf{Y}$. Une méthode classique pour obtenir un algorithme qui est ϵ-DP est le *Mécanisme Exponentiel* (McSherry et Talwar (2007)). Pour une fonction d'utilité $u : \mathfrak{X}^n \times O \to \mathbb{R}$ qui mesure la pertinence $u(\mathbf{X}, o)$ d'un résultat candidat o pour l'ensemble de données $\mathbf{X}$, le

mécanisme exponentiel $\mathcal{E}_u^{(\alpha)}$ défini par u et par un paramètre $\alpha > 0$ retourne une variable aléatoire sur O avec une densité proportionnelle à $e^{u(\mathbf{X},o)/\alpha}$ relativement à une mesure de référence sur O. Par exemple, quand O est dénombrable et relativement à la mesure de comptage,

$$\mathbb{P}\left(\mathcal{E}_u^{(\alpha)}(\mathbf{X}) = o\right) = \frac{e^{u(\mathbf{X},o)/\alpha}}{\sum_{o' \in O} e^{u(\mathbf{X},o')/\alpha}} \quad \forall o \in O \ .$$

En définissant la *sensibilité* d'une fonction d'utilité u comme

$$\Delta u := \sup_{o \in O, \mathbf{X}, \mathbf{Y} \in \mathfrak{X}^n : \mathbf{X} \sim \mathbf{Y}} |u(\mathbf{X}, o) - u(\mathbf{Y}, o)| \ ,$$

$\mathcal{E}_u^{(\alpha)}$ est ϵ-DP dès que $\alpha \geq \frac{2\Delta u}{\epsilon}$.

Fixons les notations pour l'estimation de plusieurs quantiles. Étant donné $a < b \in \mathbb{R}$, nous avons $\mathfrak{X} = [a, b]$ et $O = [a, b]^{m\nearrow}$ l'ensemble des vecteurs croissants de m points de $[a, b]$ qui représentent m quantiles. L'hypothèse que les données sont bornées est nécessaire à notre analyse et est raisonnable pour beaucoup d'applications. Considérons un vecteur de probabilités $\mathbf{p} = (p_1, \ldots, p_m) \in (0, 1)^{m\nearrow}$, l'objectif est d'estimer les quantiles empiriques associés au vecteur $\mathbf{p}$ et définis comme

$$\begin{array}{rccc} \mathcal{Q} & : & \mathfrak{X}^n & \to & O \\ & & \mathbf{X} & \mapsto & (X_{(\lceil np_1 \rceil)}, \ldots, X_{(\lceil np_m \rceil)}) \end{array}$$

où $X_{(i)}$ est la i-ème statistique d'ordre de $X_1, \ldots, X_n$. Pour tout $\mathbf{X} \in \mathfrak{X}^n$, $\mathbf{q} \in O$ et $\mathbf{p} \in (0, 1)^m$, nous adoptons les conventions $X_{i \leq 0} = q_{i \leq 0} = a$, $X_{i \geq n+1} = q_{i \geq m+1} = b$, $p_{i \leq 0} = 0$ et $p_{i \geq m+1} = 1$. Enfin, pour abréger les notations, les vecteurs seront interprétés comme les ensembles contenant leurs composantes si nécessaire.

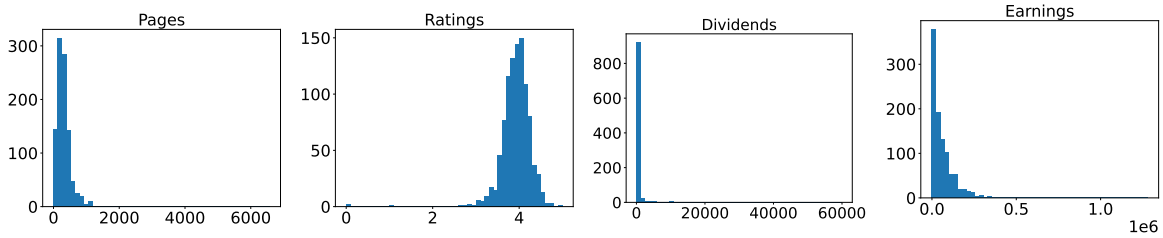
Notre but est d'estimer un ou plusieurs quantiles sous contrainte de confidentialité. Bien qu'il soit possible d'appliquer des théorèmes de composition à un algorithme d'estimation confidentielle d'un unique quantile, l'état de l'art récent nous montre, en explicitant un nouvel algorithme JointExp (Gillenwater et Joseph et Kulesza (2021)), qu'il est plus judicieux d'estimer tous les quantiles jointement. Cet algorithme est une instanciation du mécanisme exponentiel $\mathcal{E}_{u_{\mathrm{JE}}}^{(2/\epsilon)}(\mathbf{X})$ avec pour fonction d'utilité

$$-u_{\mathrm{JE}}(\mathbf{X}, \mathbf{q}) := \frac{1}{2} \sum_{i=1}^{m+1} \left| \delta^{\mathrm{JE}}(i, \mathbf{X}, \mathbf{q}) \right| \ ,$$

(dont la sensibilité est 1) où

$$\delta^{\mathrm{JE}}(i, \mathbf{X}, \mathbf{q}) := n(p_i - p_{i-1}) - \#\left(\mathbf{X} \cap (q_{i-1}, q_i]\right) \ .$$

Un résultat potentiel est pénalisé lorsque le nombre de points vivant dans chaque intervalle interquantile ($\#\left(\mathbf{X} \cap (q_{i-1}, q_i]\right)$) diverge du nombre attendu ($n(p_i - p_{i-1})$).



$n = 1000$ points représentés en histogrammes de 50 divisions.

Figure 1: Jeux de données utilisés dans nos expériences

2 Sensibilité inverse pour l'estimation de quantiles

Dans un premier temps, nous étudions l'application d'un mécanisme spécifique (le mécanisme à sensibilité inverse Asi et Duchi (2020)) à l'estimation de plusieurs quantiles. Plus précisément, lorsqu'une fonction $\mathcal{Q} : \mathfrak{X}^n \to O$ est l'objectif de l'estimation, le mécanisme à sensibilité inverse se base sur un mécanisme exponentiel avec pour fonction d'utilité

$$u_{\mathrm{IS}}(\mathbf{X}, o) := -\inf \left\{ d(\mathbf{X}, \mathbf{Y}) \text{ s.t. } \mathbf{Y} \in \mathcal{Q}^{-1}(o) \right\} .$$

Elle a pour sensibilité $\Delta u_{\mathrm{IS}} = 1$. Le mécanisme ϵ-DP qui en résulte $\mathcal{E}_{u_{\mathrm{IS}}}^{(2/\epsilon)}(\mathbf{X})$ a le comportement suivant: Un candidat de retour o a une vraisemblance qui décroit lorsque $\mathcal{Q}^{-1}(o)$ diverge de $\mathbf{X}$ en distance de Hamming. Nous explicitons le théorème suivant:

Théorème 1. *Pour tout* $\mathbf{X} \in \mathfrak{X}^n$ *et* $\mathbf{q} \in ([a, b] \setminus \mathbf{X})^{m\nearrow}$ *sans collision,*

$$\begin{aligned} -u_{IS}(\mathbf{X}, \mathbf{q}) = \frac{1}{2} \sum_{i=2}^{m+1} |\delta(i, \mathbf{X}, \mathbf{q})| + \sum_{i=2}^{m} \mathbb{1}_{\mathbb{R}_+} (\delta(i, \mathbf{X}, \mathbf{q})) \\ + \frac{1}{2} |\delta_{\texttt{closed}}(1, \mathbf{X}, \mathbf{q})| + \mathbb{1}_{\mathbb{R}_+} (\delta_{\texttt{closed}}(1, \mathbf{X}, \mathbf{q})) \end{aligned}$$

avec

$$\begin{aligned} \delta(i, \mathbf{X}, \mathbf{q}) &= \#(\mathbf{X} \cap (q_{i-1}, q_i]) - (\lceil np_i \rceil - \lceil np_{i-1} \rceil) \\ \delta_{\texttt{closed}}(i, \mathbf{X}, \mathbf{q}) &= \#(\mathbf{X} \cap [q_{i-1}, q_i]) - (\lceil np_i \rceil - \lceil np_{i-1} \rceil) . \end{aligned}$$

Nous remarquons que cette expression est très similaire à celle de la fonction d'utilité de JointExp et nous faisons ainsi un parallèle entre cet algorithme et la théorie générale de la sensibilité inverse. Ces similitudes nous permettent aussi d'adapter l'algorithme d'échantillonnage de JointExp pour la sensibilité inverse.

Nous remarquons aussi que pour ces deux algorithmes, la densité de probabilité de l'échantillonnage est constance sur les blocks $\big([X_{i_1}, X_{i_1+1}) \times \ldots \times \cdots \times [X_{i_m}, X_{i_m+1})\big) \cap$

$[a,b]^m\nearrow$ pour $\mathbf{i} = (i_1, \ldots, i_m) \in O'$ où $O' = \{\mathbf{i} \in \{0, \ldots, n\}^m, 0 \leq i_1 \leq \cdots \leq i_m \leq m\}$. Ainsi, lorsque les points sont très rapprochés au voisinage des points de grande utilité, il se peut que la probabilité de renvoyer quelque chose d'intéressant soit très faible, voire nulle en cas de collisions. Nous formalisons cette remarque dans le résultat suivant :

Proposition 1. *Supposons qu'il existe $q \in (a,b)$ et $\eta > 0$ tels que $I := (q-\eta, q+\eta) \subset [a,b]$ soit tel que $\mathbb{P}_X(\{q\}) > 0$ et $\mathbb{P}_X(I \setminus \{q\}) = 0$. Alors il existe des vecteurs de probabilité $\mathbf{p}$ tels que*

$$\mathbb{E}_{\mathbf{X}, \mathcal{E}^{(2/\epsilon)}_{u_{JE}}} \left(||\mathcal{E}^{(2/\epsilon)}_{u_{JE}}(\mathbf{X}) - F_X^{-1}(\mathbf{p})||_2^2 \right) = \Omega_n(1) \ ,$$

où nous utilisons la notation vectorielle $F_X^{-1}(\mathbf{p}) = (F_X^{-1}(p_1), \ldots, F_X^{-1}(p_n))$. De plus, la mesure de Lebesgue de l'ensemble des vecteurs qui vérifient cette condition est bornée inférieurement par $\mathbb{P}_X(\{q\})^m$.

En d'autres termes, pour des distributions avec des atomes isolés, l'estimation des quantiles n'est pas consistante. Par exemple, si la distribution est un Dirac centré en 0 et que nous essayons d'estimer sa médiane (avec $a = -1$ et $b = -2$), l'estimateur aura une distribution uniforme sur $[-1, 1]$ lorsque n est impair.

3 Lissage heuristique via perturbation des données

Bien que la théorie de la sensibilité inverse n'avait pas été appliquée à l'estimation jointe de plusieurs quantiles avant, ce comportement avait déjà été anticipé et une version lissée de mécanisme à sensibilité inverse existe pour éviter ces problèmes. Malheureusement, nous n'avons pas réussi trouver un algorithme d'échantillonnage en temps raisonnable pour ce mécanisme. Pour contourner cette difficulté, nous proposons une solution qui est extraordinairement simple d'un point de vue algorithmique: ajouter du bruit aux données. En effet, de cette manière, nous évitons d'avoir des points trop proches voire confondus et nous ne modifions que peu les statistiques d'ordre. Soit $w_1, \ldots, w_n$ des variables aléatoires i.i.d, et soit $\tilde{\mathbf{X}} = (X_1 + w_1, \ldots, X_n + w_n)$. L'algorithme de JointExp heuristiquement lissé (HSJointExp) renvoie $\mathcal{E}^{(2/\epsilon)}_{u_{\mathrm{JE}}}(\tilde{\mathbf{X}})$, le résultat de JointExp appliqué à $\tilde{\mathbf{X}}$. Nous prouvons que cette technique permet de résoudre une partie des problèmes évoqués.

Théorème 2. *Si la distribution de X est une mixture d'un nombre fini de Diracs dans (a,b) et d'une variable aléatoire Y a densité continue π_Y sur $[a,b]$ relativement a la mesure de Lebesgue telle que $\pi_Y > 0$ sur $[a,b] \setminus \mathcal{O}$ où $\mathcal{O}$ est une union finie d'intervalles et où $\pi_Y = 0$ sur $\mathcal{O}$, alors, pour toute précision δ et Lebesgue presque tout vecteur de probabilités $\mathbf{p}$, il existe un estimateur ϵ-DP $\mathbf{q}$ basé sur HSJointExp qui satisfait*

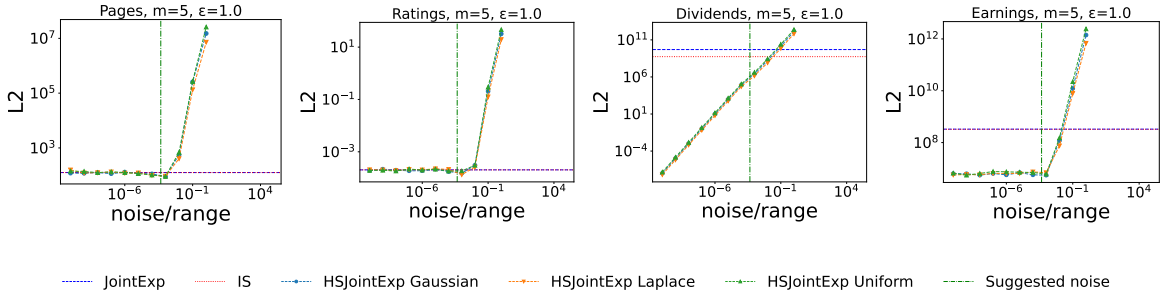
$$||\mathbf{q} - F_X^{-1}(\mathbf{p})||_\infty \leq \delta$$

avec forte probabilité (pour n grand).

En particulier, dans l'exemple du Dirac de la partie précédente, nous prouvons que l'estimateur résultant a un risque quadratique qui converge exponentiellement vite vers 0.

4 Résultats expérimentaux

Nos résultats expérimentaux confirment nos allégations théoriques. Nous appliquons HSJointExp sur quatre ensembles de données issus du monde réel (voir Figure 1). Pour les quatre ensembles, soit HSJointExp obtient des résultats similaires à JointExp lorsque ce dernier obtenait de bons résultats, soit il existe un niveau de bruit permettant à HSJointExp d'obtenir de bien meilleurs résultats que JointExp sur les jeux de données piqués (voir Figure 2).



$n = 1000$, $\epsilon = 1$, $\mathbf{p} = \left(\frac{1}{m+1}, \ldots, \frac{m}{m+1}\right)$, $L2$ = Risque quadratique entre l'estimateur confidentiel et les vrais quantiles empiriques.

Figure 2: Performances pour différentes valeurs de lissage, $\epsilon = 1$.

Bibliographie

Dwork, C. et Roth, A. (2014), The algorithmic foundations of differential privacy., *Foundations and Trends in Theoretical Computer Science.*

McSherry, F. et Talwar, K. (2007), Mechanism design via differential privacy, *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07).*

Gillenwater, J. et Joseph, M. et Kulesza, A. (2021). Differentially Private Quantiles, arXiv preprint arXiv:2102.08244.

Asi, H. et Duchi, J. (2020),Near instance-optimality in differential privacy, arXiv preprint arXiv:2005.10630ès.

GÉNÉRALISATION ET ESTIMATION DE LA Q-DISTRIBUTION GAUSSIENNE

Ben Mrad Oumaima [1,2] & Masmoudi Afif [1] & Slaoui Yousri [2]

[1] *laboratory of probability and statistics, University of Sfax, Tunisia.*
[2] *laboratory of mathematics and applications, University of Poitiers, France.*
oumaima.ben.mrad@univ-poitiers.fr & afifmasmoudi@gmail.com &
Yousri.Slaoui@math.univ-poitiers.fr

Résumé. Notre objectif est de compléter le travail de Diaz et Pariguan (2009) en introduisant la distribution q-Gaussienne $\mathcal{N}_q(0,\sigma^2)$ centrée et de variance σ^2. De plus, les notions d'entropie de Shannon et de quantile seront généralisées. Nous mettons en évidence la relation de convergence de $\mathcal{N}_q(0,\sigma^2)$ vers les distributions Uniforme sur $[-\sigma,\sigma]$ et Gaussienne sur la droite réelle lorsque q tend vers 0 et 1 respectivement. En outre, pour chaque valeur de q on obtient une q-Gaussienne donc un nouveau paramètre qui s'ajoute avec la moyenne et la variance ce qui implique plus de flexibilité. Enfin, une estimation des paramètres q et σ^2 sera effectuée en suivant l'algorithme E-M.
Mots-clés. q-calcul, distribution q-Gaussienne, q-Entropie, q-Quantile.

Abstract. The aim of this paper is to complet the work of Diaz and Pariguan (2009) by constructing the centered q-Gaussian distribution $\mathcal{N}_q(0,\sigma^2)$ with variance σ^2. In addition, the notion of entropy of Shannon and quantile will be generalized. However, we highlight the relation of convergence of $\mathcal{N}_q(0,\sigma^2)$ to the Uniform distribution on $[-\sigma,\sigma]$ and to the Gaussian distribution on the real line when q tends to 0 and 1 respectively. Furthermore, we obtain a q-Gaussian for each value of q, implying that a new parameter is added to the mean and variance, implying more flexibility. Finally an estimation of the parameters q and σ^2 will be achieved by following the E-M algorithm.
Keywords. q-Calculus, q-Gaussian distribution, q-Entropy, q-Quantil.

1 Introduction

Le calcul quantique, connu aussi sous le nom du $q-$calcul, est le nom moderne de l'étude du calcul sans limites. Les premières formules du $q-$calcul ont été obtenues par Euler au $XVIII^e$ siècle et par la suite, Jackson (2010) a introduit les notions de $q-$intégrale et $q-$dérivée. Le $q-$calcul a été développé plus tard par Kac et Cheung (2001). En analyse quantique, il n'existe pas une règle générale pour définir la notion du $q-$analogue. Une définition intuitive d'un $q-$analogue d'un tel objet mathématique A, est une famille d'objets $(A_q)_{q\in\mathbb{R}\setminus 1}$, tels que A_q tend vers A lorsque q tend vers 1. Cette condition est

nécessaire en théorie du $q-$calcul. Le calcul standard est utilisé comme un outil discret dans l'ordinateur classique. Cette théorie du calcul quantique nous permet de créer le qubits ou $q-$bits. Actuellement l'ordinateur quantique fonctionne numériquement plus rapide que l'ordinateur classique, car il fonctionne avec le $q-$bits à la place de bits. En physique mathématiques et en probabilités, la q-distribution est plus générale que la distribution classique. Elle a été introduite par Diaz et Pariguan (2009) dans le cas continu et par Charalambides (2010) dans le cas discret. La construction d'une q-distribution de probabilités revient à la construction d'un q-analogue de la distribution ordinaire. Une fonction f_q est dite q-densité de probabilité si elle satisfera les conditions $f_q(x) \geq 0 \; \forall x \in \mathbb{R}$ et $\int_{-[\infty]_q}^{-[\infty]_q} f_q(x) d_q x = 1$. Avec, $[\infty]_q = \dfrac{1}{1-q}$ est le q-analogue de l'infini et $\int_0^b f(t) d_q t = b(1-q)\sum_{j=0}^{\infty} q^j f(q^j b)$ et $\int_a^b f(t) d_q t = \int_0^b f(t) d_q t - \int_0^a f(t) d_q t$ sont les intégrales de Jackson d'une fonction $f : \mathbb{R} \to \mathbb{R}$ sur $[0, b]$ et $[a, b]$ respectivement. La fonction de q-répartition $F_q(x)$ est donnée par $F_q(x) := \mathbb{P}_q(X \leq x) \int_{-[\infty]_q}^{x} f_q(x) d_q x$. Diaz et Pariguan (2009) ont introduit le q-analogue de la distribution Gaussienne centrée réduite $\mathcal{N}_q(0, 1)$ par la donnée de sa fonction q-densité :

$$s_q(x) = \frac{1}{c(q)} E_{q^2}^{-\frac{q^2 x^2}{[2]_q}} \mathbb{I}_{[-\nu,\nu]}(x) \; \forall 0 < q < 1. \tag{1}$$

Où $c(q) := \int_{-\nu}^{\nu} E_{q^2}^{-\frac{q^2 x^2}{[2]_q}} d_q x$ est la constante de normalisation de $\mathcal{N}_q(0, 1)$, $\nu := \dfrac{1}{\sqrt{1-q}}$, E_q^x est la fonction q-exponentielle introduite par Jackson (2010) et définie par $E_q^x = \sum_{j=0}^{\infty} q^{j(j-1)/2} \dfrac{x^j}{[j]_q!}$. Avec, $[\alpha]_q = \frac{1-q^\alpha}{1-q}$, $\forall \alpha \in \mathbb{R}$ et $[n]_q! = \begin{cases} 1 & \text{si } n = 0 \\ [n]_q[n-1]_q ... [1]_q & \text{si } n \in \mathbb{N}^*. \end{cases}$

Dans tout ce qui suit on suppose que $0 < q < 1$ et $\nu = \dfrac{1}{\sqrt{1-q}}$.

2 Résultats

La proposition suivante présente un q-analogue de la distribution Gaussienne centrée et de variance σ^2, notée $\mathcal{N}_q(0, \sigma^2)$.

Proposition 1 *Soient $0 < q < 1$ et $\sigma > 0$. Si $X \sim \mathcal{N}_q(0, 1)$ alors $\sigma X \sim \mathcal{N}_q(0, \sigma^2)$ a pour q-densité :*

$$g_q(y) = \frac{1}{\sigma c(q)} E_{q^2}^{-\frac{q^2 y^2}{\sigma^2 [2]_q}} \mathbb{I}_{[-\sigma\nu,\sigma\nu]}(y). \tag{2}$$

Où, $c(\sigma,q) := \int_{-\sigma\nu}^{\sigma\nu} E_{q^2}^{-\frac{q^2y^2}{\sigma^2[2]_q}} d_qy = \sigma c(q)$ *est la constante de normalisation de* $N_q(0,\sigma^2)$.

Définition 1 *On définit la fonction q-répartition de la q-distribution* $\mathcal{N}_q(0,\sigma^2)$ *par :*

$$G_q(y) = \begin{cases} 0 & si \quad y < -\sigma\nu \\ \int_{-\sigma\nu}^{y} g_q(t)d_qt & si \quad -\sigma\nu \leq y \leq \sigma\nu \\ 1 & si \quad y > \sigma\nu. \end{cases} \tag{3}$$

Au niveau de la proposition suivante, nous donnons l'expression explicite de la fonction q-répartition de $\mathcal{N}_q(0,\sigma^2)$.

Proposition 2 *Pour* $0 < q < 1$, $\sigma > 0$ *et* $\nu = \frac{1}{\sqrt{1-q}}$, *l'expression explicite de* G_q *est :*

$$G_q(y) = \mathbb{I}_{[-\sigma\nu,\sigma\nu]}(y)\left(\frac{1}{2} + \frac{1-q}{c(\sigma,q)}\sum_{m\geq 0}\frac{q^{m(m+1)}(q-1)^m y^{2m+1}}{(1-q^{2m+1})(1-q^2)_{q^2}^m\sigma^{2m}}\right) + \mathbb{I}_{[\sigma\nu,\infty]}(y). \tag{4}$$

Définition 2 *Soient* $0 < q < 1$ *et* X *est une variable aléatoire de fonction q-densité* f_q. *Alors, la q-Entropie de* X *notée* $H_q(X)$ *est définie par :*

$$H_q(X) := \mathbb{E}_q(-ln(f_q(X))). \tag{5}$$

Définition 3 *Soient* $0 < q < 1$, $0 < \alpha < 1$ *et* X *est une variable aléatoire de fonction q-répartition* G_q. *On note le q-Quantile d'ordre* α *de* X *par* x_α *et il est défini par :*

$$\mathbb{P}_q(X \leq x_\alpha) = \alpha \Leftrightarrow x_\alpha = G_q^{-1}(\alpha). \tag{6}$$

Où $G_q^{-1}(u) = inf\{x; G_q(x) \leq u\}$ *est le pseudo inverse de la fonction q-répartition* G_q.

Ce concept du q-Quantile peut être appliqué dans le domaine de finance pour calculer la VAR (Value-At-Risque), qui représente la perte potentielle maximale d'un investisseur sur la valeur d'un actif ou d'un portefeuille d'actifs financiers qui ne devrait être atteinte qu'avec une probabilité donnée sur un horizon donné.

3 Estimation

Notre but dans ce paragraphe est d'estimer les paramètres q et σ^2 présentés dans l'Équation 2. Soient $q_1, ..., q_n$ n variables aléatoires non observables, indépendantes et de même loi Bêta $\beta(a,b)$ de paramètres $a > 0$ et $b > 0$. Soient $y_1, ..., y_n$ n observations aléatoires

indépendantes telles que $Y_i \mid q_i \sim \mathcal{N}_{q_i}(0, \sigma^2)\ \forall i = 1, ..., n$. La fonction vraisemblance complète est alors donnée par l'expression suivante :

$$l_c(y_1, ..., y_n, q_1, ..., q_n; \sigma^2, a, b) = \prod_{i=1}^{n} \left\{ \mathcal{N}_{q_i}(0, \sigma^2)(y_i) \beta(a, b)(q_i) \right\}. \tag{7}$$

Soit $\Theta := (a, b, \sigma^2)$ le vecteur des paramètres inconnus de notre modèle mélange. En partant d'un ensemble de paramètres initiaux $\Theta^{(0)} := (a_0, b_0, \sigma_0^2)$, l'algorithme E-M alterne deux étapes :
Etape 1: Espérence: $Q(\Theta \mid\mid \Theta^{(l)}) := \mathbb{E}_{\Theta^{(l)}} \left(log l_c(y_1, ..., y_n, q_1, ..., q_n; \sigma^2, a, b) \mid y_1, ..., y_n \right)$.
Etape 2: Maximisation: $\Theta^{(l+1)} = \underset{\Theta}{argmax}\ Q(\Theta \mid\mid \Theta^{(l)})$.
L'algorithme s'arrête lorsque la norme de la différence entre deux estimations successives des paramètres est suffisamment petite, i.e., $\mid\mid \Theta^{(l+1)} - \Theta^{(l)} \mid\mid \leq \varepsilon$, pour ε prédéfini et $\|\|$ est une norme choisie sur l'espace des paramètres. Les estimations obtenues sont notées $\hat{a}, \hat{b}$ et $\hat{\sigma^2}$ de a, b et σ^2 respectivement.

4 Etudes numériques

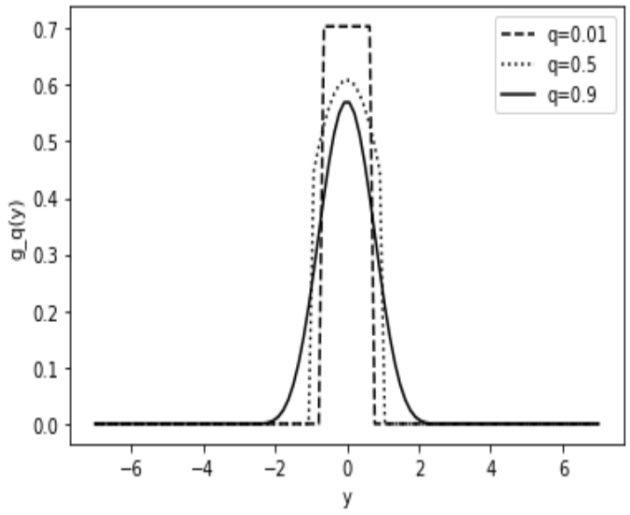


Figure 1: Représentation graphique de la q-densité de $\mathcal{N}_q(0, \sigma^2)$.

$$\mathcal{U}_{[-\sigma,\sigma]} \underset{0 \leftarrow q}{\overset{D}{\leftarrow}} \mathcal{N}_q(0, \sigma^2) \underset{q \to 1}{\overset{D}{\rightarrow}} \mathcal{N}(0, \sigma^2).$$

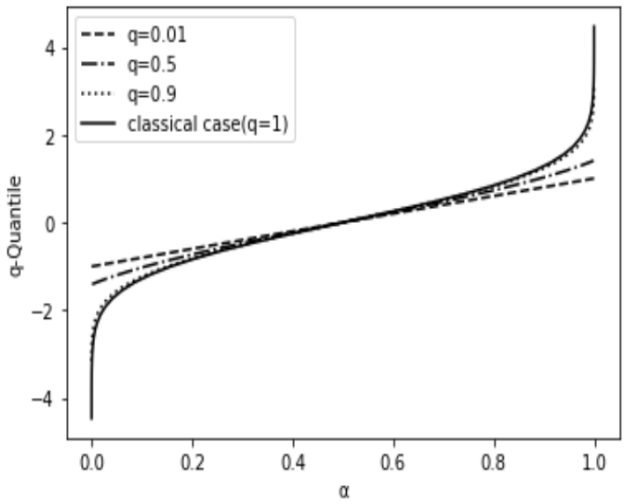


Figure 2: Représentation graphique de la fonction q-Quantile de $\mathcal{N}_q(0,1)$.

	q=0.01			q=0.5			q=0.9			classical case		
α	0.9	0.95	0.975	0.9	0.95	0.975	0.9	0.95	0.975	0.9	0.95	0.975
x_α	0.805	0.905	0.955	1.030	1.204	1.303	1.238	1.566	1.836	1.282	1.645	1.960

Table 1: Les q-Quantiles de $\mathcal{N}_q(0,1)$pour différentes valeurs de q et de α.

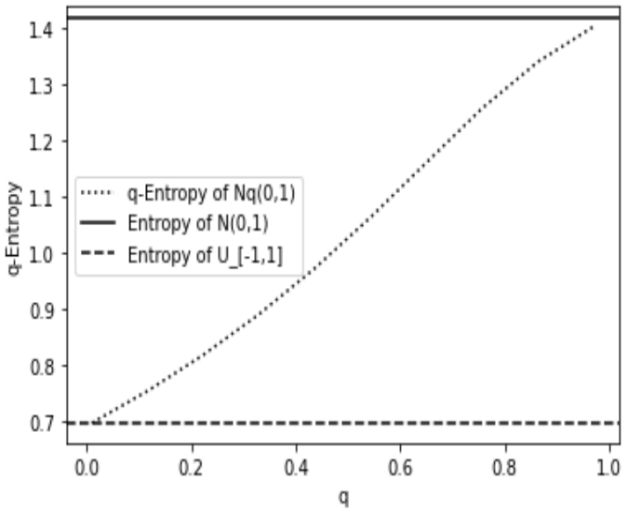


Figure 3: Représentation graphique de la q-Entropie.

$$H\left(\mathcal{U}_{[-1,1]}\right) \underset{0\leftarrow q}{\longleftarrow} H_q\left(\mathcal{N}_q(0,1)\right) \underset{q\to 1}{\longrightarrow} H\left(\mathcal{N}(0,1)\right).$$

Le tableau suivant illustre les résultats d'estimations obtenues en appliquant l'algorithme E-M présenté dans la Section 3.

a	b	σ	$\hat{a}$	$\hat{b}$	$\hat{\sigma}$
1	2.2	1	0.993	2.1	2.193
2	1	0.5	1.99	0.2	0.48
3.5	4	0.2	3.491	3.98	0.192
4.3	2	5	4.285	1.982	1.9625
5	3	1	4.991	2.986	0.98

Table 2: Les paramètres estimés.

5 Conclusion

Dans ce travail, nous avons introduit une généralisation de la distribution $\mathcal{N}(0, \sigma^2)$, qui est une extension du travail de Diaz et Pariguan (2009) (pour $\sigma = 1$). De plus, nous avons étudié ses propriétés caractéristiques et nous avons montré que cette q-distribution interpole entre la distribution Uniforme et la loi Gaussienne classique. D'autre part, nous avons proposé un q-analogue de l'Entropie de Shannon et du Quantile et nous avons donné leurs représentations graphiques en montrant à chaque fois qu'on trouve le cas classqiue lorsque q tend vers 1. Finalement, nous avons estimé q et σ^2 en suivant l'algorithme E-M qui nous a permet d'avoir de bon résultats.

Bibliographie

Diaz, R. and Pariguan, E. (2009). On the Gaussian q-distribution, *Journal of mathematical analysis and applications*, 358(1), pp. 1-9.

Jackson, FH. (1910). On q-definite integrals, *Quarterly Journal of Pure and Applied Mathematics*, 41, pp. 193-203.

Kac, V. and Cheung, P. (2001). Quantum calculus, *Springer Science & Business Media*.

Charalambides, C. A. (2010). Discrete q-distributions on Bernoulli trials with a geometrically varying success probability, *Journal of statistical planning and inference*, 140, pp. 2355-2383.

Fitouhi, A. and Bettaibi, N. (2007). Applications of the Mellin transform in quantum calculus, *Journal of mathematical analysis and applications*, 328, pp. 518-534.

De Sole, A. and Kac, V. (2003). On integral representations of q-gamma and q-beta functions, *arXiv preprint math*.

Imen, B., Imed, B., and Afif, M. (2019). On characterizing the exponential q-distribution. *Bulletin of the Malaysian Mathematical Sciences Society*, 42, pp. 3303-3322.

Diaz, R. and Teruel, C. (2005). q, k-Generalized gamma and beta functions, *Journal of Nonlinear Mathematical Physics*, pp. 118-134.

Projection de mesures de probabilité sous contraintes de quantile par distance de Wasserstein et approximation monotone polynomiale

Marouane Il Idrissi [1,2,3,*], Nicolas Bousquet [1,3,4], Fabrice Gamboa [2], Bertrand Iooss [1,2,3], Jean-Michel Loubes [2]

[1] *EDF R&D, 6 Quai Watier, 78400 Chatou, France*
[2] *Institut de Mathématiques de Toulouse, 31062, Toulouse, France*
[3] *SINCLAIR AI Laboratory, Saclay, France*
[4] *Sorbonne Université, LPSM, 4 place Jussieu, Paris, France*

[*] *Auteur correspondant - marouane.il-idrissi@edf.fr*

Résumé. Motivés par des applications en analyse de robustesse pour la quantification d'incertitude et l'apprentissage statistique, nous nous intéressons au problème de projection d'une mesure de probabilité univariée donnée. L'ensemble de projection est composé de mesures de probabilité respectant certaines contraintes sur leur quantiles. Ce travail explore le cas de la distance de Wasserstein pour des mesures de probabilité portées par $\mathbb{R}$. Ce problème particulier de projection dans des espaces de mesures sous contraintes quantile se réécrit comme un problème de projection dans $L^2([0,1])$ avec contraintes de monotonie et d'interpolation. Ce dernier problème se résout analytiquement. Nous présentons également une régularisation de la solution en nous restreignant aux fonctions polynomiales par morceaux. Finalement, nous illustrons ces résultats sur des données réelles.

Mots-clés. Projection de mesures de probabilité, quantiles, transport optimal, approximation monotone.

Abstract. Motivated by applications in robustness analysis for uncertainty quantification and statistical learning, this paper focuses on the problem of univariate probability measure projection, under quantile constraints. The choice of the Wasserstein distance is explored. The initial problem can be equivalently written as a projection in $L^2([0,1])$ with interpolation and monotonicity constraints, and can be solved analytically. We leverage this equivalence to ensure some smoothness in the solution, by further restricting the solution to be a continuous piece-wise non-decreasing polynomial. Finally, some illustrations on a dataset are presented.

Keywords. Probability measure projection, quantiles, optimal transport, isotonic approximation.

1 Projection W_2 sous contraintes quantile

Soit P une mesure de probabilité supportée par $\mathcal{X} \subseteq \mathbb{R}$. Sans distinction dans la suite, P peut être une mesure de probabilité *empirique*, i.e., composée d'atomes issus d'un n-

échantillon réel noté $X = (X_1, \ldots, X_n)$, ou une mesure de probabilité *continue*, i.e., qui ne compte aucun atome, et supposée appartenir à l'ensemble des mesures de probabilités réelles de moment d'ordre 2 fini, noté $\mathcal{P}_2(\mathbb{R})$. La fonction de répartition de P est notée $F_P(x) = P((-\infty, x])$ et on note $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, +\infty\}$. Nous pouvons définir deux inverses généralisées de F_P : son inverse à gauche, définie pour tout $y \in [0, 1]$ par

$$F_P^{\leftarrow}(y) := \sup \{x \in \overline{\mathbb{R}} \mid F_P(x) < y\} = \inf \{x \in \bar{\mathbb{R}} \mid F_P(x) \geq y\}$$

qui est croissante et continue à gauche, et qui correspond à la fonction quantile classique associée à P. Son inverse à droite, définie pour tout $y \in [0, 1]$ par

$$F_P^{\rightarrow}(y) := \sup \{x \in \overline{\mathbb{R}} \mid F_P(x) \leq y\} = \inf \{x \in \bar{\mathbb{R}} \mid F_P(x) > y\}$$

qui est croissante, continue à droite, et est égale presque partout à $F_P^{\leftarrow}$. Notons également l'espace fonctionnel

$$\mathcal{F}^{\rightarrow} = \{G : [0, 1] \to \overline{\mathbb{R}} \mid G \text{ est continue à droite, croissante, } \lim_{y \to 0} G(y) = -\infty, \lim_{y \to 1} G(y) = +\infty\}.$$

$\mathcal{F}^{\rightarrow}$ est l'espace des inverses généralisées *valides*, i.e., chaque fonction dans $\mathcal{F}^{\rightarrow}$ est l'unique inverse généralisée à droite d'une fonction de répartition, associée à une unique mesure de probabilité. La Figure 1 illustre les différences entre ces deux inverses généralisées : elles ne diffèrent qu'aux points dont l'image est un atome de P (de La Fortelle 2015).

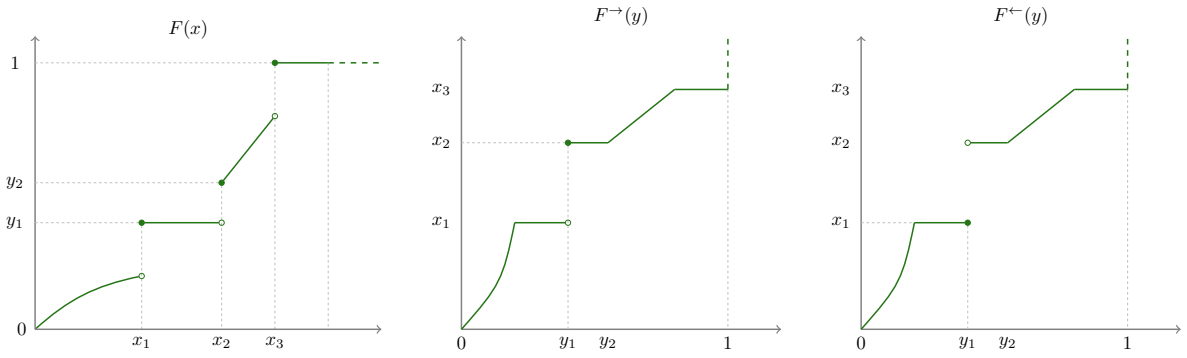


Figure 1: Une fonction de répartition et ses inverses généralisées.

Notre intérêt pour une perturbation des quantiles de F_P est issu d'une réflexion plus large sur la nécessité, dans des applications variées en quantification d'incertitude ou en apprentissage statistique, de tester la robustesse d'une prévision dépendante de X vis-à-vis de la loi de X, que celle-ci soit connue explicitement (Da Veiga et al. 2021), ou au travers de réalisations (Bachoc et al. 2020). Ce type de perturbation peut s'interpréter plus aisément que des contraintes usuelles de type moment généralisé. Plus précisément, nous imposons des contraintes sur les seuils de quantile, par essence bornés sur $[0, 1]$.

Afin de définir de telles contraintes, prenons plusieurs niveaux de probabilités $\alpha = (\alpha_1, \ldots, \alpha_K)^\top$ et $\beta = (\beta_1, \ldots, \beta_K)^\top$ avec $\alpha, \beta \in (0,1)^K$, où $K \in \mathbb{N}^*$ représente le nombre de contraintes quantiles souhaitées. Afin d'être valides, les niveaux doivent respecter la condition

$$\beta_1 < \beta_2 < \cdots < \beta_K, \quad \text{et} \quad \alpha_1 < \alpha_2 < \cdots < \alpha_K.$$

Pour $i = 1, \ldots, K$, dans le cas où $\alpha_i = \beta_i$, nous parlerons de *conservation de quantiles*, et si $\alpha_i \neq \beta_i$, nous parlerons de *perturbation de quantile*. Une mesure de probabilité Q respecte une contrainte quantile lorsque

$$\begin{cases} F_Q^{\rightarrow}(\beta) = F_P^{\leftarrow}(\alpha) & \text{si } \beta < \alpha, \\ F_Q^{\leftarrow}(\beta) = F_P^{\leftarrow}(\alpha) & \text{si } \beta \geq \alpha. \end{cases}$$

Ainsi, pour une mesure P, et des niveaux valides fixés, notons $\mathcal{C}$ la restriction de $\mathcal{P}_2(\mathbb{R})$ des mesures de probabilités respectant ces contraintes quantiles. Nous nous intéressons au problème de projection suivant

$$Q = \underset{G \in \mathcal{P}_2(\mathbb{R})}{\operatorname{argmin}} \quad W_2^2(P, G) \quad \text{s.c.} \quad G \in \mathcal{C}, \tag{1}$$

où W_2 dénote la distance de Wasserstein d'ordre 2. Essentiellement, nous recherchons une mesure de probabilité Q respectant les contraintes quantiles, et étant la plus proche de la mesure de probabilité initiale P, au sens de la distance de Wasserstein d'ordre 2. A l'aide de la reformulation de la distance de Wasserstein dans le cas particulier des mesures réelles (Villani 2003), et par unicité entre les fonctions de répartition et leurs inverses généralisées, ce problème de projection se réécrit de la manière suivante :

$$\begin{aligned} F_Q^{\rightarrow} = \underset{L \in \mathcal{F}^{\rightarrow}}{\operatorname{argmin}} \ & \left\{ \int_0^1 (F_P^{\rightarrow}(x) - L(x))^2 dx \right\} \\ \text{s.c.} \ & \begin{cases} L(\beta_i) = F^{\leftarrow}(\alpha_i) & \text{si } \beta_i < \alpha_i \\ L(\beta_i^-) = F^{\leftarrow}(\alpha_i) & \text{si } \beta_i \geq \alpha_i \end{cases}, \quad i = 1, \ldots, K, \end{aligned} \tag{2}$$

où $F_Q^{\rightarrow}$ est définie sur la mesure de probabilité Q, minimisant le problème de projection introduit en (1). Cette formulation équivalente peut être résolue de manière analytique : soient $a, b \in (0,1)^K$ définis par:

$$\begin{aligned} a_1 = \min(\alpha_1, \beta_1), \quad a_i = \min\Big[\max(\beta_{i-1}, \alpha_i), \beta_i\Big], i = 2, \ldots, K, \\ b_K = \max(\alpha_K, \beta_K), \quad b_j = \max\Big[\min(\alpha_i, \beta_{j+1}), \beta_i\Big], j = 1, \ldots, K-1, \end{aligned}$$

et notons $A_i = [a_i, b_i)$, et $\overline{A} = [0,1] \setminus \left(\bigcup_{i=1}^K A_i\right)$. Alors, la solution du problème (2) s'écrit

$$F_Q^{\rightarrow}(x) = \begin{cases} F_P^{\rightarrow}(x) & \text{si } x \in \overline{A}, \\ F_P^{\leftarrow}(\alpha_i) & \text{si } x \in A_i, \quad i = 1, \ldots, K. \end{cases}$$

En d'autres termes, Q diffère de P dans le sens où elle présente des atomes aux points $F_P^{\leftarrow}(\alpha_i)$, de masse $Q\big(\{F_P^{\leftarrow}(\alpha_i)\}\big) = b_i - a_i$, pour $i = 1, \dots, K$.

2 Polynômes monotones continus par morceaux

La reformulation (2) permet également de mettre en avant le lien entre projection de mesures de probabilité par W_2 sous contraintes quantiles et les projections de fonctions dans $L^2([0,1])$ sous contraintes conjointes d'interpolation et de monotonie. La solution peut être lissée en restreignant l'espace de recherche $\mathcal{F}^{\rightarrow}$ à des fonctions plus régulières.

Considérons une restriction de $\mathcal{F}^{\rightarrow}$ aux fonctions continues polynomiales par morceaux. Soit $\alpha, \beta \in (0,1)^K$, notons $d \in \mathbb{N}^*$ un degré de polynôme et intéressons-nous aux fonctions de la forme :

$$G(x) = \begin{cases} G_0(x) & \text{si } 0 \leq x < \beta_1, \\ \vdots & \\ G_i(x) & \text{si} \beta_i \leq x < \beta_{i+1}, \\ \vdots & \\ G_K(x) & \text{si } \beta_K \leq x \leq 1, \end{cases} \tag{3}$$

où chaque G_i est un polynôme de degré inférieur ou égal à d. Notons $\mathcal{G}$ la famille paramétrique des fonctions s'écrivant comme (3). Supposons maintenant que le support de P soit un compact $[a,b] \subset \mathbb{R}$, avec $a < b$ deux réels donnés. Le problème de projection de P par distance W_2 sur $\mathcal{G} \cap \mathcal{F}^{\rightarrow}$ peut s'écrire de manière équivalente comme

$$S = \underset{G \in \mathcal{G}}{\operatorname{argmin}} \left\{ \int_0^1 \left(F_P^{\rightarrow}(x) - G(x)\right)^2 dx \right\} \quad \text{s.c.} \quad \begin{cases} G(0) = a, \quad G(1) = b, \\ G_i(\beta_i) = G_{i+1}(\beta_i) = F_P^{\leftarrow}(\alpha_i) \\ \qquad \text{pour } i = 0, \dots K-1, \\ G'(x) \geq 0, \forall x \in [0,1]. \end{cases} \tag{4}$$

Ce problème initial est séparable sur chaque intervalle et chaque sous-problème peut être reformulé comme un problème de recherche des coefficients de chaque G_i sur l'intervalle correspondant. De manière générique, cela revient à résoudre le problème quadratique convexe, avec les contraintes convexes suivantes, pour un intervalle $[\beta_i, \beta_{i+1}]$, $i = 0, \dots, K$ (où $\beta_0 = 0$, et $\beta_{K+1} = 1$) :

$$s = \underset{g \in \mathbb{R}^{d+1}}{\operatorname{argmin}} \; g^\top M g - 2 g^\top r \quad \text{s.c.} \quad \begin{cases} g^\top \widetilde{\beta}_i = F_P^{\leftarrow}(\alpha_i), \quad g^\top \widetilde{\beta}_{i+1} = F_P^{\leftarrow} \alpha_{i+1}, \\ g \in \mathcal{K}, \end{cases}$$

où $\widetilde{\beta} = (1, \beta, \beta^2, \dots, \beta^d)^\top$, M représente la matrice $\big((d+1) \times (d+1)\big)$ des moments de la mesure de Lebesgue sur l'intervalle considéré, i.e.,

$$M_{ij} = \int_{\beta_i}^{\beta_{i+1}} x^{i+j-2} dx = \frac{(\beta_{i+i})^{i+j-1} - (\beta_i)^{i+j-1}}{i+j-1},$$

et r est le vecteur des moments de $F_P^{\rightarrow}$, i.e., pour $i = 0, \ldots, d$, $r_i = \int_{t_0}^{t_1} x^i A(x) dx$. L'ensemble $\mathcal{K}$ représente une sous-partie convexe de $\mathbb{R}^{d+1}$, étant l'image d'une composition d'applications linéaires, notée $T : \mathbb{S}_k \times \mathbb{S}_l \to \mathbb{R}^{d+1}$, pouvant s'écrire :

$$\mathcal{K} = \left\{ g \in \mathbb{R}^{d+1} \mid T(\Gamma) = g, \quad \Gamma \in \Sigma_k \times \Sigma_l \right\}$$

où $\mathbb{S}_k$ représente le sous-espace vectoriel des matrices $(k \times k)$ symmétriques et Σ_k représente le cône convexe fermé des matrices semi-définies positives dans $\mathbb{S}_k$. Les tailles k, l des matrices dépendent de la parité de d : si d est pair, $k = d$ et $l = d - 2$; si d est impair, alors $k = d - 1$ et $l = d - 1$. Ce résultat est issu de la représentation successive de la dérivée du polynôme $G_i(x)$ par le biais de polynômes *sum-of-squares* (SOS) (Lasserre 2015), garantissant sa positivité sur l'intervalle considéré, puis de la représentation des polynômes SOS par le biais de matrices symmétriques semi-définies positives (Parrilo 2010, 2012). Ce problème étant convexe, il peut être résolu par le biais d'outils de résolution numérique, tel `CVXR`, une interface en langage `R` de l'outil `CVX` (Grant and Boyd 2014).

3 Approximation continue de mesure empirique : exemple des données d'Ozone

Afin d'illustrer nos résultats, nous allons nous intéresser aux données `airquality`, et en particulier aux mesures journalières d'Ozone à New York entre Mai et Septembre 1973. Nous avons accès à 116 observations complètes, définissant notre mesure empirique initiale. Nous appliquons la projection de cette mesure, avec régularisation polynomiale par morceaux, en conservant les valeurs des quartiles empirique (Figure 2 haut) et en perturbant ces mêmes quartiles (Figure 2 bas).

Par ailleurs, cette approximation continue de la fonction quantile empirique permet, par le biais d'observations distribuées uniformément, de simuler facilement de nouveaux échantillons selon la loi Q résultante de la projection sous contraintes. Le contrôle des valeurs des quantiles permet également d'injecter des informations extérieures aux données, tout en préservant l'allure de l'échantillon observé. Des applications de cette méthode pour l'étude de la robustesse de prévisions issues de modèles *boîte noire* sont en cours.

Bibliographie

Bachoc, F., F. Gamboa, M. Halford, J-M. Loubes, and L. Risser. 2020. "Explaining Machine Learning Models using Entropic Variable Projection." *arXiv:1810.07924 [cs, stat].*

Da Veiga, S., F. Gamboa, B. Iooss, and C. Prieur. 2021. *Basics and Trends in Sensitivity Analysis. Theory and Practice in R.* SIAM.

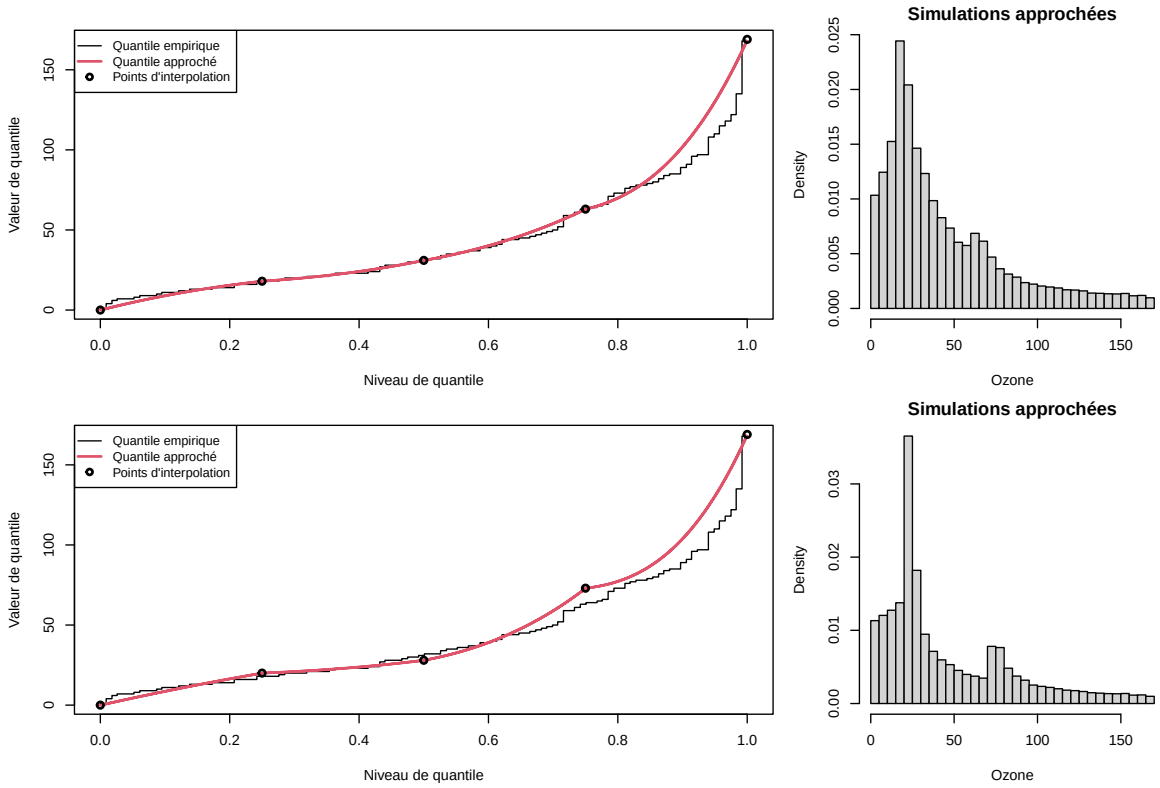


Figure 2: Approximations de la fonction de quantile empirique des données Ozone et histogrammes de simulations, par le résultat de la projection par conservation (haut) et perturbation (bas) des quartiles empiriques.

de La Fortelle, A. 2015. "A study on generalized inverses and increasing functions Part I: generalized inverses," 14.

Grant, M., and S. Boyd. 2014. *CVX: Matlab Software for Disciplined Convex Programming, version 2.1.* http://cvxr.com/cvx.

Lasserre, J-B. 2015. *An Introduction to Polynomial and Semi-Algebraic Optimization.* Cambridge University Press.

Parrilo, P. A. 2010. *Algebraic Optimization and Semidefinite Optimization : EIDMA Minicourse.*

———. 2012. "Chapter 3: Polynomial Optimization, Sums of Squares, and Applications." Chap. 3 in *Semidefinite Optimization and Convex Algebraic Geometry,* 47–157.

Villani, C. 2003. *Topics in Optimal Transportation.* Vol. 58. American Mathematical Society.

Sparse regression to anticipate the declarations of natural disaster in France

Thi Thanh Yen Nguyen [1,2] & Geoffrey Ecoto[1,3] & Antoine Chambaz [1,2]

[1] *MAP5 (UMR CNRS 8145), Université Paris Cité*
[2] *Fédération Parisienne de Modélisation Mathématique (FR CNRS 2036)*
[3] *Caisse Centrale de Réassurance*

Email: thi-thanh-yen.nguyen@u-paris.fr

Résumé. Les sécheresses (au sens du retrait-gonflement des argiles) constituent la seconde catastrophe naturelle la plus coûteuse dans le cadre légal de l'indemnisation des catastrophes naturelles en France. Nous développons une nouvelle méthodologie afin d'anticiper quelles communes bénéficieront d'une reconnaissance de catastrophe naturelle, l'un des éléments-clefs du régime d'indemnisation. La méthodologie repose sur la théorie du transport optimal et sur un algorithme proximal pour l'optimisation non-convexe.

Mots-clés. Transport optimal, algorithme proximal, algorithme de Sinkhorn, catastrophes naturelles

Abstract. Drought events are the second most expensive natural disasters within the legal framework of the natural disasters compensation scheme in France. We develop a new methodology to anticipate which cities will benefit from a declaration of natural disaster, a key step of the national compensation scheme. The methodology hinges on optimal transport theory and an inertial proximal algorithm for nonconvex optimization.

Keywords. Optimal transport, proximal algorithm, Sinkhorn algorithm, natural disasters

1 Introduction

The French state has been facing severe drought events over the past decade. For instance, the average annual cost of drought events between 2016 and 2020 amounts to approximately 1.1 billion EUROS, five times more than between 2002 and 2015.

France is one of the very few countries with a system that guarantees all its citizens adequate compensation in the event of loss and/or damage caused by a natural phenomenon such as severe droughts. The compensation scheme's solidarity is provided by the state-backed reinsurance proposed by Caisse Centrale de Réassurance (CCR), which enables the mutualisation at a national level of insurance portfolios and offers the State's guarantee.

Two prerequisites must be met to trigger the compensation scheme. First, the lost and/or damaged property must be covered by a property and casualty insurance policy (a condition of private nature). Second, a government decree declaring a disaster must be published in the Official Journal (a condition of public nature). It is the responsability of the mayors to initiate the request that the government declare a natural disaster for the cities they administer.

At the end of each civil year, CCR must anticipate the cost of natural disasters that occurred during the year. This is not an easy task. Ecoto et al. [4] developed an algorithm to anticipate the cost of drought events, the so-called one-step ahead sequential Super Learner (OSAS-SL). Based on the aggregation of a rich library of base-learners, OSAS-SL is shown to take advantage of the data consisting of a short time-series where each time-specific observation is a large network of many slightly dependent data. In their application, the authors consider only those cities that already benefit from a declaration of natural disaster. Charpentier et al. [1] also anticipate the cost of drought events. Simpler than the OSAS-SL, their algorithm relies on a generalized linear model and on tree-based machine learning algorithms. Unlike OSAS-SL, it also predicts which cities will benefit eventually from a declaration of natural disaster.

In the present study, we focus on the prediction of the declarations of natural disaster. It is a *sparse regression* task because we expect that only a small fraction of all French cities will request and benefit from such a declaration.

2 The data

The data set is obtained by merging several data sets.[1]

Given that we focus on drought events, it will not come as a surprise that we use the so-called SWI, an index provided by Météo France that quantifies soil wetness in a uniform way, see [1, section 2.3] for details. More specifically, for every year and each city, the city-level SWI is the convex average of the year-specific SWIs of the 8×8 km^2 squares that overlap the city's area. The weights are proportional to the areas of the intersections.

The city-level SWI is complemented with data from other sources, namely the National Institute for Statistical and Economic Studies (INSEE), the Geographic National Institute (IGN) and the French Geological Survey (BRGM). Eventually, for every year and each city, a city-level description encapsulates the city's profile. The description is multi-faceted. It contains: an indicator of whether or not a natural disaster was declared by the government (when available); a summary of the city's clay hazard, defined as the proportions of houses falling in each of four categories of clay hazard; a summary of the city's dwelling age, *i.e.*, how old houses are, under the form of the proportions of houses

[1]Note that, from now on, France refers to *Metropolitan* or *Mainland* France. Drought events are not a threat in Overseas France (essentially because there is little clay in these parts of the country).

falling in each of four categories; the climatic and seismic zones (a five-category and a four-category variables); a summary of the city's vegetation; the city's number of houses, population, area, average altitude, and density, defined as the ratio of the number of houses to the area. In addition, a variety of features are described by quantiles that summarize distributions (*e.g.*, the 30-quantiles of the distribution of the house-specific product of SWI and insured value, or the 30-quantile of the distribution of the house-specific product of the ground slope and insured value, to mention just a few). Overall, the city-level description consists of a little more than 430 variables.

3 The objective as an optimization problem

We denote by $x_m \in \mathcal{X}$ ($m \in [\![M]\!] := \{1, \ldots, M\}$) the mth year-specific city-level descriptions for which it is known whether or not the city benefited from a declaration of natural disaster, a piece of information denoted by $y_m \in \{0, 1\}$ (with convention $y_m = 1$ if the city with year-specific city-level description x_m did benefit from a declaration that year). Moreover, we denote by $x'_n \in \mathcal{X}$ ($n \in [\![N]\!]$) the nth year-specific city-level descriptions for which it is *not* known whether or not the city *will* benefit from a declaration of natural disaster. Predicting for each of the cities whether or not that will be the case is our objective.

To do so, we propose to solve the following optimization problem:

$$\underset{\theta \in \mathbb{R}^N}{\arg\min} \left\{ \min_{P \in \Pi} \left[\langle P, C(\theta) \rangle - \gamma E(P) \right] + \tau \|\theta\|_1 \right\}, \tag{1}$$

where

- $\Pi := \{P \in \mathbb{R}_+^{M \times N} : P\,\mathbf{1}_N = \frac{1}{M}\,\mathbf{1}_M, P^\top \mathbf{1}_M = \frac{1}{N}\,\mathbf{1}_N\}$ represents the joint laws on $[\![M]\!] \times [\![N]\!]$ with uniform marginal laws;
- the cost matrix $C(\theta) \in \mathbb{R}_+^{M \times N}$ is given by $(C(\theta))_{m,n} := d(x_m, x'_n)^2 + (y_m - \theta_n)^2$ where d is a distance or dissimilarity on $\mathcal{X}$, interpreted as a cost to move from (x_m, y_m) to (x'_n, θ_n);
- $\langle P, C(\theta) \rangle := \sum_{m \in [\![M]\!], n \in [\![N]\!]} P_{m,n} (C(\theta))_{m,n}$ is the P-specific average cost to transport $\{(x_m, y_m) : m \in [\![M]\!]\}$ onto $\{(x'_n, \theta_n) : n \in [\![N]\!]\}$;
- $E(P) := -\sum_{m \in [\![M]\!], n \in [\![N]\!]} P_{m,n} (\log P_{m,n} - 1)$ is the entropy of P;
- $\|\theta\|_1 := \sum_{n \in [\![N]\!]} |\theta_n|$ is the ℓ^1-norm of θ and a substitute for $\mathrm{card}\{n \in [\![N]\!] : \theta_n \neq 0\}$;
- $\gamma, \tau > 0$ are some user-supplied constants.

A few comments are in order.

1. The argmin is over $\mathbb{R}^N$ but could equivalently be over $[0,1]^N$. We thus view θ_n as the probability that the city described by x'_n will benefit from a declaration of natural disaster.

2. It is well known that substituting $\mathcal{W}_\gamma(C) := \min_{P\in\Pi} [\langle P, C\rangle - \gamma E(P)]$ for $\mathcal{W}_0(C) := \min_{P\in\Pi} [\langle P, C\rangle]$ has two important merits [6, Chapters 3, 4, 9]. First, $C \mapsto \mathcal{W}_0(C)$ is not differentiable whereas $C \mapsto \mathcal{W}_\gamma(C)$ is. Second, computing $\mathcal{W}_0(C)$ requires solving a costly linear program via network simplex methods whereas computing $\mathcal{W}_\gamma(C)$ can be performed easily thanks to the so-called Sinkhorn algorithm [2].

3. Adding the penalization term $+\tau\|\theta\|$ favors sparse solutions, which is in line with our prior knowledge that few cities will eventually indeed benefit from a declaration of natural disaster.

4 On solving (1)

Solving (1) is not straightforward, because the criterion to minimize is the sum of the non-convex differentiable function $f : \theta \mapsto \mathcal{W}_\gamma(C(\theta))$ and of the convex non-differentiable function $g : \theta \mapsto \tau\|\theta\|_1$. Luckily, we can rely on the so-called iPiano algorithm [5] which was developed precisely to deal with such optimization problems. The iPiano algorithm starts from an initial $\theta^0 \in]0,1[^N$ and the update scheme informally writes as (below, α, β are positive constants)

$$\theta^{k+1} = (I + \alpha\partial g)^{-1}\big(\theta^k - \alpha\nabla f(\theta^k) + \beta(\theta^k - \theta^{k-1})\big), \tag{2}$$

where the proximal map $(I + \alpha\partial g)^{-1}$ is defined by

$$(I + \alpha\partial g)^{-1}(t) := \operatorname*{arg\,min}_{\theta\in\mathbb{R}^N} \left\{ \tfrac{1}{2}\|\theta - t\|_2^2 + \alpha g(\theta) \right\}. \tag{3}$$

In our example, (3) is simply the soft-thresholding operator, given by

$$\big((I + \alpha\partial g)^{-1}(t)\big)_n = \operatorname{sign}(t_n)\left(|t_n| - \alpha\tau\right)_+.$$

Note that if $t \in [0,1]^N$ then $((I+\alpha\partial g)^{-1}(t))_n = (t_n - \alpha\tau)_+$ for every $n \in [\![N]\!]$. Moreover, following [3, Section 4.3], we show that the gradient of f is given by

$$\nabla f(\theta) = \tfrac{2}{N}\theta - \widehat{P}_\theta^\top y, \quad \text{with } \widehat{P}_\theta = \operatorname*{arg\,min}_{P\in\Pi} \left\{ \langle P, C(\theta)\rangle - \gamma E(P) \right\}.$$

We check that the assumptions of [5, Theorems 4.9 and 4.14] are met. Therefore we can assert that

- the sequence $(\theta^k)_{k\geq 0}$ converges to a critical point of $\theta \mapsto f(\theta) + g(\theta)$;

- $\min_{k \leq K} \|\theta^{k+1} - \theta^k\|_2^2 = O(K^{-1})$;
- if we set $r(\theta) := \theta - (I + \alpha \partial g)^{-1}(\theta - \alpha \nabla f(\theta))$, then $\min_{k \leq K} \|r(\theta^k)\|_2^2 = O(K^{-1})$.

The so-called proximal residual $r(\theta)$ is interesting because $r(\theta) = 0$ means that the first-order optimality condition is met at θ.

5 Application and perspectives

We already wrote a `python/pytorch` script to solve (1) and we are currently preparing the application. We will illustrate and comment on our results.

Carrying out the application is not easy. This is notably because *(i)* M and N are large ($M \approx 20 \times 36000 = 72 \times 10^4$ and $N \approx 36000$) and *(ii)* $\mathcal{X}$ is a high-dimensional space and the choice of the distance/dissimilarity d on $\mathcal{X}$ is difficult. As a consequence, we have to rely on a mini-batch procedure to evaluate the gradient of f. This poses both computational and theoretical challenges.

We also investigate a variant of (1) where we replace the penalization term $+\tau\|\theta\|_1$ by a constraint of the form $\theta \in B_{\|\cdot\|_1}(0, \tau)$. This constraint is quite interesting because it is easier to fine-tune the parameter τ, especially because experts at CCR have a rough idea of how many cities will benefit from a declaration of natural disaster.

References

[1] A. Charpentier, M. James, and H. Ali. Predicting drought and subsidence risks in france. *Natural Hazards and Earth System Sciences Discussion*, 2021. doi: 10.5194/nhess-2021-214. In review.

[2] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

[3] M. Cuturi and A. Doucet. Fast computation of wasserstein barycenters. In E. P. Xing and T. Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 685–693. PMLR, 22–24 Jun 2014.

[4] G. Ecoto, A. F. Bibaut, and A. Chambaz. One-step ahead sequential Super Learning from short times series of many slightly dependent data, and anticipating the cost of natural disasters. Technical report, 2021. URL `https://arxiv.org/abs/2107.13291`. Submitted.

[5] P. Ochs, T. Brox, and T. Pock. iPiasco: inertial proximal algorithm for strongly convex optimization. *J. Math. Imaging Vision*, 53(2):171–181, 2015.

[6] G. Peyré and M. Cuturi. Computational optimal transport, 2020.

Estimation adaptative optimale de fonctions moyennes et covariances irrégulières

Steven Golovkine[1] & Nicolas Klutchnikoff[2] & Valentin Patilea[3]

[1] *University of Limerick, steven.golovkine@ul.ie*
[2] *Université Rennes 2, nicolas.klutchnikoff@univ-rennes2.fr*
[3] *ENSAI, valentin.patilea@ensai.fr*

Résumé. Nous proposons des estimateurs non-paramétriques pour les fonctions moyenne et covariance de données fonctionnelles. Les courbes aléatoires ne sont pas nécessairement différentiables, de régularité inconnue et mesurées avec erreur sur un ensemble de points discret aléatoirement tirés. La définition de nos estimateurs non-paramétriques dépend de la régularité locale du processus stochastique générant les données. D'abord, nous proposons un estimateur simple de cette régularité locale utilisant les informations intra- et inter-courbes. Ensuite, l'approche "smoothing first, then estimate" est utilisée pour l'estimation des fonctions moyenne et covariance. Ces nouveaux estimateurs non-paramétriques atteignent des vitesses de convergence optimales.

Mots-clés. Données fonctionnelles, Exposant de Hölder, Polynômes locaux

Abstract. We propose nonparametric estimators for the mean and the covariance functions of functional data. The random trajectories are, not necessarily differentiable, have unknown regularity, and are measured with error at randomly drawn discrete design points. The definition of our nonparametric estimators depends on the local regularity of the stochastic process generating the functional data. We first propose a simple estimator of this local regularity which takes strength from the replication and regularization features of functional data. Next, we use the "smoothing first, then estimate" approach for the estimation of the mean and the covariance functions. The new nonparametric estimators achieve optimal rates of convergence.

Keywords. Functional data, Hölder exponent, Local polynomials

1 Introduction

Un intérêt croissant est porté sur la modélisation de données enregistrées sous forme de séquence de mesures discrète à travers le temps. L'analyse de données fonctionnelles (ADF) considère ces données comme des réalisations d'un processus stochastique, enregistré avec erreur à des instants aléatoires. Les fonctions moyenne et covariance jouent un rôle crucial en ADF.

Les données consistent en des réalisations d'un processus stochastique de second ordre $X = (X_t : t \in [0,1])$ ayant des trajectoires continues. Les fonctions moyenne et covariance sont définies, respectivement, par

$$\mu(t) = \mathbb{E}(X_t) \quad \text{et} \quad \Gamma(s,t) = \mathbb{E}\left\{[X_s - \mu(s)][X_t - \mu(t)]\right\}, \quad s,t \in [0,1].$$

Si des réalisations indépendantes $X^{(1)}, \ldots, X^{(N)}$ de X étaient observées sans bruit et de façon continues, les estimateurs idéaux seraient donnés par

$$\widetilde{\mu}(t) = \frac{1}{N}\sum_{i=1}^{N} X_t^{(i)} \quad \text{et} \quad \widetilde{\Gamma}(s,t) = \frac{1}{N-1}\sum_{i=1}^{N}\{X_s^{(i)} - \widetilde{\mu}(s)\}\{X_t^{(i)} - \widetilde{\mu}(t)\}, \quad s,t \in [0,1].$$

Dans le cas d'applications réelles, les courbes sont rarement observées sans erreur et jamais pour chaque $t \in [0,1]$. Pour chaque $1 \leq i \leq N$, soit M_i un entier positif et soit $T_m^{(i)} \in [0,1]$, $1 \leq m \leq M_i$, les temps d'échantillonnage aléatoires de la courbe $X^{(i)}$. Ainsi, nous observons les paires $(Y_m^{(i)}, T_m^{(i)}) \in \mathbb{R} \times [0,1]$ avec $Y_m^{(i)}$ défini comme

$$Y_m^{(i)} = X^{(i)}(T_m^{(i)}) + \epsilon_m^{(i)}, \qquad 1 \leq m \leq M_i, \;\; 1 \leq i \leq N, \tag{1}$$

et les $\epsilon_m^{(i)}$ sont des réalisations indépendantes de ϵ, une variable aléatoire de moyenne nulle et de variance σ^2.

L'idée est de calculer $\widetilde{\mu}(\cdot)$ et $\widetilde{\Gamma}(\cdot,\cdot)$ en utilisant des estimateurs non-paramétriques de $X_t^{(i)}$ et $X_s^{(i)}X_t^{(i)}$, comme celui par polynômes locaux. Cette approche est appelée "smoothing first, then estimate" [Hall *et al.* (2006), Zhang and Chen (2007)]. En général, une hypothèse de régularité est faite sur les fonctions moyenne et covariance du processus. Celles-ci doivent admettre des dérivées d'ordre 2 ou supérieure. Cependant, dans certaines applications, par exemple dans le domaine de l'énergie, la régularité est inconnue et cette hypothèse peut sembler irréaliste. Nous proposons donc des estimateurs non-paramétriques des fonctions moyenne et covariance d'un processus de régularité inconnue atteignant des vitesses de convergence optimale pour le risque minimax, basé sur l'estimation de la régularité locale du processus.

2 Estimation locale de la régularité

Soit t_0 un point de $[0,1]$ et soit $\mathcal{O}_*$ un voisinage de t_0. Pour $H_{t_0} \in (0,1]$, nous faisons l'hypothèse que le processus X vérifie la condition suivante :

$$\mathbb{E}\left((X_u - X_v)^2\right) \asymp L^2|v-u|^{2H_{t_0}}, \quad \text{avec } u \text{ et } v \in \mathcal{O}_*. \tag{2}$$

La quantité H_{t_0} correspond à la *régularité locale* du processus X sur $\mathcal{O}_*$. Elle est associée à la vitesse de décroissance des valeurs propres ordonnées de l'opérateur de covariance

du processus. En effet, en notant $\lambda_j, j \geq 1$, ces valeurs propres ordonnées, alors pour $\nu > 1, \lambda_j \sim j^{-\nu}$, et nous avons $2H = \nu - 1$. Nous présentons ensuite les principales idées conduisant à la construction de l'estimateur local de la régularité.

Pour $u, v \in \mathcal{O}_*$, notons

$$\theta(u,v) = \mathbb{E}\left[(X_u - X_v)^2\right] \approx L^2|v-u|^{2H_{t_0}} \quad \text{si } v-u \text{ est petit.}$$

Considérons t_1 et t_3 tels que $[t_1, t_3] \subset \mathcal{O}_*$ et notons t_2 le point central de $[t_1, t_3]$. On remarque facilement que

$$H_{t_0} \approx \frac{\log\big(\theta(t_1,t_3)\big) - \log\big(\theta(t_1,t_2)\big)}{2\log(2)} \quad \text{si } t_3 - t_1 \text{ est petit.}$$

Considérons un échantillon d'apprentissage de N courbes, générées suivant (1). Soit $\widetilde{X}_u$, un estimateur non-paramétrique de X_u, pour $u \in \mathcal{O}_*$. Nous définissons un estimateur naturel de H_{t_0} par

$$\widehat{H}_{t_0} = \frac{\log\big(\widehat{\theta}(t_1,t_3)\big) - \log\big(\widehat{\theta}(t_1,t_2)\big)}{2\log(2)}, \quad \text{où} \quad \widehat{\theta}(u,v) = \frac{1}{N}\sum_{i=1}^{N}\left(\widetilde{X}_u^{(i)} - \widetilde{X}_v^{(i)}\right)^2.$$

En particulier, sous certaines conditions sur le processus X, sur son estimateur non-paramétrique $\widetilde{X}$, et, en notant $\mathfrak{m} = \mathbb{E}(M_i)$, sur les vitesses de croissance de N et $\mathfrak{m}$, nous montrons que, pour $\mathfrak{m}$ suffisamment grand,

$$\mathbb{P}\left(|\widehat{H}_{t_0} - H_{t_0}| > \log^{-2}(\mathfrak{m})\right) \leq \exp(-\mathfrak{m}).$$

Le résultat non-asymptotique sur la régularité locale de X permet de construire une fenêtre ayant une vitesse de convergence optimale pour l'estimation des fonctions moyenne et covariance.

3 Estimation des fonctions moyenne et covariance

Nous expliquons maintenant comment selectionner une fenêtre ayant une vitesse de convergence optimale pour le lissage par polynômes locaux des trajectoires, et ensuite construire les estimateurs des fonctions moyenne et covariance.

Pour chaque $1 \leq i \leq N$, nous utilisons une approche par polynômes locaux pour contruire des estimateurs non-paramétrique $\widehat{X}^{(i)}$ de $X^{(i)}$. Pour $t \in [0,1]$, si la régularité locale H_t est connue, alors en utilisant les mesures $(Y_m^{(i)}, T_m^{(i)})$, $1 \leq m \leq M_i$, d'une trajectoire $X^{(i)}$, nous considérons l'estimateur défini par

$$\widehat{X}_t^{(i)} = \widehat{X}_t^{(i)}(h_t) = \sum_{m=1}^{M_i} Y_m^{(i)} W_m^{(i)}(t), \qquad 1 \leq i \leq N,$$

avec h_t une fenêtre dépendante de H_t, et $W_m^{(i)}(t)$ des poids dépendants de h_t.

Soit k_0 un entier. Notons $\mathbf{1}\{\cdot\}$ la fonction indicatrice. Pour $t \in [0,1]$, définissons $\mathcal{W}(t;h) = \sum_{i=1}^N w_i(t;h)$ et $\mathcal{W}(s,t;h) = \sum_{i=1}^N w_i(s;h)w_i(t;h)$ tels que

$$w_i(t;h) = 1 \quad \text{si} \quad \sum_{m=1}^{M_i} \mathbf{1}\{|T_m^{(i)} - t| \leq h\} \geq k_0, \qquad \text{et} \quad w_i(t;h) = 0 \text{ sinon.}$$

L'estimateur adaptatif de la fonction moyenne est donné par $\widehat{\mu}^*(t) = \widehat{\mu}(t;h_\mu)$ avec

$$\widehat{\mu}(t;h) = \mathcal{W}(t;h)^{-1} \sum_{i=1}^N w_i(t;h) X_t^{(i)}, \quad t \in [0,1]. \tag{3}$$

L'estimateur $\widehat{\mu}(t;h)$ est un estimateur de la moyenne prenant en compte le fait qu'il puisse y avoir moins de k_0 observations entre $t - h_\mu$ et $t + h_\mu$ pour certaines trajectoires. Nous définissons la fenêtre optimale pour calculer $\widehat{\mu}^*(t)$ telle que la différence au carré entre $\widehat{\mu}(t;h)$ et $\widetilde{\mu}(t)$ soit minimisé. La fenêtre optimale est donc définie par

$$h_\mu(t) = \arg\min_{h>0} \mathcal{R}_\mu(t;h), \quad \text{avec} \quad \mathcal{R}_\mu(t;h) = q_1^2 h^{2\hat{H}_t} + \frac{q_2^2}{\mathcal{N}_\mu(t;h)} + q_3^2 \left[\frac{1}{\mathcal{W}(t;h)} - \frac{1}{N}\right],$$

et q_1, q_2 et q_3 des constantes dépendantes de la variance du processus au point t, de la variance de ϵ, de la constante L de l'équation (2), du noyau utilisé et tel que

$$\mathcal{N}_\mu(t;h) = \left[\frac{1}{\mathcal{W}^2(t;h)} \sum_{i=1}^N \frac{w_i(t;h)}{\mathcal{N}_i(t;h)} \sum_{m=1}^{M_i} |W_m^{(i)}(t;h)|\right]^{-1} \text{ avec } \mathcal{N}_i(t;h) = \frac{w_i(t;h)}{\max\limits_{1\leq m\leq M_i} |W_m^{(i)}(t;h)|}.$$

Nous montrons que $\mathcal{R}_\mu(t;h)$ est une borne de $\mathbb{E}\left[\{\widehat{\mu}(t;h) - \widetilde{\mu}(t)\}^2\right]$. La minimisation de $\mathcal{R}_\mu(t;h)$ peut se faire sur une grille de valeurs pour h. En particulier, sous certaines conditions sur la densité des points d'observations $T_m^{(i)}$, sur la vitesse de convergence de H_t et sur les vitesses de croissance de N et $\mathfrak{m}$, nous montrons que, pour $t \in [0,1]$,

$$\widehat{\mu}^*(t) - \widetilde{\mu}(t) = O_{\mathbb{P}}\left((N\mathfrak{m})^{-\frac{H_t}{1+2H_t}}\right) \quad \text{et} \quad \widehat{\mu}^*(t) - \mu(t) = O_{\mathbb{P}}\left((N\mathfrak{m})^{-\frac{H_t}{1+2H_t}} + N^{-1/2}\right).$$

En ce qui concerne la fonction de covariance, la fonction variance $\Gamma(s,s)$ induit une singularité lors de l'estimation de la covariance $\Gamma(\cdot,\cdot)$, nous distinguons donc les points sur la diagonale de ceux en dehors [Zhang and Wang (2016)]. Pour les points hors-diagonale, de façon similaire à l'estimation de la moyenne, l'estimateur adaptatif de la fonction covariance est donné par

$$\widehat{\Gamma}^*(s,t) = \widehat{\Gamma}(s,t;h_\Gamma) \quad \text{avec} \quad \widehat{\Gamma}(s,t;h) = \widehat{\gamma}(s,t;h) - \widehat{\mu}^*(s)\widehat{\mu}^*(t), \tag{4}$$

où $\widehat{\mu}^*(s)$ et $\widehat{\mu}^*(t)$ sont définis suivant (3) avec les fenêtres correspondantes, et

$$\widehat{\gamma}(s,t;h) = \mathcal{W}(s,t;h)^{-1} \sum_{i=1}^{N} w_i(s;h) w_i(t;h) \widehat{X}_s^{(i)} \widehat{X}_t^{(i)}, \quad s,t \in [0,1], s \neq t. \tag{5}$$

Nous définissons la fenêtre optimale pour calculer $\widehat{\gamma}(s,t;h)$ telle que la différence au carré entre $\widehat{\gamma}(s,t;h)$ et $\widetilde{\gamma}(s,t)$, l'estimateur de $\mathbb{E}(X_s^{(i)} X_t^{(i)})$, soit minimisé. La fenêtre optimale est donc définie par

$$h_\Gamma = h_\Gamma(s,t) = \arg\min_{h>0} \{\mathcal{R}_\Gamma(s|t;h) + \mathcal{R}_\Gamma(t|s;h)\}, \tag{6}$$

avec

$$\mathcal{R}_\Gamma(t|s;h) = q_1^2 h^{2\widehat{H}_t} + \frac{q_2^2}{\mathcal{N}_\Gamma(t|s;h)} + q_3^2 \left[\frac{1}{W(s,t;h)} - \frac{1}{N}\right]. \tag{7}$$

et q_1, q_2 et q_3 des constantes dépendantes de différents moments du processus, de la variance de ϵ, de la contante L de l'équation (2), du noyau utilisé et tel que

$$\mathcal{N}_\Gamma(t|s;h) = \left[\frac{1}{\mathcal{W}^2(s,t;h)} \sum_{i=1}^{N} \frac{w_i(s;h) w_i(t;h)}{\mathcal{N}_i(t|s;h)} \sum_{m=1}^{M_i} |W_m^{(i)}(t;h)|\right]^{-1}$$

avec $\mathcal{N}_i(t|s;h) = w_i(s;h) w_i(t;h) \min_{1\leq m\leq M_i} |W_m^{(i)}(t,h)|^{-1}$. La définition de $\mathcal{R}_\Gamma(s|t;h)$ est la même que dans l'équation (7) avec s et t échangés. Nous montrons la fonction de h minimisée en (6) est une borne de $\mathbb{E}\left[\{\widehat{\gamma}(s,t;h) - \widetilde{\gamma}(s,t)\}^2\right]$. La minimisation de $\mathcal{R}_\Gamma(s|t;h) + \mathcal{R}_\Gamma(t|s;h)$ peut se faire sur une grille de valeurs pour h. En particulier, en notant $H(s,t) = \min\{H_s, H_t\}$ et $H(s,t) < 1$, sous certaines conditions sur la densité des points d'observations $T_m^{(i)}$, sur la vitesse de convergence de H_t, sur les vitesses de croissance de N et $\mathfrak{m}$ et sur les moments du processus X, nous montrons que, pour $s,t \in [0,1], s \neq t$,

$$\widehat{\Gamma}^*(s,t) - \widetilde{\Gamma}(s,t) = O_{\mathbb{P}}\left((N\mathfrak{m})^{-\frac{H(s,t)}{1+2H(s,t)}}\right)$$

et

$$\widehat{\Gamma}^*(s,t) - \Gamma(s,t) = O_{\mathbb{P}}\left((N\mathfrak{m})^{-\frac{H(s,t)}{1+2H(s,t)}} + N^{-1/2}\right).$$

Nous proposons d'utiliser cet estimateur (5) seulement en dehors de la bande diagonale définie par $\{(s,t) : |s-t| \leq \mathfrak{d}\}$, pour $\mathfrak{d} > 0$.

Il nous reste à donner un estimateur de cette bande diagonale pour la fonction de covariance. Pour cela, nous proposons une règle basée sur les données pour le choix de $\mathfrak{d}$ et ensuite donner un estimateur de $\mathbb{E}(X_s X_t)$ quand s et t sont sur la diagonale. Notons H la régularité locale au point $u = (s+t)/2$. Sous certaines hypothèses sur les moments de X_t et de $X_t - X_s$, nous montrons que

$$\int\int_{t-\mathfrak{d}\leq s\leq t} \left\{\widetilde{\Gamma}(u-\mathfrak{d}/2, u+\mathfrak{d}/2) - \widetilde{\Gamma}(s,t)\right\}^2 dsdt = O_{\mathbb{P}}\left(\mathfrak{d}^{2H+1}\right).$$

Pour (s,t) dans la bande diagonale, nous pouvons prendre $\widehat{\Gamma}(u-\mathfrak{d}/2, u+\mathfrak{d}/2)$ définie par l'équation (4) quand $s \leq t$ et utiliser la symétrie de la covariance pour $s > t$. Si $\widehat{H}$ est un estimateur de H, nous pouvons, par exemple, prendre

$$\mathfrak{d} = \left\{ N^{-2} \sum_{i=1}^{N} (1/M_i) \right\}^{c} \quad \text{avec} \quad c = \frac{2\widehat{H}+1/2}{(2\widehat{H}+1)^2}.$$

Nous avons étudié la qualité de ces estimateurs de la fonction moyenne et covariance par rapport à deux estimateurs classiques, des estimateurs de lissage par splines sur les fonctions moyenne et covariance empiriques [Cai et Yuan (2010, 2011)] et des estimateurs par polynômes locaux pondérés nécessitant l'ensemble des courbes [Zhang and Wang (2016)]. Nous obtenons de bons résultats au regard de l'erreur quadratique intégrée avec des temps de calculs raisonnables.

La méthode peut s'étendre aux cas où la régularité est supérieur à 1 en estimant la régularité des dérivées successives. Enfin, le cadre peut aussi s'élargir aux cas où la variance du bruit n'est plus supposée constante mais d'esperance contionnnelle constante, ce qui permet de considérer un bruit hétéroscédastique. Une version complète de l'article est disponible à l'adresse suivante : arxiv:2108.06507.

Bibliographie

Cai, T. and Yuan, M. (2010) *Nonparametric covariance function estimation for functional and longitudinal data.* University of Pennsylvania and Georgia Institute of Technology.

Cai, T. T. and Yuan, M. (2011) *Optimal estimation of the mean function based on discretely sampled functional data: Phase transition. Ann. Statist.*, 39, 2330–2355.

Hall, P., Müller, H.-G. and Wang, J.-L. (2006) *Properties of principal component methods for functional and longitudinal data analysis. Ann. Statist.*, 34, 1493–1517.

Zhang, J.-T. and Chen, J. (2007) *Statistical inferences for functional data. Ann. Statist.*, 35, 1052–1079.

Zhang, X. and Wang, J.-L. (2016) *From sparse to dense functional data and beyond. Ann. Statist.*, 44, 2281–2321.

Réduction d'une fonction aléatoire cyclostationnaire définie sur $\mathbb{R}$

Alain Boudou [1,†] & Sylvie Viguier-Pla [2,1,‡]

[1] *Institut de Mathématiques de Toulouse, Université Paul Sabatier, 31062 Toulouse Cedex, France,* [†]*palain@gmail.com*

[2] *Laboratoire de Modélisation Pluridisciplinaire et Simulations, Université de Perpignan via Domitia, 66000 Perpignan, France,* [‡]*viguier@univ-perp.fr*

Résumé. Considérant une fonction aléatoire cyclostationnaire, nous proposons une transformation pour en déduire une série multidimensionnelle stationnaire, dont on peut éventuellement faire une analyse en composantes principales. Nous définissons ce qu'est une série résumé de plus faible dimension, et nous en étudions une particulière.

Mots-clés. Analyse en composantes principales, cyclostationnarité, opérateur unitaire, série stationnaire.

Abstract. Considering a cyclostationary random function, we propose a method for its conversion into a multidimensional stationary series, for which a principal components analysis can be processed. We then define what is a lower-dimensional summary series, and we study a particular one.

Keywords. Principal components analysis, cyclostationarity, unitary operator, stationary series.

1 Introduction

La cyclostationnarité, encore nommée corrélation périodique, permet de modéliser de phénomènes naturels pour lesquels certaines statistiques sont périodiques.

Dans Boudou and Viguier-Pla (2021), nous étudions des fonctions aléatoires (f.a.) cyclostationnaires définies comme des applications de $\mathbb{R}^k$ dans un espace de variables aléatoires (v.a.) de type $L^2(\Omega, \mathcal{A}, P)$ (que nous noterons pour simplifier $L^2(\mathcal{A})$). A une telle fonction nous associons une série stationnaire. On peut ainsi procéder à une Analyse en Composantes Principales (ACP) dans le domaine des fréquences de cette dernière, et donc d'une certaine façon à une analyse de la f.a. cyclostationnaire. Ici, nous considérons et approfondissons le cas particulier où $k = 1$. Après avoir précisé la définition d'un résumé $p-$dimensionnel d'une série, l'évaluation de la qualité d'un tel résumé, et l'ACP dans le domaine des fréquences, nous rappelons comment on déduit une série stationnaire d'une f.a. cyclostationnaire. Ensuite, nous étudions les qualités d'un résumé $p-$dimensionnel que l'on peut qualifier de naturel. A l'occasion de l'exposé, nous présenterons une simulation de ces méthodes.

Les v.a. que nous considérons peuvent être supposées centrées, ainsi la covariance devient un produit scalaire. Cette hypothèse de centrage est justifiée du point de vue statistique mais n'est pas nécessaire pour les développements mathématiques.

2 Résumé $p-$dimensionnel d'une série stationnaire

Désormais, $(\Omega, \mathcal{A}, P)$ est un espace probabilisé tel que le $\mathbb{C}-$espace de Hilbert $L^2(\mathcal{A})$ soit séparable.

Une série $(Y_n)_{n\in\mathbb{Z}}$ d'éléments de $L^2_H(\mathcal{A})$ (H étant un $\mathbb{C}-$espace de Hilbert séparable) est stationnaire lorsque $E(Y_n \otimes Y_m) = E(Y_{n-m} \otimes Y_0)$, pour tout couple (n, m) d'éléments de $\mathbb{Z}$ (si H est de dimension finie, cela peut matriciellement s'écrire $EY_n {}^t\overline{Y_m} = EY_{n-m} {}^t\overline{Y_0}$). Nous avons ainsi une famille de séries coordonnées stationnaires et stationnairement corrélées deux à deux.

Étant donnée une série $p-$dimensionnelle stationnaire $(Y'_n)_{n\in\mathbb{Z}}$ stationnairement corrélée à $(Y_n)_{n\in\mathbb{Z}}$ (c'est-à-dire telle que $E(Y_n \otimes Y'_m) = E(Y_{n-m} \otimes Y'_0)$, pour tout couple (n, m) d'éléments de $\mathbb{Z}$), elle peut être considérée comme un résumé $p-$dimensionnel de $(Y_n)_{n\in\mathbb{Z}}$. La qualité de ce résumé peut être évaluée de la façon suivante. Si l'on note $P_{Y'}$ le projecteur de $L^2_H(\mathcal{A})$ sur $\overline{\text{vect}}\{K \circ Y'_n; (n, K) \in \mathbb{Z} \times L(\mathbb{C}^p, H)\}$, où $L(\mathbb{C}^p, H)$ est l'espace vectoriel des applications linéaires de $\mathbb{C}^p$ dans H, on montre que $(P_{Y'}Y_n)_{n\in\mathbb{Z}}$ est une série stationnaire stationnairement corrélée à $(Y_n)_{n\in\mathbb{Z}}$. Cette dernière propriété permet d'écrire $\|Y_n - P_{Y'}Y_n\| = \|Y_0 - P_{Y'}Y_0\|$, quantité qui sera retenue comme la mesure de la qualité du résumé $p-$dimensionnel de $(Y_n)_{n\in\mathbb{Z}}$ qu'est $(Y'_n)_{n\in\mathbb{Z}}$. D'où les définitions.

Définition 2.1. *Une série d'éléments de $L^2_{\mathbb{C}^p}(\mathcal{A})$, $(Y'_n)_{n\in\mathbb{Z}}$, est un résumé $p-$dimensionnel de $(Y_n)_{n\in\mathbb{Z}}$ lorsqu'elle est stationnaire et stationnairement corrélée à $(Y_n)_{n\in\mathbb{Z}}$. La qualité de ce résumé $p-$dimensionnel est mesurée par la quantité $\|Y_0 - P_{Y'}Y_0\|$.*

L'ACP dans le domaine des fréquences de $(Y_n)_{n\in\mathbb{Z}}$ est la recherche du meilleur résumé $p-$dimensionnel.

Définition 2.2. *Les p premières étapes de l'ACP dans le domaine des fréquences d'une série stationnaire $(Y_n)_{n\in\mathbb{Z}}$ sont la recherche d'un résumé $p-$dimensionnel $(Y'_n)_{n\in\mathbb{Z}}$, de telle sorte que $\|Y_0 - P_{Y'}Y_0\|$ soit le plus petit possible.*

Pour plus de développements concernant cette analyse, on pourra consulter Brilliger (2001) et Boudou and Dauxois (1994).

La solution $(Y'_n)_{n\in\mathbb{Z}}$ est un filtre de $(Y_n)_{n\in\mathbb{Z}}$ de la forme $(\sum_m A_m \circ Y_{n-m})_{n\in\mathbb{Z}}$. Soit $\{Z_j; j \in J\}$ une famille finie d'éléments de $L^2_H(\mathcal{A})$ telle que $E(Z_j \otimes Z_{j'}) = 0$, pour tout couple (j, j') d'éléments distincts de J. On peut vérifier que $(\sum_{j\in J} \mathrm{e}^{i\lambda_j n} Z_j)_{n\in\mathbb{Z}}$ est une série stationnaire dont l'ACP dans le domaine des fréquences revient à effectuer l'ACP de chacun des vecteurs aléatoires Z_j, d'où l'appellation d'ACP dans le domaine des fréquences.

3 La cyclostationnarité

Classiquement, une f.a. cyclostationnaire est une famille $\{x_t; t \in \mathbb{R}\}$ d'éléments de $L^2(\mathcal{A})$ telle que $\langle x_t, x_{t'}\rangle = \langle x_{t+\Delta}, x_{t'+\Delta}\rangle$, pour tout couple (t, t') de réels. Ce n'est pas restreindre la généralité que de considérer $\Delta = 1$, on peut toujours se ramener à ce cas de figure par un changement de variable.

Désormais, μ est une mesure de probabilité définie sur ξ, tribu de parties de $[0; 1[$, trace de $\mathcal{B}_{\mathbb{R}}$ sur $[0; 1[$. On peut montrer que $L^2([0; 1[, \xi, \mu)$ (que nous noterons pour simplifier $L^2(\xi)$) est un $\mathbb{C}-$espace de Hilbert séparable. Précisons le sens que l'on donne au terme f.a. cyclostationnaire dans ce texte.

Définition 3.1. *Une f.a. cyclostationnaire $\{x_t; t \in \mathbb{R}\}$ est une famille d'éléments de $L^2(\mathcal{A})$ telle que*
i) $\langle x_t, x_{t'}\rangle = \langle x_{t+1}, x_{t'+1}\rangle$, pour tout couple (t, t') ;
ii) l'application $t \in [0; 1[\mapsto x_t \in L^2(\mathcal{A})$ est mesurable et de norme carrée $\mu-$intégrable.

Notons que le point *ii)* n'est pas trop contraignant, il est vérifié dès lors que l'application $t \in [0; 1[\mapsto x_t \in L^2(\mathcal{A})$ est continue. Remarquons qu'une f.a. stationnaire est une f.a. cyclostationnaire et se prête donc aux traitements que nous proposons dans ce texte.

4 Série stationnaire déduite d'une fonction aléatoire cyclostationnaire

Dans ce paragraphe, nous construisons une série stationnaire à partir de la f.a. cyclostationnaire $\{x_t; t \in \mathbb{R}\}$. L'outil mathématique utilisé est développé dans Boudou and Viguier (2021). Nous allons le rappeler succinctement. Il se trouve simplifié du fait que $k = 1$.

Pour tout X (resp. Y) de $L^2_{L^2(\mathcal{A})}(\xi)$ (resp. $L^2_{L^2(\xi)}(\mathcal{A})$), on note $\widetilde{X}$ (resp. $\widetilde{Y}$) l'opérateur de Hilbert-Schmidt défini par $\widetilde{X} : h \in L^2(\xi) \mapsto \int h(\lambda)X(\lambda)\mathrm{d}\mu(\lambda) \in L^2(\mathcal{A})$ (resp. $\widetilde{Y} : y \in L^2(\mathcal{A}) \mapsto \int y(\omega)Y(\omega)\mathrm{d}P(\omega) \in L^2(\xi)$). On montre que $E(Y_1 \otimes Y_2) = \widetilde{Y_2} \circ \widetilde{Y_1}^*$, pour tout couple (Y_1, Y_2) d'éléments de $L^2_{L^2(\xi)}(\mathcal{A})$. Il est bien connu qu'il existe une isométrie, et une seule, $\mathcal{I}$, de $L^2_{L^2(\mathcal{A})}(\xi)$ sur $L^2_{L^2(\xi)}(\mathcal{A})$, telle que $\mathcal{I}(h.y) = y.h$, pour tout couple (y, h) de $L^2(\mathcal{A}) \times L^2(\xi)$.

Pour tout X de $L^2_{L^2(\mathcal{A})}(\xi)$, on a alors $\gamma \circ \widetilde{X} \circ \Gamma = \widetilde{\mathcal{I}X}^*$, où γ (resp. Γ) est l'application antilinéaire qui conserve la norme définie par $y \in L^2(\mathcal{A}) \mapsto \overline{y} \in L^2(\mathcal{A})$ (resp. $h \in L^2(\xi) \mapsto \overline{h} \in L^2(\xi)$).

Du point i) de la définition de la cyclostationnarité, nous déduisons l'existence d'un opérateur unitaire U de $L^2(\mathcal{A})$ tel que $U^n x_t = x_{t+n}$, pour tout (t, n) de $\mathbb{R} \times \mathbb{Z}$.

L'application $t \in [0; 1[\mapsto x_{t+n} \in L^2(\mathcal{A})$, donc l'application $U^n \circ (t \in [0; 1[\mapsto x_t \in L^2(\mathcal{A}))$ est mesurable (U^n est continu donc mesurable) et de norme carrée $\mu-$intégrable (U^n conserve la norme), c'est donc un élément X_n de $L^2_{L^2(\mathcal{A})}(\xi)$.

On a alors $\widetilde{X_n}h = \int h(\lambda)x_{\lambda+n}\mathrm{d}\mu(\lambda) = \int h(\lambda)U^n x_\lambda \mathrm{d}\mu(\lambda) = U^n \int h(\lambda)x_\lambda \mathrm{d}\mu(\lambda) = U^n\widetilde{X_0}h$, d'où, puisque cela est exact pour tout h de $L^2(\xi)$, $\widetilde{X_n} = U^n\widetilde{X_0}$. Il vient alors $E\mathcal{I}X_n \otimes \mathcal{I}X_m = \widetilde{\mathcal{I}X_m}\widetilde{\mathcal{I}X_n}^* = (\gamma \circ \widetilde{X_m} \circ \Gamma)^* \circ \gamma \circ \widetilde{X_n} \circ \Gamma = \Gamma \circ \widetilde{X_m}^* \widetilde{X_n} \circ \Gamma = \Gamma \circ \widetilde{X_0}^* \circ U^{-m} \circ U^n \circ \widetilde{X_0} \circ \Gamma = \Gamma \circ \widetilde{X_0}^* \circ U^{n-m} \circ \widetilde{X_0} \circ \Gamma = E\mathcal{I}X_{n-m} \otimes \mathcal{I}X_0$, pour tout couple (n,m) d'éléments de $\mathbb{Z}$.

La série d'éléments de $L^2_{L^2(\xi)}(\mathcal{A})$, $(\mathcal{I}X_n)_{n\in\mathbb{Z}}$, est donc stationnaire, c'est cette série que l'on appelle série stationnaire déduite de la f.a. cyclostationnaire $\{x_t; t \in \mathbb{R}\}$.

5 Étude de la série $\left((x_n, x_{n+\frac{1}{p}}, \ldots, x_{n+\frac{p-1}{p}})\right)_{n\in\mathbb{Z}}$

Etant donnée une f.a. cyclostationnaire $\{x_t; t \in \mathbb{R}\}$, il est naturel, suite à une partition de $[n; n+1[$ en p segments égaux, de considérer la série d'éléments de $L^2_{\mathbb{C}^p}(\mathcal{A})$, $(Y'_n)_{n\in\mathbb{Z}}$, où $Y'_n = \sum_{j=1}^p x_{n+\frac{j-1}{p}}.e_j$, $\{e_1, \ldots, e_p\}$ étant la base canonique de $\mathbb{C}^p$.

Cette série $p-$dimensionnelle est stationnaire, en effet, reprenant les notations du paragraphe précédent, on a

$\widetilde{Y'_n} = \sum_{j=1}^p (\gamma x_{n+\frac{j-1}{p}}) \otimes e_j = \sum_{j=1}^p (\gamma U^n x_{\frac{j-1}{p}}) \otimes e_j$

$= \sum_{j=1}^p ((\gamma x_{\frac{j-1}{p}}) \otimes e_j) \circ \gamma \circ U^{-n} \circ \gamma = \widetilde{Y'_0} \circ \gamma \circ U^{-n} \circ \gamma$

et donc, pour tout couple (n,m) d'éléments de $\mathbb{Z}$,

$EY'_m \otimes Y'_n = \widetilde{Y'_n} \circ \widetilde{Y'_m}^* = \widetilde{Y'_0} \circ \gamma \circ U^{-n} \circ U^m \circ \gamma \circ \widetilde{Y'_0}^* = \widetilde{Y'_0} \circ \gamma \circ U^{m-n} \circ \gamma \circ \widetilde{Y'_0}^* = EY'_{m-n} \otimes Y'_0$.

Elle est également stationnairement corrélée à $(\mathcal{I}X_n)_{n\in\mathbb{Z}}$ (série stationnaire déduite de $\{x_t; t \in \mathbb{R}\}$). Pour tout (n,m) de $\mathbb{Z}\times\mathbb{Z}$:

$E\mathcal{I}X_m \otimes Y'_n = \widetilde{Y'_n} \circ \widetilde{\mathcal{I}X_m}^* = \widetilde{Y'_0} \circ \gamma \circ U^{-n} \circ \widetilde{X_m} \circ \Gamma = \widetilde{Y'_0} \circ \gamma \circ U^{m-n} \circ \widetilde{X_0} \circ \Gamma = E\mathcal{I}X_{m-n} \otimes Y'_0$.

Compte-tenu de ces propriétés de stationnarité, la série $(Y'_n)_{n\in\mathbb{Z}}$ peut être considérée comme un résumé $p-$dimensionnel de $(\mathcal{I}X_n)_{n\in\mathbb{Z}}$, donc de la f.a. cyclostationnaire $\{x_t; t \in \mathbb{R}\}$. C'est la raison pour laquelle nous consacrons ce paragraphe à son étude.

Examinons quelques résultats préliminaires.

Bien entendu, $P_{Y'}$ désigne le projecteur de $L^2_{L^2(\xi)}(\mathcal{A})$ sur $\overline{\text{vect}}\{K \circ Y'_n; (n,K) \in \mathbb{Z} \times L(\mathbb{C}^p, L^2(\xi))\}$. Si l'on note $\mathcal{L}$ l'isométrie définie par $Y \in L^2_{L^2(\xi)}(\mathcal{A}) \mapsto \widetilde{Y} \in \sigma_2(L^2(\mathcal{A}), L^2(\xi))$, les propriétés d'identification de l'isométrie d'un espace à un autre permettent d'écrire, avec des notations évidentes :

$P_{Y'} = \mathcal{L}^{-1} \circ P_{\mathcal{L}(\overline{\text{vect}}\{K\circ Y'_n;(n,K)\in\mathbb{Z}\times L(\mathbb{C}^p,L^2(\xi))\})} \circ \mathcal{L}$.

Si l'on remarque que $\mathcal{L}(\overline{\text{vect}}\{K\circ Y'_n; (n,K) \in \mathbb{Z}\times L(\mathbb{C}^p, L^2(\xi))\}) = \mathcal{L}(\overline{\text{vect}}\{x_{\frac{n}{p}}.h; (n,h) \in \mathbb{Z} \times L^2(\xi)\}) = \overline{\text{vect}}\{\gamma x_{\frac{n}{p}} \otimes h; (n,h) \in \mathbb{Z} \times L^2(\xi)\}$,

ce qui précède devient

$P_{Y'} = \mathcal{L}^{-1} \circ P_{\overline{\text{vect}}\{\gamma x_{\frac{n}{p}} \otimes h;(n,h)\in\mathbb{Z}\times L^2(\xi)\}} \circ \mathcal{L}$.

Si l'on note D le projecteur de $L^2(\mathcal{A})$ sur $\overline{\text{vect}}\{x_{\frac{n}{p}}; n \in \mathbb{Z}\}$, on peut vérifier que $\gamma \circ D \circ \gamma$ est le projecteur de $L^2(\mathcal{A})$ sur $\overline{\text{vect}}\{\gamma x_{\frac{n}{p}}; n \in \mathbb{Z}\}$. On peut alors énoncer le

résultat préliminaire suivant.

Lemme 5.1. *Pour tout Y de $L^2_{L^2(\xi)}(\mathcal{A})$, $\widetilde{P_{Y'}Y} = \widetilde{Y} \circ \gamma \circ D \circ \gamma$.*

Démonstration. Nous rappelons que, si H_1 et H_2 sont deux $\mathbb{C}-$espaces de Hilbert séparables, $\{h_\lambda; \lambda \in \Lambda\}$ une famille d'éléments de H_1 et T le projecteur de H_1 sur $\overline{\text{vect}}\{h_\lambda; \lambda \in \Lambda\}$, alors la projection d'un élément K de $\sigma_2(H_1, H_2)$ sur $\overline{\text{vect}}\{h_\lambda \otimes h; (\lambda, h) \in \Lambda \times H_2\}$ est $K \circ T$.

Donc $\widetilde{Y} \circ \gamma \circ D \circ \gamma$ est la projection de $\widetilde{Y} = \mathcal{L}(Y)$ sur $\overline{\text{vect}}\{\gamma x_{\frac{n}{p}} \otimes h; (n, h) \in \mathbb{Z} \times L^2(\xi)\}$. Comme $P_{Y'} = \mathcal{L}^{-1} \circ P_{\overline{\text{vect}}\{\gamma x_{\frac{n}{p}} \otimes h;(n,h)\in\mathbb{Z}\times L^2(\xi)\}} \circ \mathcal{L}$, donc comme $\mathcal{L}P_{Y'} = P_{\overline{\text{vect}}\{\gamma x_{\frac{n}{p}} \otimes h;(n,h)\in\mathbb{Z}\times L^2(\xi)\}} \circ \mathcal{L}$, il vient $\widetilde{P_{Y'}Y} = \mathcal{L}P_{Y'}Y = P_{\overline{\text{vect}}\{\gamma x_{\frac{n}{p}} \otimes h;(n,h)\in\mathbb{Z}\times L^2(\xi)\}}\widetilde{Y}$.

Donc, compte-tenu de ce qui précède, $\widetilde{P_{Y'}Y} = \widetilde{Y} \circ \gamma \circ D \circ \gamma$. □

Nous avons maintenant les outils nécessaires pours établir une relation entre la mesure de la qualité du résumé $p-$dimensionnel $(Y'_n)_{n\in\mathbb{Z}}$ et le projecteur sur $\overline{\text{vect}}\{x_{\frac{n}{p}}; n \in \mathbb{Z}\}$.

Proposition 5.1. *Pour tout n de $\mathbb{Z}$, on a $\|\mathcal{I}X_n - P_{Y'}\mathcal{I}X_n\|^2 = \int \|x_t - Dx_t\|^2_{L^2(\mathcal{A})} d\mu(t)$.*

Démonstration. Cela est une conséquence du lemme précédent.

$\|\mathcal{I}X_n - P_{Y'}\mathcal{I}X_n\|^2 = \|\mathcal{I}X_0 - P_{Y'}\mathcal{I}X_0\|^2 = \|\widetilde{\mathcal{I}X_0} - \widetilde{P_{Y'}\mathcal{I}X_0}\|^2 = \|\widetilde{\mathcal{I}X_0} - \widetilde{\mathcal{I}X_0} \circ \gamma \circ D \circ \gamma\|^2 = \|\widetilde{\mathcal{I}X_0}^* - \gamma \circ D \circ \gamma \circ \widetilde{\mathcal{I}X_0}^*\|^2 = \|\gamma \circ \widetilde{X_0} \circ \Gamma - \gamma \circ D \circ \gamma \circ \gamma \circ \widetilde{X_0} \circ \Gamma\|^2 = \|\gamma \circ \widetilde{X_0} \circ \Gamma - \gamma \circ D \circ \widetilde{X_0} \circ \Gamma\|^2 = \|\widetilde{X_0} - D \circ \widetilde{X_0}\|^2 = \|X_0 - D \circ X_0\|^2 = \int \|x_t - Dx_t\|^2_{L^2(\mathcal{A})} \mathrm{d}\mu(t)$. □

Remarque 5.1. *Ce dernier résultat se prête à une interprétation évidente : le résumé $p-$dimensionnel $((x_n, x_{n+\frac{1}{p}}, \ldots, x_{n+\frac{p-1}{p}}))_{n\in\mathbb{Z}}$ est d'autant plus performant que la famille de v.a.'s $\{x_t; t \in \mathbb{R}\}$ est proche de $\overline{vect}\{x_{\frac{n}{p}}; n \in \mathbb{Z}\}$.*

Au paragraphe 4, on associe une série stationnaire d'élément de $L^2_{L^2(\xi)}(\mathcal{A})$ à une f.a. cyclostationnaire. Le résultat qui suit définit la f.a. cyclostationnaire dont la série stationnaire associée est $(P_{Y'}\mathcal{I}X_n)_{n\in\mathbb{Z}}$.

Proposition 5.2. *La famille de v.a. $\{Dx_t; t \in \mathbb{R}\}$, est une f.a. cyclostationnaire qui a pour série stationnaire associée $(P_{Y'}\mathcal{I}X_n)_{n\in\mathbb{Z}}$.*

Démonstration. Commençons par démontrer que D et l'opérateur unitaire U (tel que $U^n x_t = x_{t+n}$) commutent. Comme U est une isométrie, on peut écrire $D = U^{-1} \circ P_{U\text{Im}D} \circ U$. Mais $U\text{Im}D = U\overline{\text{vect}}\{x_{\frac{n}{p}}; n \in \mathbb{Z}\} = \overline{\text{vect}}\{Ux_{\frac{n}{p}}; n \in \mathbb{Z}\} = \overline{\text{vect}}\{x_{\frac{n+p}{p}}; n \in \mathbb{Z}\} = \overline{\text{vect}}\{x_{\frac{n}{p}}; n \in \mathbb{Z}\} = \text{Im}D$. Ce qui précède s'écrit $D = U^{-1} \circ P_{\text{Im}D} \circ U$, soit $D = U^{-1} \circ D \circ U$ et donc $U \circ D = D \circ U$.

D'où le premier point de la définition de la cyclostationnarité :
$\langle Dx_t, Dx_{t'}\rangle = \langle UDx_t, UDx_{t'}\rangle = \langle DUx_t, DUx_{t'}\rangle = \langle Dx_{t+1}, Dx_{t'+1}\rangle$, pour tout couple (t, t') de réels.

De plus, l'application $t \in [0;1[\mapsto Dx_t \in L^2(\mathcal{A})$, donc l'application $D \circ (t \in [0;1[\mapsto x_t \in L^2(\mathcal{A}))$ est mesurable, car composée de deux applications mesurables (D est mesurable car continue). En outre, on a $\int \|Dx_t\|^2 \mathrm{d}\mu(t) \leq \int \|x_t\|^2 \mathrm{d}\mu(t) < +\infty$. D'où le point ii) de la définition de la cyclostationnarité.

Etant donné que $\{Dx_t; t \in \mathbb{R}\}$ est une f.a. cyclostationnaire, d'après le paragraphe 4, l'application $t \in [0;1[\mapsto Dx_{t+n} \in L^2(\mathcal{A})$, donc l'application $D \circ (t \in [0;1[\mapsto x_{t+n} \in L^2(\mathcal{A}))$, est un élément X'_n de $L^2_{L^2(\mathcal{A})}(\xi)$ et $(\mathcal{I}X'_n)_{n\in\mathbb{Z}}$ est la série d'éléments de $L^2_{L^2(\xi)}(\mathcal{A})$ stationnaire déduits de la f.a. cyclostationnaire $\{Dx_t; t \in \mathbb{R}\}$.

Il est clair que $X'_n = D \circ X_n$, et donc $\widetilde{\mathcal{I}X'_n} = (\widetilde{\mathcal{I}X'_n}^*)^* = (\gamma \circ \widetilde{X'_n} \circ \Gamma)^* = \Gamma \circ \widetilde{X_n}^* \circ D \circ \gamma = \Gamma \circ \widetilde{X_n}^* \circ \gamma \circ \gamma \circ D \circ \gamma = \widetilde{\mathcal{I}X_n} \circ \gamma \circ D \circ \gamma = \widetilde{P\mathcal{I}X_n}$, d'où $\mathcal{I}X'_n = P\mathcal{I}X_n$, ce qui termine la démonstration. □

6 Une hypothèse supplémentaire

On pourrait penser intuitivement que plus p est grand, plus $\|\mathcal{I}X_n - P\mathcal{I}X_n\|^2 = \int \|x_t - Dx_t\|^2 \mathrm{d}\mu(t)$ est petit. C'est ce que nous allons confirmer dans cette étude. Supposons que l'application $t \in \mathbb{R} \mapsto x_t \in L^2(\mathcal{A})$ est uniformément continue. Soit ε un réel strictement positif. Il existe un élément α_ε de $\mathbb{R}^*_+$ tel que $|t - t'| < \alpha_\varepsilon \Rightarrow \|x_t - x_{t'}\| < \varepsilon$. Choisissons un entier p tel que $\frac{1}{p} < \alpha_\varepsilon$. Comme $0 \leq t - \frac{1}{p}[pt] < \frac{1}{p} < \alpha_\varepsilon$ ($[pt]$ partie entière de pt), nous avons $\|x_t - x_{\frac{1}{p}[pt]}\| \leq \varepsilon$.

Mais $\|x_t - Dx_t\| \leq \|x_t - x_{\frac{1}{p}[pt]}\|$, donc $\|x_t - Dx_t\|^2 \leq \varepsilon^2$, pour tout t de $[0;1[$. Nous pouvons alors écrire $\|\mathcal{I}X_n - P_{Y'}\mathcal{I}X_n\|^2 = \int \|x_t - Dx_t\|^2 \mathrm{d}\mu(t) \leq \varepsilon^2$. Nous venons de prouver la propriété suivante.

Proposition 6.1. *Si $\{x_t; t \in \mathbb{R}\}$ est une f.a. cyclostationnaire telle que l'application $t \in \mathbb{R} \mapsto x_t \in L^2(\mathcal{A})$ soit uniformément continue, alors pour tout ε de $\mathbb{R}^*_+$, il existe un entier $p > 0$ tel que $\|\mathcal{I}X_n - P_{Y'}\mathcal{I}X'_n\|^2 \leq \varepsilon$.*

Bibliographie

Boudou, A. and Viguier-Pla, S. (2021), Principal Components Analysis and Cyclostationarity. *J. Multivariate Anal.*, to appear.

Boudou, A. and Dauxois, J. (1994). Principal Component Analysis for a Stationary Random Function Defined on a Locally Compact Abelian Group. *J. Multivariate Anal.* **51** 1-16.

Brillinger, D.M. (2001). *Time Series Data Analysis and Theory.* Reprint of the 1981 edition. Classics in Applied Mathematics, 36. Society for Industrial Applied Mathematics (SIAM), Philadelphia.

PLENIERE : Frédéric Chazal

Quelques propriétés statistiques des descripteurs topologiques des données

Frédéric Chazal

INRIA Univ Paris Saclay

L'Analyse Topologique des Données (TDA) est un domaine récent qui connait un succès croissant depuis quelques années. Il vise à comprendre, analyser et exploiter la structure topologique et géométrique de données complexes. Avec l'émergence de la théorie de la persistance homologique, la géométrie et la topologie ont fourni des outils nouveaux et efficaces pour aborder ces questions. Dans cet exposé, nous introduirons quelques outils permettant de construire des descripteurs robustes de la topologie des données. Nous nous intéresserons en particulier à leurs propriétés statistiques et nous illustrerons, sur quelques exemples concret, l'intérêt des approches topologiques pour l'analyse des données et l'apprentissage statistique.

PLENIERE : Gersende Fort

Algorithmes Majoration-Minoration stochastiques pour l'Apprentissage Statistique grande échelle

Gersende Fort

CNRS IMT Toulouse

En Apprentissage Statistique, notamment pour la minimisation de fonctions de risque, on s'intéresse à la minimisation de fonctions de type 'sommes finies' i.e. s'exprimant comme la somme d'un grand nombre de termes, eux-mêmes pouvant ne pas avoir d'expressions explicites. Il est alors nécessaire de définir des procédures d'optimisation stochastiques, capables de réduire le coût computationnel lié à la gestion de grands ensembles d'apprentissage, mais aussi d'intégrer des approximations consistantes de quantités incalculables.

Cet exposé sera consacré aux procédures d'optimisation de type Majoration-Minoration (MM). Les algorithmes de gradient stochastique et leurs extensions proximales, ou encore les algorithmes Expectation-Maximization pour l'apprentissage dans les modèles à données latentes, sont des exemples de procédures MM très populaires en Statistique.

Cet exposé présentera de nouveaux algorithmes d'approximations stochastiques accélérés pour répondre à des problèmes d'apprentissage statistique dans le contexte usuel de grands ensembles d'apprentissage, mais aussi celui de l'apprentissage en ligne et de l'apprentissage fédéré. Des éléments d'analyse de convergence et d'analyse de complexité seront aussi discutés.

Séries Temporelles

Une nouvelle méthode pour les propriétés asymptotiques de modèles à seuil autoexcités

Guy Mélard [1] & Marcella Niglio [2]

[1] *Université libre de Bruxelles, SBS-EM & ECARES, Belgique - Guy.Melard@ulb.be*
[2] *Universitá degli Studi di Salerno, Department of Economics and Statistics, Italy - mniglio@unisa.it*

Résumé. Une nouvelle méthode pour obtenir les propriétés asymptotiques des estimateurs de modèles à seuil autoexcités autorégressifs (self-exciting threshold autoregressive, SETAR) est introduite. Les modèles à seuil sont des modèles non linéaires pour une série temporelle $x_t, t = 1, ..., n$, avec k régimes, où le régime dépend de la valeur d'une variable y_{t-d}, $d > 0$, appelée variable de seuil, par rapport à une (quand $k = 2$) ou à plusieurs (quand $k > 2$) valeur(s) de seuil. Comme pour la plupart des modèles non linéaires, la méthode usuelle pour établir les propriétés asymptotiques de tels modèles consiste à obtenir une solution stationnaire et ergodique de l'équation du modèle et à utiliser l'ergodicité pour prouver la consistance et la normalité asymptotique d'un estimateur des paramètres du modèle. Une nouvelle méthode pour établir ces propriétés asymptotiques a été récemment proposée par les auteurs quand la variable de seuil y_{t-d} est exogène et indépendante des innovations du modèle. La méthode est basée sur la théorie asymptotique pour des modèles tdARMA scalaires ou vectoriels, où les coefficients ne sont pas constants mais sont des fonctions déterministes du temps et d'un petit nombre de paramètres. La méthode est donc valable bien au-delà de modèles autorégressifs à seuil (TAR), comme des modèles TARMA et leurs généralisations multivariées, mais toujours en supposant une variable de seuil exogène. Les modèles SETAR, c'est-à-dire quand $y_{t-d} \equiv x_{t-d}$, par exemple, ont des coefficients aléatoires de sorte que la théorie n'est plus directement applicable. Néanmoins, il est possible d'adapter les résultats fondamentaux d'Alj et al. (2017) au cas de coefficients aléatoires tels qu'ils apparaissent dans les modèles SETAR ainsi que dans leurs généralisations SETARMA scalaires ou vectorielles. Le seul problème avec cette nouvelle méthode est que l'existence et la non-singularité de la matrice d'information doivent être supposées ou prouvées.

Mots-clés. modèle SETAR, modèle SETARMA, série temporelle

Abstract. A new method for obtaining the asymptotic properties of estimators for self-exciting threshold autoregressive (SETAR) models is introduced. Threshold models are non-linear models for time series $x_t, t = 1, ..., n$, with k regimes, where the regime depends on the value of a variable y_{t-d}, $d > 0$, called the threshold variable, with respect to one (when $k = 2$) or several (when $k > 2$) threshold values. Like for most non-linear models, the usual method for obtaining the asymptotic properties for such models consists in exhibiting a stationary and ergodic solution of the model equation and using

ergodicity to prove the consistency and the asymptotic normality of an estimator of the model parameters. A new method for obtaining these asymptotic properties was recently proposed by the authors when the threshold variable y_{t-d} is exogenous and independent of the innovations of the time series models. That method is based on an asymptotic theory for scalar or vector tdARMA models, where the coefficients are not constant but are deterministic functions of time and of a small number of parameters. The method is thus valid well beyond threshold autoregressive (TAR) models, like TARMA models and their multivariate counterparts, but still assuming an exogenous threshold variable. SETARMA models, i.e. when $y_{t-d} \equiv x_{t-d}$, have random coefficients so that the theory is not directly applicable. Nevertheless, it is possible to adapt the fundamental results of Alj et al. (2017) to the case of the randomly-varying coefficients that appear in SETAR models and in their SETARMA scalar or vector generalizations. The only problem with the new method is that the existence and the non-singularity of the information matrix have to be assumed or proved.

Keywords. SETAR model, SETARMA model, self-exciting threshold model, time series, asymptotic properties

1 Introduction

Une nouvelle méthode pour obtenir les propriétés asymptotiques de modèles à seuil autoexcités autorégressifs (self-exciting threshold autoregressive, SETAR) est introduite. Voir par exemple Tong (1983, 1990) et De Gooijer (2017, Section 2.6). Les modèles à seuil sont des modèles non linéaires pour une série temporelle $x_t, t = 1, ..., n$, avec k régimes, où le régime dépend de la valeur d'une variable y_{t-d}, $d > 0$, appelée variable de seuil, par rapport à une (quand $k = 2$) ou à plusieurs (quand $k > 2$) valeur(s) de seuil. L'exemple le plus simple est donné par le modèle TAR(1) où l'observation x_t au temps t est obtenue par

$$x_t = \begin{cases} \phi^{(1)} x_{t-1} + \epsilon_t, & \text{si } y_{t-1} \leq r, \\ \phi^{(2)} x_{t-1} + \epsilon_t, & \text{si } y_{t-1} > r, \end{cases} \tag{1}$$

où $\phi^{(1)}$ and $\phi^{(2)}$ sont des paramètres, r est le seuil et les ϵ_t sont des variables aléatoires independantes de moyenne 0 et de variance constante σ^2. Dans (1), on a $k = 2$ et $d = 1$. Si on remplace y_{t-1} par x_{t-1}, on a un modèle SETAR(1).

Comme pour la plupart des modèles non linéaires, la méthode usuelle pour établir les propriétés asymptotiques de tels modèles consiste à obtenir une solution stationnaire et ergodique de l'équation du modèle et à utiliser l'ergodicité pour prouver la consistance et la normalité asymptotique d'un estimateur des paramètres du modèle (Chan, 1993).Une nouvelle méthode pour établir ces propriétés asymptotiques a été récemment proposée par les auteurs, voir Mélard et Niglio (2021, 2022), quand la variable de seuil y_{t-d} est

exogène et indépendante des innovations du modèle. La méthode est basée sur la théorie asymptotique pour des modèles ARMA à coefficients dépendant du temps ou tdARMA, scalaires ou vectoriels, où les coefficients ne sont pas constants mais sont des fonctions déterministes du temps et d'un petit nombre de paramètres. La méthode est donc valable bien au-delà de modèles autorégressifs à seuil (TAR), comme des modèles TARMA et leurs généralisations multivariées, mais toujours en supposant une variable de seuil exogène.

Les modèles SETAR, c'est-à-dire quand $y_{t-d} \equiv x_{t-d}$, par exemple, ont des coefficients aléatoires de sorte que la théorie n'est plus directement applicable. Néanmoins, il est possible d'adapter les résultats fondamentaux d'Alj et al. (2017) au cas de coefficients aléatoires tels qu'ils apparaissent dans les modèles SETAR ainsi que dans leurs généralisations SETARMA scalaires et vectorielles. Le seul problème avec cette nouvelle méthode est que l'existence et la non-singularité de la matrice d'information doivent être supposées ou prouvées. Pour la simplicité nous ne traitons ici que le cas autorégressif et univarié, laissant les modèles ARMA et VARMA pour un article futur.

2 Préliminaires

2.1 Introduction des modèles TAR, SETAR et tdAR

Considérons un modèle TAR(p) qui peut se décrire par l'équation suivante

$$x_t = \sum_{i=1}^{k} \left[\sum_{j=1}^{p} \phi_j^{(i)} x_{t-j} \right] \mathbb{I}_{\{y_{t-d} \in \mathbb{R}_i\}} + \epsilon_t, \tag{2}$$

où les $\phi_j^{(i)}$ sont les paramètres d'intérêt, les ϵ_t sont comme ci-dessus et les $\mathbb{R}_i$ sont une partition de la droite réelle $\mathbb{R}$ en intervalles. Si $y_{t-d} \equiv x_{t-d}$, on a un modèle SETAR(p).

Les modèles autorégressifs d'ordre p à coefficients dépendant du temps ou tdAR se définissent par l'équation suivante

$$x_t = \sum_{j=1}^{p} \phi_{tj} x_{t-j} + \epsilon_t, \tag{3}$$

où les ϕ_{tj} sont des fonctions déterministes du temps dépendant du temps et d'un vecteur $\boldsymbol{\beta}$ de paramètres d'intérêt et les ϵ_t sont comme ci-dessus. Il est clair que (2) peut s'écrire comme (3) en posant $\phi_{tj}(\boldsymbol{\beta}) = \sum_{i=1}^{k} \phi_j^{(i)} \mathbb{I}_{\{y_{t-d} \in \mathbb{R}_i\}}$ et en notant $\boldsymbol{\beta} = (\phi_1^{(1)} ... \phi_p^{(1)} \phi_1^{(2)} ... \phi_p^{(k)})'$, où $'$ désigne la transposition. Notons m le nombre de paramètres, $e_t(\boldsymbol{\beta})$ le résidu du modèle et $\boldsymbol{\beta}_0$ la vraie valeur de $\boldsymbol{\beta}$ de sorte que $e_t(\boldsymbol{\beta}_0) = \epsilon_t$, pour tout t.

2.2 Mise du modèle sous forme moyenne mobile

Comme dans Mélard et Niglio (2021, 2022), on met le modèle sous forme tdVAR(1)

$$\mathbf{X}_t(\boldsymbol{\beta}) = \mathbf{A}_t(\boldsymbol{\beta}) \mathbf{X}_{t-1}(\boldsymbol{\beta}) + \mathbf{E}_t(\boldsymbol{\beta}),$$

avec

$$\mathbf{A}_t(\boldsymbol{\beta}) = \begin{bmatrix} \boldsymbol{\Phi}_t(\boldsymbol{\beta}) & \underset{p\times 1}{\mathbf{0}} \\ \underset{1\times p}{\mathbf{0}} & 0 \end{bmatrix}, \quad \boldsymbol{\Phi}_t(\boldsymbol{\beta}) = \begin{bmatrix} \phi_{t1}(\boldsymbol{\beta}) & \cdots & \phi_{tp}(\boldsymbol{\beta}) \\ & \underset{(p-1)}{\mathbf{I}} & \underset{(p-1)\times 1}{\mathbf{0}} \end{bmatrix},$$

et

$$\mathbf{X}_t(\boldsymbol{\beta}) = \begin{bmatrix} X_t \\ \cdots \\ X_{t-p+1} \\ e_t(\boldsymbol{\beta}) \end{bmatrix}, \quad \mathbf{E}_t(\boldsymbol{\beta}) = \begin{bmatrix} e_t(\boldsymbol{\beta}) \\ \underset{(p-1)}{\mathbf{0}} \\ e_t(\boldsymbol{\beta}) \end{bmatrix},$$

puis sous forme moyenne mobile (voir Mélard, 2022) :

$$\mathbf{X}_t(\boldsymbol{\beta}) = \sum_{\ell=0}^{t-1} \boldsymbol{\Psi}_{t\ell}(\boldsymbol{\beta}) \mathbf{E}_{t-\ell}(\boldsymbol{\beta}_0), \tag{4}$$

où $\boldsymbol{\Psi}_{t\ell}(\boldsymbol{\beta}) = \sum_{s=0}^{\ell} \left(\prod_{j=0}^{\ell-s-1} \mathbf{J}\mathbf{A}_{t-j}(\boldsymbol{\beta}) \right) \mathbf{K} \left(\prod_{j=0}^{s-1} \mathbf{A}_{t-\ell+s-j}(\boldsymbol{\beta}_0) \right)$, $\mathbf{J}$ et $\mathbf{K}$ sont des matrices carrées de dimension $(p+1)$, $\mathbf{J}$ est une matrice identité avec deux remplacements $J_{1,1} = 0$ and $J_{(p+1),1} = -1$ et $\mathbf{K}$ est une matrice nulle avec deux remplacements $K_{1,1} = K_{(p+1),1} = 1$. Ensuite, sachant que $e_t(\boldsymbol{\beta})$ est l'élément $(p+1)$ de $\mathbf{X}_t(\boldsymbol{\beta})$:

$$e_t(\boldsymbol{\beta}) = \sum_{\ell=1}^{t-1} \psi_{t\ell}(\boldsymbol{\beta}) \epsilon_{t-\ell}, \tag{5}$$

avec $\psi_{t\ell}(\boldsymbol{\beta}) = \mathbf{U}'_{p+1} \Psi_{t\ell}(\boldsymbol{\beta}) \mathbf{U}_1$, où $\mathbf{U}_1$ et $\mathbf{U}_{p+1}$ sont deux vecteur nuls $(p+1) \times 1$ où les éléments 1 et $(p+1)$ sont remplacés par 1, respectivement. On peut définir la dérivée de $e_t(\boldsymbol{\beta})$ par rapport à β_i par

$$\frac{\partial e_t(\boldsymbol{\beta})}{\partial \beta_i} = \sum_{\ell=1}^{t-1} \psi_{t\ell i}(\boldsymbol{\beta}) \epsilon_{t-\ell}, \tag{6}$$

où $\psi_{t\ell i} = \partial \psi_{t\ell}(\boldsymbol{\beta}) / \partial \beta_i$, $i = 1, ..., m$. La théorie nécessite aussi les dérivées d'ordre 2 et 3 de $e_t(\boldsymbol{\beta})$ qu'on peut traiter de même.

2.3 Estimation des paramètres

Supposons que la série $x_t, t = 1, ..., n$, de longueur n soit une réalisation du processus SETAR défini par (2) pour $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ et avec $y \equiv x$. On estime $\boldsymbol{\beta}$ par la méthode du quasi-maximum de vraisemblance comme si la distribution des innovations ϵ_t était normale $(0, \sigma^2)$. Notons $\widehat{\boldsymbol{\beta}}_n$ l'estimateur de $\boldsymbol{\beta}$. On doit montrer, sous certaines conditions, que $\widehat{\boldsymbol{\beta}}_n \to \boldsymbol{\beta}_0$ presque sûrement quand $n \to \infty$ et $\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \to N(\mathbf{0}, \boldsymbol{V}^{-1})$, en loi, quand $n \to \infty$.

La nouveauté de nos résultats en comparaison avec ceux de Chan (1993), où les estimateurs conditionnels des moindres carrés sont obtenus pour les paramètres d'un modèle

SETAR, est que x_t n'est pas supposé ici être stationnaire et ergodique, de sorte que nos résultats inférentiels sont plus généraux.

3 Contribution

Dans Mélard et Niglio (2021, 2022), nous avons traité le cas d'une variable de seuil exogène et appliqué la théorie d'Alj et al. (2017) mais ce n'est pas possible dans le cas autoexcité parce que les coefficients $\psi_{t\ell}(\boldsymbol{\beta})$ dans (5) et donc aussi les coefficients $\psi_{ti\ell}(\boldsymbol{\beta})$ dans (6) sont des variables aléatoires qui ne sont pas indépendantes des ϵ_t.

Il faut donc remplacer les suppositions pour un modèle *td*VAR de manière à pouvoir étendre ces résultats au cas de coefficients aléatoires. Par manque de place, nous n'allons pas examiner toutes les suppositions, voir Mélard (2022) et Mélard et Niglio (2022), mais seulement celles qui sont critiques. Tout d'abord, deux suppositions doivent être maintenues : l'existence des moments d'ordre 4 des ϵ_t et celle d'existence et d'inversibilité de la matrice d'information de Fisher V dont les éléments V_{ij} sont définis par les limites $\lim_{n\to\infty} \frac{1}{n\sigma^2} \sum_{t=1}^{n} E_{\boldsymbol{\beta}_0} \left(\frac{\partial \epsilon_t(\boldsymbol{\beta})}{\partial \beta_i} \frac{\partial \epsilon_t(\boldsymbol{\beta})}{\partial \beta_j} \right)$.

Pour le reste, il apparaît qu'il faut trouver des bornes supérieures pour des formes quadratiques du type $E_{\boldsymbol{\beta}_0}\{(\partial e_t(\boldsymbol{\beta})/\partial \beta_i)(\partial e_t(\boldsymbol{\beta})/\partial \beta_j)\}$ (et aussi quartiques que nous ne donnerons pas ici). L'idée est de remplacer les dérivés en employant (6) et de transformer l'expression à borner par

$$\left| E \left(\sum_{\ell_1=1}^{t-1} \sum_{\ell_2=1}^{t-1} \epsilon_{t-\ell_1} \psi_{ti\ell_1} \psi_{tj\ell_2} \epsilon_{t-\ell_2} \right) \right|,$$

où $\psi_{ti\ell} = \psi_{ti\ell}(\boldsymbol{\beta}_0)$, et celle-ci devrait être finie pour satisfaire les conditions d'Alj et al. (2017, Technical Appendix Lemma 4.7). A cette fin, on se basait sur des bornes avec l'existence de constantes N et Φ, $0 < \Phi < 1$ telles que $\sum_{k=\nu}^{t-1} \psi_{ti\ell}^2 < N\Phi^{\nu-1}$ pour $\nu = 1, ..., t-1$ et $i = 1, ..., m$.

Il n'est pas possible d'employer les mêmes arguments si les $\psi_{ti\ell}(\boldsymbol{\beta})$ sont aléatoires mais bien de trouver d'autres majorations. Notons $C_{tij\ell_1\ell_2} = \psi_{ti\ell_1}\psi_{tj\ell_2}$ et $D_{t\ell_1\ell_2} = \epsilon_{t-\ell_2}\epsilon_{t-\ell_1}$. On a

$$E \left(\left| \sum_{\ell_1=1}^{t-1} \sum_{\ell_2=1}^{t-1} C_{tij\ell_1\ell_2} D_{t\ell_1\ell_2} \right| \right) \leq \sum_{\ell_1=1}^{t-1} \sum_{\ell_2=1}^{t-1} \{E(|C_{tij\ell_1\ell_2}|^2)\}^{1/2} \{E(|D_{t\ell_1\ell_2}|^2)\}^{1/2}.$$

Mais $\{E(|D_{t\ell_1\ell_2}|^2)\}^{1/2} \leq \sigma^2$ et $|C_{tij\ell_1\ell_2}|^2 = \psi_{ti\ell_1}^2 \psi_{tj\ell_2}^2$.

Nous faisons la supposition suivante concernant les régimes du modèle SETAR : les racines de tous les polynômes autoréregressifs $(1 - \phi_1^{(i)} z - ... - \phi_p^{(i)} z^p)$ sont supérieures à 1 en module, $i = 1, ..., k$. Notons que cette condition de stationnarité sur chacun des régimes ne doit pas être confondue avec la stationnarité du modèle SETAR complet pour

lequel les conditions sont plus restrictives. Comme il existe alors Φ tel que $0 < \Phi < 1$ et que la matrice $\boldsymbol{\Phi}_t(\boldsymbol{\beta})$ a toutes ses valeurs propres inférieures à Φ, il résulte de la forme des coefficients dans (4) que la norme de Frobenius de $\boldsymbol{\Psi}_{t\ell}$ est bornée presque sûrement par Φ^ℓ à un facteur constant près. On peut vérifier que cela implique une borne similaire pour les coefficients dans (6) donc les $\psi_{ti\ell}$, $i = 1, ..., m$, d'où une borne pour $E(|C_{tij\ell_1\ell_2}|^2)$. Un raisonnement similaire tient pour les autres formes quadratiques avec les autres dérivées de $e_t(\boldsymbol{\beta})$ et aussi pour les formes quartiques et, enfin, il en est de même pour les autres conditions employées par Alj et al. (2017). Pour les modèles ARMA, on doit recourir à une représentation tdVAR(1) plus générale, voir Mélard et Niglio (2021, 2022). Ce n'est pas clair dans ce qui précède, mais on peut adapter le développement avec $\boldsymbol{e}_t(\boldsymbol{\beta})$ vectoriel et une matrice $\boldsymbol{\psi}_{ti\ell}$ pour traiter des modèles VARMA à seuil avec une variable de seuil scalaire (par exemple une combinaison linéaire des composantes) ou même vectorielle.

En conclusion, contrairement aux approches existantes dans la littérature, on n'a pas besoin d'une solution stationnaire et ergodique de l'équation du modèle SETAR ou SETARMA mais seulement de la supposition ci-dessus en plus de l'existence des moments d'ordre 4 et de l'existence et l'inversibilité de la matrice d'information de Fisher V. En outre, comme indiqué, on peut traiter les modèles SETARMA scalaires et vectoriels de la même manière que les modèles SETAR.

Bibliographie

Alj, A., Azrak, R., Ley, C. & Mélard, G. (2017), Asymptotic properties of QML estimators for VARMA models with time-dependent coefficients. *Scandinavian Journal of Statistics*, 44, 617-635.

Chan, K.S. (1993), Consistency and limiting distribution of the least squares estimator of a threshold autoregressive model, *The Annals of Statistics*, 21, 520-533.

De Gooijer, J. G. (2017), *Elements of Nonlinear Time Series Analysis and Forecasting*. New York: Springer-Verlag.

Mélard G. (2022), An indirect proof for the asymptotic properties of VARMA model estimators, *Econometrics and Statistics*, 21, 96-111.

Mélard G. and Niglio, M. (2021), Link between threshold ARMA and tdARMA models, in Recueil des soumissions, 53èmes Journées de Statistique de la Société Française de Statistique (SFdS), Université Côte d'Azur, Nice, France, 7-11 juin 2021, pp. 770-775.

Mélard G. and Niglio, M. (2022), Another approach for asymptotic properties of threshold models. En préparation.

Tong, H. (1983), *Threshold Models in Non-linear Time Series Analysis*, Springer-Verlag, New York.

Tong, H. (1990), *Non-linear Time Series: A Dynamical System Approach*, Oxford University Press, Oxford.

Barycentres de séries temporelles : une nouvelle approche basée sur la méthode de la signature

Raphaël Mignot [1] & Konstantin Usevich [2] & Marianne Clausel [1]
& Georges Oppenheim [3] & Laure Coutin [4] & Antoine Lejay [1]

[1] *Université de Lorraine, CNRS, IECL, Nancy, France*
[2] *Université de Lorraine, CNRS, CRAN, Nancy, France.*
[3] *Université Paris-Est-Marne la Vallée, Département de Mathématiques, Marne-la-Vallée, France*
[4] *Université Paul Sabatier, IMT, Toulouse, France*

Résumé. La méthode de la signature a été largement utilisée pour l'analyse des séries temporelles. Cette approche a prouvé son efficacité pour de nombreuses applications en apprentissage statistique. La définition d'une notion de barycentre dans l'espace des signatures est un premier pas prometteur permettant de développer de nouvelles extensions de l'analyse en composantes principales (ACP) ou de l'algorithme des k-moyennes aux séries temporelles.

Mots-clés. Séries temporelles multivariées, clustering, signature par intégrales itérées.

Abstract. The signature method is a widely used method for time series analysis. This novel approach has shown to perform well in various machine learning contexts. Designing a notion of barycenter in the signature space would allow the use of the signature transform in ubiquitous strategies in data science such as the Principal Component Analysis (PCA) for data compression or k-means for clustering.

Keywords. Multidimensional time series, clustering, iterated integrals signature.

1 Introduction

L'analyse statistique de séries temporelles multivariées est un problème difficile, avec de nombreuses applications dans des domaines divers tels la finance, l'environnement ou le domaine médical. Une des principales difficultés est d'encoder de manière pertinente la dépendance temporelle inhérente à chaque composante ainsi que les dépendances possiblement non linéaires entre les composantes. Une approche prometteuse introduite tout récemment vise à utiliser la méthode dite de la signature [Chevyrev and Kormilitzin, 2016], introduite initialement dans les années 1950 par K-T. Chen [Chen, 1957] dans le cadre de la théorie du contrôle et qui a connu de nombreux développements liés à la théorie des trajectoires rugueuses [Lyons, 1998].

La méthode de la signature a souvent été utilisée en amont des méthodes usuelles d'analyse de séries temporelles multivariées, comme une première étape d'encodage des

dépendances inter et intra composantes. Cette approche s'est révélée être redoutablement efficace pour de nombreuses applications, comme la reconnaissance d'idéogrammes manuscrits [Yang et al., 2016], la détection du trouble bipolaire [Perez Arribas et al., 2018] ou encore en océanographie [Sugiura and Hosoda, 2020]. Ici, nous nous intéressons à la résolution d'un problème important dans le domaine de l'apprentissage : celui du calcul de barycentres de séries temporelles. Nous montrons que notre approche basée sur le calcul de la signature a de bonnes propriétés sur le plan computationnel et statistique.

2 Processus multivariés et méthode de la signature

Soit $X : [0,1] \to \mathbb{R}^d$ un processus multivarié continu à variations bornées. Pour tout $0 \leq t_1 < t_2 \leq 1$ et m un entier naturel, on définit sa signature d'ordre m sur le segment $[t_1, t_2]$ de la manière suivante :

$$S^{(m)}_{[t_1,t_2]}(X) \overset{\text{déf}}{=} \int\cdots\int_{t_1<u_1<\cdots<u_m<t_2} dX_{u_1} \otimes \cdots \otimes dX_{u_m} \tag{1}$$

où $\otimes$ est le produit tensoriel. La signature d'ordre m est un élément de $(\mathbb{R}^d)^{\otimes m}$ dont chaque coefficient est intuitivement une mesure d'association entre m composantes du processus multivarié considéré.

La signature du processus multivarié X est alors la collection infinie des signatures de tous les ordres :

$$S_{[t_1,t_2]}(X) = (1, S^{(1)}_{[t_1,t_2]}(X), S^{(2)}_{[t_1,t_2]}(X), \dots). \tag{2}$$

où par convention on a fixé $S^{(0)}_{[t_1,t_2]}(X) = 1$.

La signature possède trois propriétés principales : l'invariance par reparamétrisation du temps qui en fait un bon outil pour analyser des courbes, la relation de Chen, qui permet de calculer la signature de deux chemins concaténés et enfin le produit de *shuffle*, qui permet de calculer le produit de deux signatures d'une même trajectoire.

Lorsqu'on analyse des séries temporelles multivariées, on peut leur associer par interpolation linéaire un processus multivarié. Les intégrales itérées (1) peuvent être alors calculées numériquement. En pratique, un ordre de troncature m est fixé, et l'on récupère les signatures d'ordres inférieurs.

3 Barycentre de séries temporelles multivariées et signature

Soit $(X_i)_{1\leq i\leq n}$ un jeu de données de séries temporelles discrètes, avec n le nombre d'observations. Notre objectif est de définir une notion de barycentre de ces séries temporelles,

et pour cela une étape intermédiaire va être de trouver le barycentre de leurs signatures respectives.

Pour se fixer les idées, nous supposerons que chaque série temporelle est de taille $d \times l$ avec d le nombre de dimensions et l le nombre d'instants de temps. Dans le vocabulaire de l'apprentissage automatique (n, l, d) sont respectivement appelés *batch, stream* et *channels*. On cherche à définir un barycentre des n signatures $\{S(X_1), \ldots, S(X_n)\}$ de poids $(w_i)_{1 \leq i \leq n}$. Nous présentons trois stratégies qui ont toutes pour point de départ la nature de groupe de Lie G de l'espace des signatures. On notera $\mathfrak{g}$ son algèbre de Lie.

3.1 Barycentre log-euclidien

La méthode naïve consistant à calculer la moyenne euclidienne $\bar{S} = \frac{1}{n}\sum_{i=1}^{n} S(X_i)$ n'est pas pertinente sur le groupe de Lie des signatures car $\bar{S}$ peut ne pas être une signature. Pour obtenir une moyenne qui reste dans l'espace des signatures, une première stratégie consiste à effectuer une moyenne euclidienne dans l'algèbre de Lie $\mathfrak{g}$ puis de repasser dans le groupe de Lie G :

$$\bar{S}_{LE} = \exp\left(\frac{1}{n}\sum_{i=1}^{n} w_i \log S(X_i)\right). \tag{3}$$

Cependant, ce barycentre n'est pas invariant par translation à gauche ou à droite. Ceci implique que si l'on concatène à une famille de séries multivariées $X_1, \cdots, X_n$ une même série multivariée Y, on n'a aucune assurance que le barycentre des $X_i \star Y$ soit le barycentre des X_i auquel on aurait concaténé Y. Cette notion de barycentre amène donc à des phénomènes non intuitifs que nous souhaitons éviter. Ceci motive ainsi la stratégie suivante que nous proposons.

3.2 Barycentre exponentiel de groupe

La première stratégie vise à adapter [Pennec and Arsigny, 2013, Algorithme 1] au contexte de la signature afin de définir de manière itérative un barycentre sur groupe de Lie, invariant par translation à gauche ou bien à droite. Après initialisation de $\bar{S}^0$ (par exemple $\bar{S}^0 = S(X_i)$ pour un i tiré aléatoirement), l'étape de mise à jour de l'algorithme est la suivante pour $k = 1, 2, \ldots,$

$$\bar{S}^{k+1} = \bar{S}^k \otimes \exp\left(\sum_{i=1}^{n} w_i \log\left((\bar{S}^k)^{-1} \otimes S(X_i)\right)\right). \tag{4}$$

On s'arrête quand le critère de convergence choisi est atteint, par exemple un nombre d'itérations maximal ou lorsque $d(\bar{S}^k, S^{k+1})$ est inférieur à un seuil fixé. La procédure converge vers une solution, à condition que les données $(S(X_i))_{1 \leq i \leq n}$ soient suffisamment proches de l'élément neutre, et que l'initialisation $\bar{S}^0$ soit suffisamment proche des données

[Pennec and Arsigny, 2013, Corollaire 5]. En pratique, on observe que lorsque l'algorithme converge, il le fait en moins de 10 itérations.

3.3 Optimisation sur l'espace des trajectoires

Dans les deux méthodes précédentes, le barycentre obtenu est une signature. Si l'on souhaite obtenir un chemin correspondant à $\bar{S}$, c'est-à-dire un chemin X tel que $S(X) = \bar{S}$, une inversion est nécessaire. Cette inversion est de mieux en mieux maîtrisée mais reste loin d'être exacte et nécessite un ordre élevé de la signature [Chang and Lyons, 2019].

Dans la méthode suivante, le barycentre obtenu est directement une trajectoire discrète. Celui-ci est obtenu par optimisation sur l'espace des matrices. Soit $X \in \mathbb{R}^{d\times\tilde{l}}$ une trajectoire discrète avec $1 \leq \tilde{l} \leq l$ et soit m un ordre de troncature pour la signature. On définit notre fonction objectif à minimiser :

$$f(X) = \sum_{i=1}^{n}\sum_{p=1}^{m} d(S^{(p)}(X), S^{(p)}(X_i)) \tag{5}$$

avec d une distance à choisir, par exemple $d(x,y) = ||x-y||_F^2$, avec $||.||_F$ la norme de Frobenius ou bien $d(x,y) = ||x^{-1}\otimes y||_{CC}$ avec $||.||_{CC}$ la norme de Carnot-Carathéodory. La fonction objectif peut être minimisée par exemple par descente de gradient. Numériquement, le gradient peut facilement être calculé par différentiation automatique.

Afin d'accroître les performances et la robustesse de cette méthode, on peut jouer sur plusieurs paramètres. La longueur $\tilde{l}$ de la matrice X est un hyper-paramètre à régler afin d'éviter un sur ou sous-apprentissage. On peut jouer sur le nombre d'initialisations. En choisissant un unique point de départ pour l'optimisation, on peut tomber dans un minimum local, surtout si le point choisi se trouve dans une région éloignée du minimum global. Effectuer p initialisations accroît les chances de se trouver proche du minimum global au départ. On choisit ensuite le plus petit minimum obtenu parmi les p solutions.

3.4 Comparaison des trois stratégies

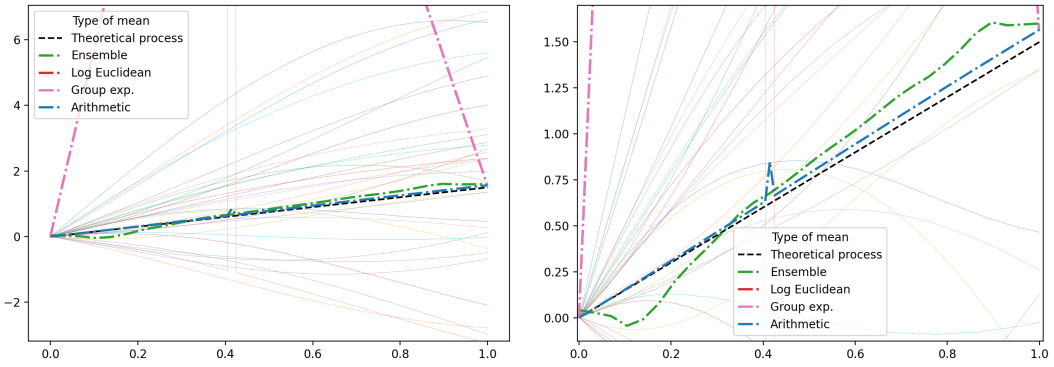
Les trois méthodes précédentes ont des cas d'utilisations préférentiels : selon que l'on souhaite obtenir une signature ou une trajectoire. Seules les deux premières méthodes tirent vraiment parti de la structure intrinsèque de l'espace des signatures (groupe de Lie). En pratique, le temps de calcul des deux premières méthodes est quasi-instantané. Alors que celui de la troisième méthode varie selon les paramètres choisis et la convergence n'est pas assurée.

4 Expériences numériques

Pour illustrer les différentes stratégies de barycentre, on simule $n = 500$ processus gaussiens sur $l = 100$ indices $0 = t_0 < t_1 < \cdots < t_l = 1$ avec pas de temps $h = t_{i+1} - t_i$

constant. On calcule la signature jusqu'à l'ordre $m = 8$. Pour le barycentre exponentiel de groupe, l'initialisation est faite à partir de $\bar{S}^0 = S(X_i)$ avec i tiré aléatoirement, puis on effectue 10 itérations. Pour la méthode par optimisation sur l'espace des trajectoires, on effectue 50 initialisations différentes avec la même stratégie : on part de $\bar{X}^0 = X_i$ avec i tiré aléatoirement. On choisit comme critère d'arrêt pour la descente de gradient un nombre maximum d'itérations fixé à 1000. Par ailleurs, pour évaluer la robustesse des stratégies, on ajoute une valeur aberrante.

On tire une observation i et un indice de temps t_j aléatoirement. On obtient un point $X^i_{t_j} \in \mathbb{R}$ et on remplace cette valeur par une valeur aberrante. Le résultat est illustré en Figure 1. On observe une bonne résistance à la valeur aberrante par la stratégie d'optimisation (courbe verte), qui semble même meilleure que la robustesse de la moyenne arithmétique (courbe bleue). Les deux premières stratégies, barycentre log-euclidien et barycentre exponentiel de groupe sont moins robustes et sortent de la figure.



(a) Zoom arrière (b) Zoom avant

Figure 1: Illustration des trois stratégies de barycentre avec une valeur aberrante (pointillés fins en rouge). La courbe *Ensemble* (verte) est la méthode Optimisation sur l'espace des trajectoires. La courbe *Arithmetic* (bleue) est une moyenne arithmétique classique, sans utilisation de la transformée par signature. La courbe *Theoretical process* (noire) est la moyenne théorique utilisée pour simuler les processus gaussiens.

En outre, il faut souligner le fait que la bonne représentation n'est peut-être pas dans l'espace des trajectoires mais dans l'espace des signatures. Or, celui-ci étant un groupe de Lie de grande dimension, il est difficile de l'illustrer sur un plan.

5 Conclusion

Plusieurs stratégies de barycentre pour la transformée par signature ont été définies, avec des cas d'utilisations qui peuvent différer. Ce travail n'est encore que préliminaire. Dans de futurs travaux, les performances des trois stratégies seront étudiées par exemple pour

le *clustering* de séries temporelles multivariées, et mis en concurrence avec des méthodes de pointe comme le Dynamic Time Warping (DTW) sur des données réelles.

Références

[Chang and Lyons, 2019] Chang, J. and Lyons, T. (2019). Insertion algorithm for inverting the signature of a path. *arXiv preprint arXiv:1907.08423*.

[Chen, 1957] Chen, K.-T. (1957). Integration of paths, geometric invariants and a generalized baker-hausdorff formula. *Annals of Mathematics*, pages 163–178.

[Chevyrev and Kormilitzin, 2016] Chevyrev, I. and Kormilitzin, A. (2016). A primer on the signature method in machine learning.

[Lyons, 1998] Lyons, T. J. (1998). Differential equations driven by rough signals. *Revista Matemática Iberoamericana*, 14(2):215–310.

[Pennec and Arsigny, 2013] Pennec, X. and Arsigny, V. (2013). *Exponential Barycenters of the Canonical Cartan Connection and Invariant Means on Lie Groups*, page 123–166. Springer Berlin Heidelberg.

[Perez Arribas et al., 2018] Perez Arribas, I., Goodwin, G. M., Geddes, J. R., Lyons, T., and Saunders, K. E. (2018). A signature-based machine learning model for distinguishing bipolar disorder and borderline personality disorder. *Translational psychiatry*, 8(1):1–7.

[Sugiura and Hosoda, 2020] Sugiura, N. and Hosoda, S. (2020). Machine learning technique using the signature method for automated quality control of argo profiles. *Earth and Space Science*, 7(9):e2019EA001019.

[Yang et al., 2016] Yang, W., Jin, L., and Liu, M. (2016). Deepwriterid: An end-to-end online text-independent writer identification system. *IEEE Intelligent Systems*, 31(2):45–53.

Réduction de dimension pour les séries temporelles avec des auto-encodeurs variationnels

William Todo [1,2], Béatrice Laurent[1], Jean-Michel Loubes [1], Selmani Merwann [2]

[1] *Institut de Mathématiques de Toulouse*
[2] *Liebherr Aerospace Toulouse*
William.Todo@liebherr.com

Résumé. Dans cette présentation, nous explorons des techniques de réduction de dimensions sur des séries temporelles univariées et multivariées. En particulier, une comparaison détaillée entre les auto-encodeurs variationnels et la décomposition en ondelettes pour la réduction de dimension. On montre que les auto-encodeurs variationnels sont performants pour réduire la dimension de données de grandes dimensions comme les ECG. Ses comparaisons sont faites sur des jeux de données réels et disponibles publiquement qui présentent une grande variabilité. On utilise l'erreur de reconstruction comme métrique de performance. Enfin nous montrerons la robustesse de ces modèles avec des données bruitées que ce soit pour l'entrainement ou l'inférence. Ces tests permettant de simuler des problèmes souvent rencontré avec des séries temporelles et les VAE sont robustes à ces perturbations.

Mots-clés. VAE, Série temporelles multivariées, Reduction de dimension, ECG ...

Abstract. In this work, we explore dimensionality reduction techniques for univariate and multivariate time series data. We especially conduct a comparison between wavelet decomposition and convolutional variational autoencoders for dimension reduction. We show that variational autoencoders are a good option for reducing the dimension of high dimensional data like ECG. We make these comparisons on a real world, publicly available, ECG dataset which has lots of variability and use the reconstruction error as the metric. We then explore the robustness of these models with noisy data whether for training or inference. These tests are intended to reflect the problems that exist in real-world time series data and the VAE was robust to both tests.

Keywords. VAE, Multivariate Time Series, Dimension Reduction, ECG ...

1 Proposition de présentation

1.1 Contexte

Le fléau de la dimension est au coeur de décennies de recherches en statistiques et machine leaning, empêchant l'utilisation de nombreuses méthodes lorsque la dimension des données augmente. Il y a pourtant de plus en plus de données disponibles et étudiées de grandes dimensions comme les images, les series temporelles ou les données fonctionnelles. Ce type de données n'est pas compatible avec le machine learning traditionnel car celles-ci perdent en précision en grandes dimensions. Pour régler ce problème on ajoute souvent une étape de preprocessing où l'on utilise une technique de réduction de dimension comme par exemple l'ACP, voir : Wold, Esbensen, and Geladi (1987). Plus récemment, grace à l'émergence de techniques de deep learning comme les auto-encodeurs ou les auto-encodeurs variationnels (VAE) Kingma and Welling (2013) des techniques de réduction de dimensions plus complexes ont vu le jour et sont analysées dans Mahmud, Huang, and Fu (2020) et Dai et al. (2017).

Les series temporelles multivariées sont des données de grandes dimensions mais elles nécessitent une attention particulière. En effet les techniques de réduction de dimensions classiques ne parviennent pas à capter les aspects temporels des observations. C'est pourquoi il y a des techniques de représentation pour ce type de données. En particulier des projections dans des bases adaptées comme par exemple dans des bases d'ondelettes comme dans Yihui Liu (2009). Ou des bases basées sur la donnée comme les PCA fonctionnelles (FPCA) Di et al. (2009). La transformée en base d'ondelettes permet contrairement à la transformée de Fourier de garder une information localisée.

Depuis quelques années, des méthodes de machine learning et deep learning sont utilisées comme alternatives aux méthodes précédemment cotées. Elles sont capables de construire un embedding en faible dimension, notamment en utilisant la dernière couche d'un réseau de neurones. Il y a aussi les auto-encodeurs variationnels qui vont créer une représentation structurée des données dans un espace de faibles dimensions. L'utilisation des VAE pour des series temporelles pour réduire la dimension est un sujet encore peu traité et c'est le sujet de cette présentation.

Nous montrons ici les avantages des VAE du point de vue de la réduction de dimensions mais aussi de leur robustesse face à une méthode classique et éprouvée : la transformée en ondelettes.

1.2 Methodologie

1.2.1 Données

Ces comparaisons sont faites grace à des données d'électrocardiogrammes multivariés issues du monde réel. Le jeu de données utilisées et PTB-XL Wagner et al. (2020), il

est composé de 21837 enregistrements cliniques de 10 secondes échantillonné à 100 hz. La dimension de ces données nécessite de trouver une représentation plus petite.

1.2.2 VAE

Nous entrainons des VAE à convolution avec le framework Pytorch Paszke et al. (2019). Notons que des VAE utilisant des réseaux récurrents fonctionnent aussi, nous avons choisi les réseaux de convolution principalement car ils sont plus rapides à entrainer. Il n'y a de plus que peu de différences dans les performances de ces deux architectures.

Pour l'entrainement des modèles, nous utilisons des fenêtres de 2.56 secondes prisent aléatoirement dans les 10 secondes. Ce fenêtrage permet d'entrainer les VAE à modéliser les variations dans les ECG quelque soi la position des pics (battements de coeur). Aussi, nous changeons artificiellement le rythme cardiaque en sur/sous échantillonnant les series temporelles, rendant ainsi le modèle robuste aux changements de rythme cardiaque.

1.2.3 Transformée en ondelettes

Nous comparons les VAE avec une décomposition en base d'ondelettes comme dans Hilton (1997). Cette décomposition concentre l'information contenue dans le signal vers un petit nombre de coefficients. On peut ensuite facilement faire de la compression (avec pertes) en éliminant les coefficients trop petits. Cette compression de l'information est si puissante qu'elle est encore utilisée aujourd'hui dans le standard de compression JPEG2000 Rabbani and Joshi (2002). On utilise deux méthodes différentes pour faire les comparaisons avec les VAE. La première, qui correspond à ce qui serait fait en vrai, consiste à garder les n coefficients d'ondelettes avec la plus grande moyenne calculée sur le dataset. On appelle cette méthode la méthode globale. Les données du dataset n'étant pas synchronisées et donc les pics ne l'étant pas non plus cette technique s'apparentent en fait à prendre les n premiers coefficients d'ondelettes comme on peut le voir sur la figure 1.

La deuxième méthode consiste à prendre les n plus grands coefficients d'ondelettes pour chaque signal. Cette méthode prend en fait la meilleure configuration possible pour chaque signal. On surestime ici le taux de compression car en pratique il faudrait stocker la position de ces n coefficients. C'est pourquoi on appelle cette méthode l'Oracle.

1.3 Compression de données

Les VAE utilisent les filtres de convolutions pour combiner les informations des 12 dimensions de ces series temporelles multivariées. Les 12 signaux sont très corrélés car synchronisés autour du rythme cardiaque. C'est pourquoi les VAE sont très performants et surtout lorsqu'on impose un fort taux de compression. Les performances sont d'ailleurs même supérieures à celles de l'oracle à ondelettes on le voit Figure 2. Lorsque le taux de compression diminue les ondelettes redeviennent intéressantes. On peut expliquer

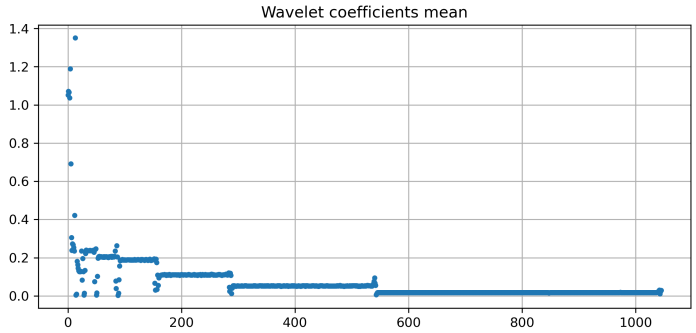


Figure 1: Moyenne des coefficients d'ondelettes

ce phénomène par une difficulté du VAE à désenchevêtrer l'information contenue dans l'espace lattent.

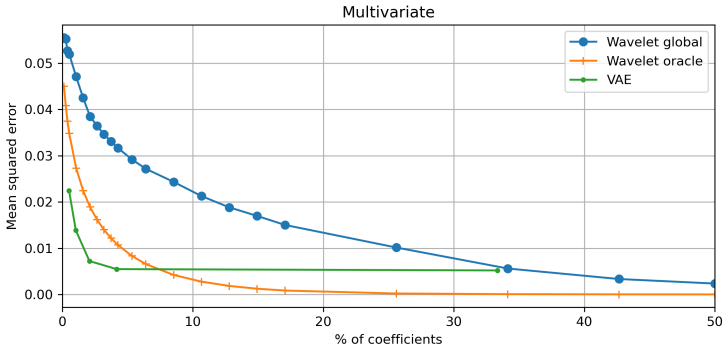


Figure 2: Comparaison des performances de compression

Pour mettre en évidence les capacités de généralisation des VAE nous testons ceux-ci sur d'autres datasets d'ECG disponible issue du challenge physionet Alday et al. (2020) composé de 10344 enregistrements de Géorgie nommé "Georgia" et 3453 enregistrements du challenge CPSC2018 F. Liu et al. (2018). Les résultats sont présentés dans la table 1 et sont similaires aux résultats obtenus sur le dataset initial (PTB-XL).

1.4 Robustesse

Pour mettre en évidence la robustesse des VAE face au bruit nous comparons les performances de reconstruction des VAE d'une part avec un VAE entrainé sur des données bruitées (pour simuler une acquisition bruitée des données) mais aussi un VAe entrainé

Dataset	Method	Mean Squared Error				
PTB	VAE	**0.02201**	**0.01381**	**0.00723**	**0.00528**	**0.00487**
	Global	0.05074	0.04625	0.03815	0.03149	0.00603
	Oracle	0.03624	0.02859	0.02002	0.01160	0.00008
Georgia	VAE	**0.01759**	**0.01223**	**0.00723**	**0.00540**	**0.00531**
	Global	0.04057	0.03752	0.03220	0.02637	0.00576
	Oracle	0.02871	0.02254	0.01612	0.00962	0.00008
China	VAE	**0.03632**	**0.02905**	**0.01972**	**0.01387**	0.01424
	Global	0.06356	0.05981	0.05313	0.04382	0.00710
	Oracle	0.04053	0.03150	0.02175	0.01239	0.00009
% Coefficients		0.5	1	2	4	33

Table 1: Performances sur d'autres datasets

sur les données saines mais testé avec des données bruitées. Les résultats sont présentés Figure 3. Dans ces deux cas de figure il n'y a pas de chute des performances ce qui confirme la robustesse au bruit des VAE.

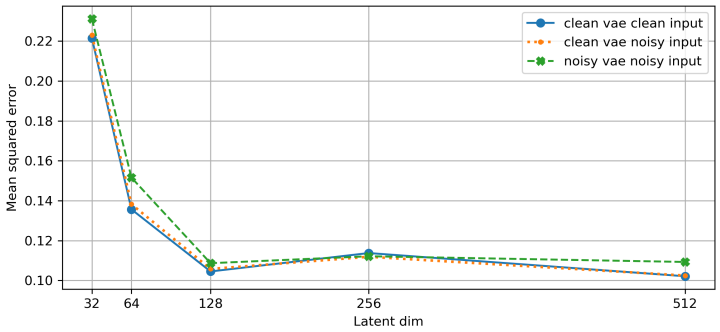


Figure 3: Etude de robustesse au bruit

1.5 Autres applications du VAE

La spécificité des VAE pour la réduction de dimension est que, contrairement à la transformée en ondelettes, les filtres sont appris et donc adaptés aux données. Comme vu précédemment c'est utile pour garder une erreur de reconstruction faible et surtout avec un fort taux de compression. Une fois que l'on a un VAE qui a bien capté les structures des données on peut utiliser la représentation en faible dimension pour appliquer des méthodes de machine learning classique. On peut notamment faire de la détection d'anomalies en combinant cette représentation en faible dimension et l'erreur de reconstruction.

References

Alday, Erick A Perez et al. (2020). "Classification of 12-lead ecgs: the physionet/computing in cardiology challenge 2020". In: *Physiological measurement* 41(12), p. 124003.

Dai, Bin et al. (2017). "Hidden talents of the variational autoencoder". In: *arXiv preprint arXiv:1706.05148*.

Di, Chong-Zhi et al. (2009). "Multilevel functional principal component analysis". In: *The annals of applied statistics* 3(1), p. 458.

Hilton, Michael L (1997). "Wavelet and wavelet packet compression of electrocardiograms". In: *IEEE Transactions on Biomedical Engineering* 44(5), pp. 394–402.

Kingma, Diederik P and Max Welling (2013). "Auto-encoding variational bayes". In: *arXiv preprint arXiv:1312.6114*.

Liu, Feifei et al. (2018). "An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection". In: *Journal of Medical Imaging and Health Informatics* 8(7), pp. 1368–1373.

Liu, Yihui (2009). "Feature extraction and dimensionality reduction for mass spectrometry data". In: *Computers in Biology and Medicine* 39(9), pp. 818–823.

Mahmud, Mohammad Sultan, Joshua Zhexue Huang, and Xianghua Fu (2020). "Variational autoencoder-based dimensionality reduction for high-dimensional small-sample data classification". In: *International Journal of Computational Intelligence and Applications* 19(01), p. 2050002.

Paszke, Adam et al. (2019). "Pytorch: An imperative style, high-performance deep learning library". In: *Advances in neural information processing systems* 32, pp. 8026–8037.

Rabbani, Majid and Rajan Joshi (2002). "An overview of the JPEG 2000 still image compression standard". In: *Signal processing: Image communication* 17(1), pp. 3–48.

Wagner, Patrick et al. (2020). "PTB-XL, a large publicly available electrocardiography dataset". In: *Scientific data* 7(1), pp. 1–15.

Wold, Svante, Kim Esbensen, and Paul Geladi (1987). "Principal component analysis". In: *Chemometrics and intelligent laboratory systems* 2(1-3), pp. 37–52.

ML et Extrêmes

In-store traffic flow forecasting using high dimensional time series methods

Siham Alaoui Belghiti [1] & Badih Ghattas [2]

[1] *AskLocala siham.alaoui@s4m.io 1*
[2] *I2M badih.ghattas@univ-amu.fr 2*

Abstract. Predicting the flow of visits to stores is of great importance to businesses since it can help to optimize and streamline their operations by adapting to expected traffic volumes. In this article, we frame the problem of predicting traffic flow as a high dimensional time series prediction problem. We pre-process the data in order to enable us to utilize state of the art high dimensional time series algorithms to predict future time points. We then analyze and compare in detail, several state-of-the-art forecasting methods. We found that a high dimensional approach utilising deep learning and learning both global and local structure works the best in our case.

Keywords. High dimensional time series, Sequential models, Deep learning, Visit in store, Ads.

1 Introduction

Predicting the flow of traffic to physical stores enables the optimization of store operations (Lam, Vandenbosch, and Pearce (1998) and Kabak et al. (2008)) and thereby enhances business outcomes (Perdikaki, Kesavan, and Swaminathan (2012)). This is a non-trivial problem since traffic can be affected by several complex and dynamic factors such as weather, daily events and holidays. Furthermore, data used to predict future traffic consists of many related rather than independent samples of time series. Therefore, the available datasets may contain a large number of correlated time series over several thousand time points. Such type of data can be termed high-dimensional time series.
Classical statistical time series forecasting methods, such as ARIMA (Luceño and Peña (2008)), exponential smoothing (Gardner Jr (1985)), Box-Jenkins and other variants, have been developed with a focus on forecasting individual or small number of time-series. Similarly, several traditional machine learning approaches such as tree based algorithms as well as neural network (NN) based models have also been applied to the problem of time series forecasting (Ismail Fawaz et al. (2019)). Studies have focused on recurrent neural networks (RNN) and their many variants e.g. LSTM and GRU (Fu, Zhang, and Li (2016)). These NN-based approaches have also been applied to the specific problem of predicting traffic flow to stores. The typical approach taken is to create an individual model per store or treat times series data as independent samples rather than considering common structure. (Abrishami and Kumar (2018)) attempted to capture some global

structure, the training data of a given store was augmented with features from its nearest neighbors. In this manner, some information across time series was shared. However, the shared information were supplied to the model rather than learnt. Further, the model is limited to extracting information from the shared data based on the specific type of feature engineering applied.
Recently, several studies have observed the importance of learning global structure in high dimensional time series forecasting problems. Temporal regularized matrix factorization (TRMF) (Yu, Rao, and Dhillon (2016)) is a variation on Matrix factorization (MF) that introduces a temporally sensitive regularization component to the objective function. TRMF is able to learn latent structure while preserving local dynamics and temporal dependencies within the individual time series.
DeepAR (Salinas et al. (2020)) presents a probablistic approach to learning global structure that utilizes deep neural networks to predict the parameters of a conditional distribution from which forecasts for future time points are sampled. DeepGLO (Sen, Yu, and Dhillon (2019)) is a method that attempts to capture the common structure within the data by utilizing TRMF to learn the global component while attempting to also capture local nuances using a NN-based model. The results from this study have shown the importance of learning both local and global structure in the data for performant forecasting.
In this paper, we collected traffic data from store locations. We then explored and applied state-of-the-art high dimensional time series forecasting methods on our data set. Specifically, we used the techniques of Probabilistic Forecasting with Autoregressive Recurrent Networks (deepAR) (Salinas et al. (2020)), a Deep Neural Network Approach to High-Dimensional Time Series Forecasting (DeepGLO) (Sen, Yu, and Dhillon (2019)) and Temporal Regularized Matrix Factorization technique (TRMF) (Yu, Rao, and Dhillon (2016)). We analyzed each of these approaches along several metrics.
The rest of the article is organized as follows: in section 2 we formulate the time series problem more formally and then present the details of the state-of-the-art methods applied to our datasets; section 3 lists details of the dataset, the metrics utilized as well as the results of our experimentation; in section 4 we present a summary and conclusion of the study.

2 Forecasting high dimensional times series

Given N univariate time series of the form $Y = \{\mathbf{y}_{i,1:T}\}_{i=1}^{N}$ where $\mathbf{y}_{i,1:T} = [y_{i,1}, y_{i,2}, \cdots, y_{i,T}]$ and $y_{i,t} \in R$ represents the value of the i-th time series at time t. Our goal is to forecast $[y_{i,t_0}, y_{i,t_0+1}, \ldots, y_{i,T}]$ given its past $[y_{i,1}, \ldots, y_{i,t_0-2}, y_{i,t_0-1}]$. In certain cases, a high dimensional time series can contain a set of corresponding time varying covariates that can be item-dependent (e.g., store type, store zipcode), time-dependent (e.g day of week), or both (e.g. day of the week with a specific store type). We compare several state of the art high dimensional time series forecasting methods and use a vector autoregressive

model (VAR) as a baseline since this is a variation on a standard model for time series forecasting. VAR generalizes the traditional Auto-Regressive Integrated Moving Average (ARIMA) model for multivariate time series by forecasting a vector of values using a linear function (Hyndman and Athanasopoulos (2018)).

Deep AutoRegression (DeepAR) is a probabilistic forecasting model developed by Amazon. It combines probablistic modeling with neural networks where the neural network is trained to output the best parameters of a conditional probability distribution from which a future time point to be forecasted is sampled. Given a time series $\{\mathbf{y}_{i,1:t_0}\}_{i=1}^{N}$ and covariate $\mathbf{x}_{i,1:T}$, the conditional probability distribution is defined as follows:

$$P(y_{i,t_0:T} \mid y_{i,1:t_0-1}, \mathbf{x}_{i,1:T}) = \prod_{t=t_0}^{T} p\left(y_{i,t} \mid y_{i,1:t-1}, \mathbf{x}_{i,1:T}\right) = \prod_{t=t_0}^{T} l_\theta\left(y_{i,t}; \mathbf{h}_{i,t}\right) \tag{1}$$

Where l_θ is a likelihood function parameterized by θ, and $\mathbf{h}_{i,t}$ is considered to be the output of a recurrent network with LSTM cells given by: $\mathbf{h}_{i,t} = h(\mathbf{h}_{i,t-1}, y_{i,t}, \mathbf{x}_{i,t})$.

Temporal Regularized Matrix Factorization (TRMF) is a variation on the MF approach that respects the temporal order of the columns in the time series input data. The set of N time series of length T which are presented by: $\mathbf{Y} = \{y_{i,1:T}\}_{i=1}^{N}$ may be factorized as: $\mathbf{Y} = F \times X$, where F and X represent the latent factors where X is time dependent. F and X may be obtained by minimizing the penalized quadratic loss:

$$\sum_{(i,t)\in\Omega} \left(Y_{it} - f_i^T x_t\right)^2 + \lambda_f R_f(F) + \lambda_x \mathcal{T}_M(X|\Theta) \tag{2}$$

Where $R_f(F)$ and $\mathcal{T}_M(X|\Theta)$ are the regularizers for F and X respectively. TRMF proposed a regularizer $\mathcal{T}_M(X|\Theta)$ which encourages the structure induced by M_Θ. Temporal dependencies among x_t are modeled as: $x_t = M_\Theta\left(\{x_{t-l} : l \in \mathcal{L}\}\right) + \boldsymbol{\epsilon}_t$. Where $\boldsymbol{\epsilon}_t$ is a Gaussian noise vector, and M_Θ is the time-series model parameterized by $\mathcal{L}$ and Θ. $\mathcal{L}$ is a set containing the lag indices l, l is a the temporal indices of time shifts which refer to a dependence between t-th and $(t-l)$-th time points. When the Θ is unknown, Θ is treated as another set of variables and include another regularizer $R_\theta(\Theta)$.

A deep neural network approach to high-dimensional time series forecasting (DeepGLO) is a hybrid model that consists of a combination of two parts namely: i) a local component represented by a temporal convolution network (TCN) (Lea et al. (2017)) that captures the local dynamics of individual time series, ii) a global component represented by the TRMF method which uses MF to capture latent global structure. The output of the global model is fed as a covariate, along with other covariates to the local TCN based model and in this manner DeepGLO presents a hybrid global-local approach to model high dimensional time series.

3 Empirical results and Analysis

In this section, we provide details on the experimental setup, the metrics used for evaluating and comparing the various methods as well as the results.

3.1 Experimental Setup

We collected five months of traffic flow data from May to September 2021 from 52 stores of a single brand around the United States. To determine the number of people considered visitors to a point of sale, we rely on geolocation data supplied by a third party. We aggregate the data to 15 minutes before conducting our analysis. After aggregation, the dataset contains a total of 14,688 time points for each time series. The time series are of varying magnitude and variances. We used data from 2021-05-01 to 2021-09-23 as a training set and the data of the following 7 days to 2021-09-30 as a test set. We set the size of the forecast horizon to 672 time points, which it corresponds to one week of data. Due to the scale sensitive nature of the models analyzed in our study, we compare the performance of our models on raw vs standardized data where standardization simply removes the mean and divides by the standard deviation. There are several implementations available for each of the models studied. For DeepAR, we choose the Gluonsts package [1]and utilized LSTM units. For TRMF and DeepGlo, we used the DeepGLO code base[2]. All models are trained five times on the initial training set and used for subsequent prediction for all moving windows including a seven day horizon.
The models were evaluated using several metrics including root mean squared error (RMSE) which is insensitive to outliers, mean absolute error (MAE) and mean absolute percentage error (MAPE) that provide a more intuitive sense of the quality of the models.

3.2 Results

The objective for this set of tests was to find the model that yielded the lowest error metrics, including RMSE, MAE and MAPE. As can be seen in table 1 the best performance is achieved by using a dataset that has high variance outliers removed. On this dataset the best performing model was deepGLO. In general, most models perform better with raw data with the exception of VAR and TRMF, that perform better with standardized data. These findings are consistent with earlier studies that defined these methods (Sen, Yu, and Dhillon (2019) and Salinas et al. (2020)) where it was found that leaving the data unnormalized/unstandardized yielded the best results.

In order to determine the sensitivity of each of the forecasting methods studied to high variance outliers, we trained the models with and without 10 high variance outliers

[1]https://github.com/awslabs/gluon-ts.git
[2]https://github.com/rajatsen91/deepglo

Data	Algorithm	RMSE/ MAE / MAPE	
		Non Standardized	Standardized
Stores N=52	DeepGLO	**84.80** / **62.43** / **54.45**	90.10 / 68.39 / 61.64
	DeepAR	162.76 / 133.59 / 40.15	441.15 / 117.74 / 40.39
	TRMF	183.95 / 157.69 / 113.18	148.16 / 124.79 / 122.44
	VAR	442.85 / 354.43 / 379.64	183.96 / 153.38 / 137.26
	AR (local)	166.87 / 139.90 / 86.46	448.94 / 120.58 / 79.75
Stores N=42 Time series with same variance	DeepGLO	21.41 / **15.58** / **43.80**	20.99 / 16.00 / 44.95
	DeepAR	25.33 / 18.64 / 63.03	24.68 / 14.36 / 93.60
	TRMF	24.88 / 19.31 / 44.96	103.05 / 86.78 / 19.09
	VAR	60.15 / 48.62 / 192.79	47.70 / 37.60 / 160.97
	AR (local)	26.73 / 21.51 / 95.79	27.13 / 18.12 / 89.81

Table 1: Empirical Results

that we found in our data. As can be seen in table 1 removing these outliers improves the performance of all of the models significantly. There is a drop in error using all of the metrics, and the winning model namely deepGLO, achieves the best performance among all the models in this condition when high variance outliers are removed.

4 Conclusion

In this study we framed the task of predicting traffic to stores as a high-dimensional time series prediction problem and applied state of the art methods to forecast traffic flow. The high dimensional time-series prediction methods presented in this article are able to handle the large datasets collected even with large-sized store visit data (a number of stores greater than 52 studied here, and also over a period greater than five months). We explored these models by considering different evaluation metrics MAE, RMSE and MAPE. Based on these metrics and the data analyzed for a single brand of stores we found that deepGLO outperformed other methods for forecasting traffic flow. Further, we found that all of the methods applied were sensitive to the scale of the data as well as the presence of outliers. For deepGLO and some of the other methods we found that leaving the data un-standardized performed the best and this is consistent with what was found in (Sen, Yu, and Dhillon (2019)). Further, we found a considerable gain in forecasting performance by removing time series that had very high variance, demonstrating that the methods studied here are sensitive to large variance difference between time series that are used for training.

We demonstrate that high-dimensional time series methods are a promising approach to forecasting traffic data and specifically deep learning based methods outperform classical autoregressive approaches.

References

Abrishami, Soheila and Piyush Kumar (2018). "Using real-world store data for foot traffic forecasting". In: *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, pp. 1885–1890.

Fu, Rui, Zuo Zhang, and Li Li (2016). "Using LSTM and GRU neural network methods for traffic flow prediction". In: *2016 31st Youth Academic Annual Conference of Chinese Association of Automation (YAC)*. IEEE, pp. 324–328.

Gardner Jr, Everette S (1985). "Exponential smoothing: The state of the art". In: *Journal of forecasting* 4.1, pp. 1–28.

Hyndman, Rob J and George Athanasopoulos (2018). *Forecasting: principles and practice*. OTexts.

Ismail Fawaz, Hassan et al. (2019). "Deep learning for time series classification: a review". In: *Data mining and knowledge discovery* 33.4, pp. 917–963.

Kabak, Özgür et al. (2008). "Efficient shift scheduling in the retail sector through two-stage optimization". In: *European Journal of Operational Research* 184.1, pp. 76–90.

Lam, Shunyin, Mark Vandenbosch, and Michael Pearce (1998). "Retail sales force scheduling based on store traffic forecasting". In: *Journal of Retailing* 74.1, pp. 61–88.

Lea, Colin et al. (2017). "Temporal convolutional networks for action segmentation and detection". In: *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 156–165.

Luceño, Alberto and Daniel Peña (2008). "Autoregressive Integrated Moving Average (ARIMA) Modeling". In: *Encyclopedia of Statistics in Quality and Reliability* 2.

Perdikaki, Olga, Saravanan Kesavan, and Jayashankar M Swaminathan (2012). "Effect of traffic on sales and conversion rates of retail stores". In: *Manufacturing & Service Operations Management* 14.1, pp. 145–162.

Salinas, David et al. (2020). "DeepAR: Probabilistic forecasting with autoregressive recurrent networks". In: *International Journal of Forecasting* 36.3, pp. 1181–1191.

Sen, Rajat, Hsiang-Fu Yu, and Inderjit S Dhillon (2019). "Think globally, act locally: A deep neural network approach to high-dimensional time series forecasting". In: *Advances in neural information processing systems* 32.

Yu, Hsiang-Fu, Nikhil Rao, and Inderjit S Dhillon (2016). "Temporal regularized matrix factorization for high-dimensional time series prediction". In: *Advances in neural information processing systems* 29.

Estimation de paramètres pour un modèle de propagation de la fièvre typhoïde à Mayotte

Ibrahim Bouzalmat [1] & Benoîte de Saporta [1] & Solym Manou-Abi [2]

[1] *IMAG, Univ Montpellier, CNRS, Montpellier, France,*
ibrahim.bouzalmat@umontpellier.fr , benoite.de-saporta@umontpellier.fr
[2] *IMAG, Univ Montpellier, CNRS, CUFR Mayotte, France,*
solym.manou-abi@univ-mayotte.fr

Résumé. L'objectif de ce travail est de modéliser la propagation de la fièvre typhoïde à Mayotte à partir d'un jeu de données d'hospitalisations fourni par l'Agence Régionale de Santé. Nous utilisons un processus de naissance et mort linéaire avec immigration comptabilisant les personnes infectées par la maladie. L'objectif est alors d'estimer les taux de contamination de personne à personne, contamination par l'environnement et guérison à partir des données. L'originalité de notre approche et la difficulté du problème provient de deux sources. D'une part, les observations ne sont pas disponibles en temps continu mais seulement à des dates fixes (hospitalisations journalières), et d'autre part à ces dates le nombre total de personnes infectées n'est pas observé directement, seuls les nouveaux cas depuis la date précédente sont comptabilisés. Pour traiter ces spécificités, nous obtenons d'abord une expression explicite des lois de transition à pas fixe pour construire des estimateurs de nos paramètres basés sur les fréquences des transitions pour le nombre d'infectés. Ensuite nous nous plaçons dans le cadre des chaînes de Markov cachées et adaptons l'algorithme de Baum-Welch à notre cas.

Mots-clés. Processus de naissance et mort linéaire avec immigration, Modèle de Markov Caché, Algorithme de Baum-Welch, Estimation paramétrique.

Abstract. The objective of this work is to model the spread of typhoid fever in Mayotte from a set of hospitalization data provided by the Regional Health Agency. We use a linear birth and death process with immigration counting people infected with the disease. The objective is then to estimate the rates of contamination from person to person, contamination by the environment and recovery from the data. The originality of our approach and the difficulty of the problem comes from two sources. On the one hand, the observations are not available in continuous time but only at fixed dates (daily hospitalizations), and on the other hand at these dates the total number of infected people is not observed directly, only the new cases since the previous date are counted. To deal with these specificities, we first obtain an explicit expression of the fixed-step transition distributions to construct estimators of our parameters based on the frequencies of the transitions for the number of infected. Then we place ourselves in the framework of hidden Markov chains and adapt the Baum-Welch algorithm to our case.

Keywords. Linear birth-death process with immigration, Hidden Markov Model, Baum-Welch algorithm, Parametric estimation.

1 Introduction.

L'objectif de ce travail est de modéliser la propagation de la fièvre typhoïde à Mayotte à partir d'un jeu de données d'hospitalisations entre 2018 et 2020 fourni par l'Agence Régionale de Santé. La fièvre typhoïde est une maladie à déclaration obligatoire, endémique à Mayotte avec une trentaine de cas par an. Elle se transmet par contact de personne à personne ou par ingestion d'eau ou d'aliments contaminés. Comme le nombre de personnes infectées simultanément est faible, et que l'environnement peut être source de contamination, nous modélisons sa propagation à l'aide d'un processus de naissance et mort linéaire avec immigration comptabilisant les personnes infectées par la maladie. L'objectif est alors d'estimer les taux de contamination de personne à personne (naissance), contamination par l'environnement (immigration) et guérison (mort) à partir des données.

L'originalité de notre approche et la difficulté du problème provient de deux sources. D'une part, les observations ne sont pas disponibles en temps continu mais seulement à des dates fixes (hospitalisations journalières). De plus, à ces dates le nombre total de personnes infectées n'est pas observé, seuls les nouveaux cas depuis la date précédente sont comptabilisés. L'inférence de paramètres pour des processus de naissance et mort linéaires est facile à mener par maximum de vraisemblance lorsqu'on observe les dates de survenue des naissances et des morts et en l'absence d'immigration puisqu'on peut alors utiliser une propriété de branchement (Keiding et al. (1975)). Pour des processus de naissance et mort plus généraux, Crawford et Suchard (2012) et Crawford et al (2014) proposent d'utiliser la théorie des fractions continues et l'algorithme espérance-maximisation (EM), Xu et al (2015) utilisent l'algorithme EM pour l'inférence de processus de branchement multi-types. Dans tous ces travaux, les dates de naissance et de mort sont toujours observées, ce qui les rend inapplicables dans notre cas. Dans une première étape, pour prendre en compte le fait que les observations sont discrètes, nous utilisons les équations de Kolmogorov et les fonctions génératrices des moments des lois de transition à pas fixe pour calculer l'expression explicite de ces probabilités de transition en fonction de nos trois paramètres d'intérêt. Ces formules peuvent ensuite être inversées pour construire des estimateurs basés sur les fréquences des transitions pour le nombre d'infectés. Dans une deuxième étape, pour prendre en compte le fait que seules les nouvelles infections sont comptabilisées, nous nous plaçons dans le cadre des chaînes de Markov cachées et nous adaptons l'algorithme de Baum-Welch (Baum et Petrie (1966)) à nos données. Ce problème est non standard car la loi d'émission des observations ne dépend pas que de l'état courant de la chaîne mais de toute sa trajectoire (continue) depuis la dernière observation.

Ce document est organisé comme suit. Dans la section 2, nous modélisons la dynamique de transmission de la typhoïde par un modèle de naissance et mort avec immigration. Dans la section 3, nous dérivons la procédure d'estimation des paramètres de notre modèle en utilisant la méthode de Markov caché et un algorithme de Baum-Welch adapté à nos observations. Dans la section 4, nous appliquons notre approche d'estimation aux données de Mayotte pour obtenir les estimations désirées et discutons les résultats obtenus.

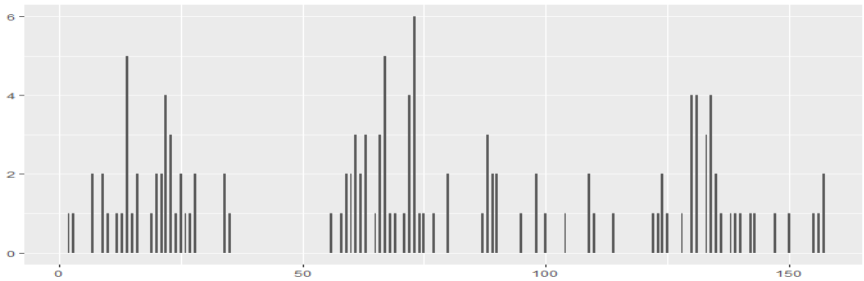


FIGURE 1 – Nombre hebdomadaire de nouveaux cas confirmés de fièvre typhoïde entre 2018 et 2020 à Mayotte (données Agence Régionale de Santé).

2 Modélisation de la dynamique de la typhoïde

La prévalence de la fièvre typhoïde à Mayotte est caractérisée par des niveaux faibles et fluctuants, et des résurgences de la maladie après des plages sans nouvelles infections, voir Fig. 1. Un modèle stochastique de type processus de comptage est donc bien adapté pour modéliser ces données. Parmi ces modèles, l'un des plus simples et pertinent, avec peu de paramètres est le processus de naissance et mort linéaire avec immigration. Un tel processus $(X_t)_{t\geq 0}$ comptabilise le nombre d'individus infectés. Chaque individu contamine un nouvel individu à un taux (de naissance) $\lambda > 0$ et guérit avec un taux (de mort) $\mu > 0$ indépendamment des autres individus, ce qui génère des taux de naissance et mort linéaires pour la population globale. La contamination via l'environnement génère également de nouveaux infectés via un processus de Poisson de taux (d'immigration) $\nu > 0$. Notre objectif est d'estimer ces trois paramètres λ, μ, ν.

Le nombre d'infectés X_t n'est pas directement observé. Nous avons à notre disposition dans le jeu de données uniquement les cumuls de nouvelles infections par jour. Rappelons qu'un processus de naissance et mort ne peut faire que des sauts d'amplitude $+1$ (nouvelle contamination) ou -1 (nouvelle guérison). Notons $N_{]s,t]}$ le nombre des sauts d'amplitude $+1$ du processus $(X_t)_{t\geq 0}$ sur l'intervalle $]s,t]$. Alors le processus joint $(X_t, N_{]0,t]})_{t\geq 0}$ est un processus de Markov de saut pur et les probabilités $p_{i,(j,n)}(t) := \mathbb{P}\left(X_t = j, N_{]0,t]} = n \mid X_0 = i\right)$ satisfont l'équation de Kolmogorov suivante : si $j > i + n$, alors $p_{i,(j,n)}(t) = 0$ et si $j \leq i + n$ alors

$$\frac{dp_{i,(j,n)}(t)}{dt} = -((\lambda+\mu)i+\nu)p_{i,(j,n)}(t) + (\lambda i+\nu)p_{i+1,(j,n-1)}(t) + (\mu i)p_{i-1,(j,n)}(t), \qquad (1)$$

avec la condition initiale $p_{i,(j,n)}(0) = \delta_{i=j}\delta_{n=0}$. Cette équation n'a pas de solution analytique. Nous ne pouvons pas exprimer la distribution marginale des observations en fonction de λ, μ et ν. Ainsi, il n'est pas possible de mettre en oeuvre une méthode d'estimation directe par maximum de vraisemblance avec ce schéma d'observations. Pour pallier ce

problème, dans un premier temps nous allons considérer le problème d'estimation lorsque X_t est observé à des dates régulières. Dans un second temps, nous replacerons notre cadre d'observation dans la famille des modèles de Markov cachés.

Soit Δt un pas de temps fixé ($\Delta t = 1$ jour pour nos données). Alors, le processus $(X_{n\Delta t})_{n\in\mathbb{N}}$ est une chaîne de Markov de matrice de transition $P = (p_{i,j})$. Les coefficients de P peuvent être calculés explicitement en fonction de λ, μ, ν en résolvant une équation de Kolmogorov et en passant par les fonctions génératrices des moments. Comme nous avons 3 paramètres à estimer, et que les nombres de nouveaux cas journaliers observés sont fréquemment nuls, nous utilisons l'expression explicite des paramètres λ, μ et ν en fonction uniquement des probabilités de transition $p_{0,0}, p_{0,1}$ et $p_{1,0}$:

$$\lambda = \frac{ln\left(\frac{u}{p}\right)(p-1)}{\Delta t\,(p-u)}, \quad \mu = \frac{ln\left(\frac{u}{p}\right)(u-1)}{\Delta t\,(p-u)}, \quad \nu = r\lambda, \tag{2}$$

avec

$$p = \frac{p_{0,1}}{p_{0,0}ln(p_{0,0})} W\left(\frac{p_{0,0}^{\left(\frac{p_{0,0}}{p_{0,1}}+1\right)} ln(p_{0,0})}{p_{0,1}}\right), \quad r = \frac{ln(p_{0,0})}{ln(p)}, \quad u = 1 - \frac{p_{1,0}}{p_{0,0}},$$

où W est la fonction W de Lambert. Ainsi, avec une estimation des probabilités de transition $p_{0,0}, p_{0,1}$ et $p_{1,0}$, on obtient une estimation de nos paramètres d'intérêt par plug-in.

3 Estimation des taux de contamination et de guérison

Notons $O_n := N_{[(n-1)\Delta t, n\Delta t[}$ pour $n \in \mathbb{N}^*$ le nombre de nouveaux infectés sur l'intervalle de temps $[(n-1)\Delta t, n\Delta t[$ qui correspond aux observations disponibles. Ici, O_n dépend de toute la trajectoire continue de X_t sur l'intervalle $[(n-1)\Delta t, n\Delta t[$, et pas seulement de $X_{n\Delta t}$. Le schéma d'émission est donc non standard, et l'approche classique doit être adaptée en conséquence. Par la propriété de Markov, on peut montrer que la loi de O_n dépend en fait de $X_{n-1\Delta t}$ et $X_{n\Delta t}$, et que conditionnellement aux $(X_{k\Delta t})_{k\leq n}$, les $(O_k)_{k\leq n}$ sont indépendants, ce qui donne le schéma d'émission de la Figure 2. Posons $Z_n = (X_{(n-1)\Delta t}, X_{n\Delta t})$ pour $n \in \mathbb{N}^*$. Alors $(Z_n)_{n\geq 1}$ est une chaîne de Markov à valeurs dans $\mathbb{N}^2$, et pour cette chaîne, le schéma d'émission des observations (O_n) redevient standard puisque O_n ne dépend que de Z_n. Soit $M = (Q, \psi, \rho)$ les paramètres de ce modèle de Markov caché où Q est la matrice de transition du processus d'état (Z_n), calculée à partir de la propriété de Markov sur le processus $(X_{n\Delta t})$: $Q_{(i,j),(i',j')} = p_{i',j'}\delta_{i'=j}$, ψ est la probabilité d'émission du processus O sachant Z

$$\psi_{(i,j)}(o) = \mathbb{P}(O_n = o | Z_n = (i,j)) = \frac{p_{i,(j,o)}(\Delta t)}{p_{i,j}},$$

et ρ est la loi de l'état initial Z_1 : $\rho_{i,j} = p_{i,j}\pi_i$, où $\pi_i := \mathbb{P}(X_0 = i)$ est la loi initiale de (X_t). En résolvant numériquement (1), nous pouvons calculer la probabilité $p_{i,(j,o)}(\Delta t)$ et en déduire la probabilité d'émission $\psi_{(i,j)}(o)$.

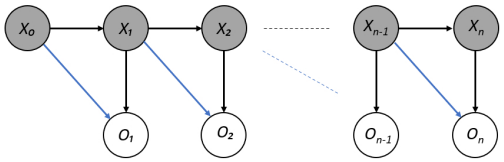


FIGURE 2 – Diagramme d'émission entre les nombre d'infectés X_k à la date $k\Delta t$ et les nombre de nouveaux cas cumulés O_k entre les dates $(k-1)\Delta t$ et $k\Delta t$.

Étant donné une suite d'observations $(O_n)_{0\leq n\leq T}$, nous cherchons à déterminer par maximum de vraisemblance les paramètres $M = (Q, \psi, \rho)$ du modèle, et plus précisément les trois coefficients de la matrice de transition P qui apparaissant dans les équations (2). Malheureusement, l'utilisation directe de l'algorithme de Baum-Welsh sur la chaîne (Z_n) sans tenir compte de sa structure particulière ne donne pas de résultats cohérents sur données simulées. Nous avons donc récrit toutes les formules de récurrence pour les quantités d'intérêt dans le cas particulier où Z_n correspond à deux états successifs d'une chaîne de Markov. Finalement, partant d'un modèle initial $M^{(0)} = (Q^{(0)}, \psi^{(0)}, \rho^{(0)})$, on obtient à chaque étape une nouvelle estimation des paramètres du modèle par les formules

$$p_{i,j}^{(n)} = \frac{1}{N}\sum_{i'=0}^{N} Q_{(i',i),(i,j)}^{(n)},$$

où

$$Q_{(i,j),(i',j')}^{(n)} = \frac{\sum_{t=0}^{T-1}\xi_{(i,j),(i',j')}(t)}{\sum_{t=0}^{T-1}\gamma_{(i,j)}(t)}\delta_{i'=j}, \quad \psi_{(i,j)}^{(n)}(o) = \frac{\sum_{t=0}^{T}\mathbb{1}_{\{O_t=o\}}\gamma_{(i,j)}(t)}{\sum_{t=0}^{T}\gamma_{(i,j)}(t)}, \quad \rho_{i,j}^{(n)} = \gamma_{(i,j)}(0),$$

avec

$$\gamma_{(i,j)}(t) = \frac{\alpha_{(i,j)}(t)\beta_{(i,j)}(t)}{\mathbb{P}(O_{0:T} = o_{0:T}|M^{n-1})}, \quad \xi_{(i,j),(i',j')}(t) = \frac{\alpha_{(i,j)}(t)p_{(i',j')}^{(n-1)}\psi_{(i',j')}^{(n-1)}(o_{t+1})\beta_{(i',j')}(t+1)}{\mathbb{P}(O_{0:T} = o_{0:T}|M^{(n-1)})}\delta_{i'=j}$$

$$\mathbb{P}(O_{0:T} = o_{0:T}|M^{(n-1)}) = \sum_{i=0}^{N}\sum_{j=0}^{N}\alpha_{(i,j)}(t)\beta_{(i,j)}(t)$$

$$\alpha_{(i,j)}(0) = p_{i,j}\pi_i\psi_{(i,j)}^{(n-1)}(o_0), \quad \alpha_{(i,j)}(t) = \psi_{(i,j)}^{(n-1)}(o_t)\, p_{i,j}^{(n-1)}\sum_{i'=0}^{N}\alpha_{(i',i)}(t-1) \quad \text{(Forward probability)}$$

$$\beta_{(i,j)}(T) = 1, \quad \beta_{(i,j)}(t) = \sum_{j'=0}^{N} p_{j,j'}\psi_{(j,j')}^{(n-1)}(o_{t+1})\,\beta_{(j,j')}(t+1) \quad \text{(backward probability)}.$$

L'algorithme recalcule les opérations de ré-estimation M fois jusqu'à un certain critère d'arrêt. Les estimateurs des paramètres λ, μ et ν sont obtenues par le plug-in à partir de $p_{0,0}^{(M)}, p_{0,1}^{(M)}$ et $p_{1,0}^{(M)}$ et des équations (2).

4 Résultats sur les données d'hospitalisation à Mayotte

Nous disposons de 1086 cumuls journaliers de nouveaux cas positifs de la fièvre typhoïde à Mayotte pour les années 2018 à 2020. Ces cumuls compris entre 0 et 3. Comme il s'agit d'une maladie à déclaration obligatoire, nous considérons que tous les nouveaux cas positifs sont bien observés. A partir de la littérature, nous choisissons les paramètres initiaux dans les intervalles suivants $\lambda^{(0)} \in [0.05, 0.08], \mu^{(0)} \in [0.11, 0.25]$ et $\nu^{(0)} \in [0.015, 0.03]$. Pour l'initialisation de $\rho^{(0)}$, nous avons supposé arbitrairement que X_0 suit la loi invariante du processus de naissance et mort de paramètres $\lambda^{(0)}, \mu^{(0)}$ et $\nu^{(0)}$. Les valeurs des paramètres maximisant la vraisemblance des observations pour toutes les valeurs initiales testées sont $\hat{\lambda} = 0.05, \hat{\mu} = 0.11$ et $\hat{\nu} = 0.015$. En particulier, on a $\hat{\lambda} < \hat{\mu}$ qui correspond à un régime récurrent positif sous lequel nos estimations sont valables, et où le processus retourne à 0 infiniment souvent ce qui est cohérent avec les observations. Le temps moyen de guérison $1/\hat{\mu} = 9$ est cohérent avec les durées d'hospitalisations et de traitements, alors que celles-ci n'ont pas été utilisées dans l'estimation.

En conclusion, nous avons présenté une méthode d'estimation générique pour les paramètres d'un processus de naissance et mort avec immigration lorsque seuls les cumuls de nouveaux cas à date périodiques sont observés. Nous avons implémenté cette méthode sur des données de fièvre typhoïde à Mayotte et obtenu des résultats cohérents avec la réalité du terrain. Ces estimations pourront permettre par exemple de déterminer des seuils critiques de nombres de malades à partir desquels une saturation hospitalière risque d'arriver avec forte probabilité dans une fenêtre de temps donnée.

Bibliographie

L. E. Baum and T. Petrie (1966), Statistical inference for probabilistic functions of finite state markov chains, *The annals of mathematical statistics*, vol. 37, no. 6, pp. 1554-1563.

N. Keiding et al. (1975), Maximum likelihood estimation in the birth-and-death process, *The Annals of Statistics*, vol. 3, no. 2, pp. 363-372.

F. W. Crawford and M. A. Suchard (2012), Transition probabilities for general birth ?death processes with applications in ecology, genetics, and evolution, *J. of mathematical biology*, vol. 65, no. 3, pp. 553-580.

F. W. Crawford, V. N. Minin, and M. A. Suchard (2014), Estimation for general birth-death processes, *J.of the American Statistical Association*, vol. 109, no. 506, pp. 730-747.

J. Xu, P. Guttorp, M. Kato-Maeda, and V. N. Minin (2015), Likelihood-based inference for discretely observed birth-death-shift processes, with applications to evolution of mobile genetic elements, *Biometrics*, vol. 71, no. 4, pp. 1009-1021.

Crowdsourcing label noise simulation on image classification tasks

Tanguy Lefort [1,a] & Benjamin Charlier[1,b] & Joseph Salmon[1,2,c] & Alexis Joly[3,d]

[1] *IMAG, Univ. Montpellier, CNRS, Montpellier, France*
[2] *Institut Universitaire de France (IUF)* [3] *Inria, LIRMM, CNRS Montpellier*
[a] *tanguy.lefort@umontpellier.fr* [b] *benjamin.charlier@umontpellier.fr*
[c] *joseph.salmon@umontpellier.fr* [d] *alexis.joly@inria.fr*

Résumé. Il est courant de collecter des données étiquetées en ayant recourt à la production participative *crowdsourcing*. Cependant, la qualité des étiquettes dépend fortement de la difficulté des tâches et des capacités des contributeurs. Avec ces données, le manque de vérité terrain rend difficile l'évaluation de la qualité des annotations. Il existe peu de données de ce type disponibles publiquement, et encore moins sur des tâches de difficulté hétérogène avec les réponses de tous les contributeurs avant l'étape d'agrégation. Nous proposons ici un nouveau cadre de simulation de *crowdsourcing* où la qualité peut être contrôlée. Cela permet d'évaluer l'apprentissage de différentes stratégies de manière empirique à partir des étiquettes collectées. Notre objectif est de séparer les différentes sources de bruit dans l'agrégation des étiquettes: les contributeurs ne fournissant aucune information sur la vraie étiquette et ceux peu performants, utiles pour des tâches faciles.

Mots-clés. Apprentissage participatif, simulation de foule, détection de spammeur

Abstract. It is common to collect labelled datasets using crowdsourcing. Yet, labels quality depends deeply on the task difficulty and on the workers abilities. With such datasets, the lack of ground truth makes it hard to assess the quality of annotations. There are few open-access crowdsourced datasets, and even fewer that provide both heterogeneous tasks in difficulty and all workers answers before the aggregation. We propose a new crowdsourcing simulation framework with quality control. This allows us to evaluate different empirical learning strategies empirically from the obtained labels. Our goal is to separate different sources of noise: workers that do not provide any information on the true label against poorly performing workers, useful on easy tasks.

Keywords. Crowdsourcing, crowd simulation, spam detection

1 Introduction

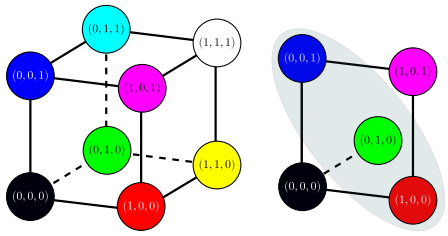
Let us consider a supervised learning setting, with a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n_t}$ containing n_t tasks $x_i \in \mathcal{X}$ (vectorized images) with their corresponding label $y_i \in \mathcal{Y}$ (with K card$(\mathcal{Y})$ classes), commonly annotated by a crowd. With a crowd of n_w workers, it yields multiple labels $\{y_i^{(j)}\}_{j=1}^{n_w}$ for each image x_i. Learning from multiple labels is challenging (Rodrigues

et al., 2013; Khetan et al., 2018), since the ground truth is unknown and trust needs to be evaluated.

Platforms like ThePlantGame[1] expect users to label (plant) images hard to identify, possibly inducing weak estimators and very imbalanced datasets (in the number of instances per class but also in the number of proposed labels per tasks). We first need to consider more controlled crowdsourced experiments, like CIFAR-10H (Peterson et al., 2019) (relying on Amazon's Mechanical Turk[1]): the most complete open-access dataset with a control over the number of tasks annotated by each worker. Note that the images were very curated (Aitchison, 2020), thus most reaching consensus among the workers. Yet, obtaining such data is long, raises ethical or possibly legal issues, and a relevant toy simulation framework is missing.

We introduce a toy-dataset simulator that provides visually corrupted tasks and confusion-aware workers with control over the confusion using the task difficulty. Most simulations use the class-conditional assumption (Patrini et al., 2017), *i.e.,* one only observes the confused version y_{conf} of y where $\mathbb{P}_{y_{conf}|x} = \pi^\top \mathbb{P}_{y|x}$ for $\pi \in \mathbb{R}^{K\times K}$ a confusion matrix with $\pi_{k,k'} = \mathbb{P}(y_{conf} = k'|y = k)$ the probability that the true label k is confused with k'. This assumption overlooks the essential factor that is the instance-dependence, as labelling a clear image is easier than a blurry one. Whitehill et al. (2009) created difficulty-aware labelling simulations, without generating visualizable tasks, with either hard or easy instances and good or bad workers for a prescribed confusion matrix.

In our framework, confusion is reflected through the colors' variability in the output images: pure colors are easy to label while blended ones are more difficult, see Figure 2. Each worker provides an answer depending on their abilities to distinguish colors. The framework allows testing different aggregations and workers evaluations strategies while knowing the ground truth. The architecture also authorizes poor labelling such as noisy workers (answering tasks at random) that we identify before label aggregation.



(a) (b)

Figure 1: a) RGB graph; b) neighborhood example for $y = (0, 0, 0)$: $\mathcal{N}_y = \{(0, 0, 1), (0, 1, 0), (1, 0, 0)\}$.

2 Crowdsourced dataset

2.1 Simulated crowdsourced data

To provide visual intuition for our simulated crowdsourced dataset, we rely on the RGB cube (Qin et al., 2019). We denote by G the unit cube graph with vertices $\mathcal{Y} = \{0, 1\}^3$, with each vertex represent an RGB color as in Figure 1. The neighborhood of $y \in \mathcal{Y}$ is defined as $\mathcal{N}_y := \{y' \in \mathcal{Y},\ \sum_{i=1}^{3} \mathbb{1}_{\{y_i \neq y'_i\}} = 1\}$ and the edges are $\{(y, y') \in \mathcal{Y} \times \mathcal{N}_y\}$. Let

[1] http://theplantgame.com/, https://www.mturk.com/

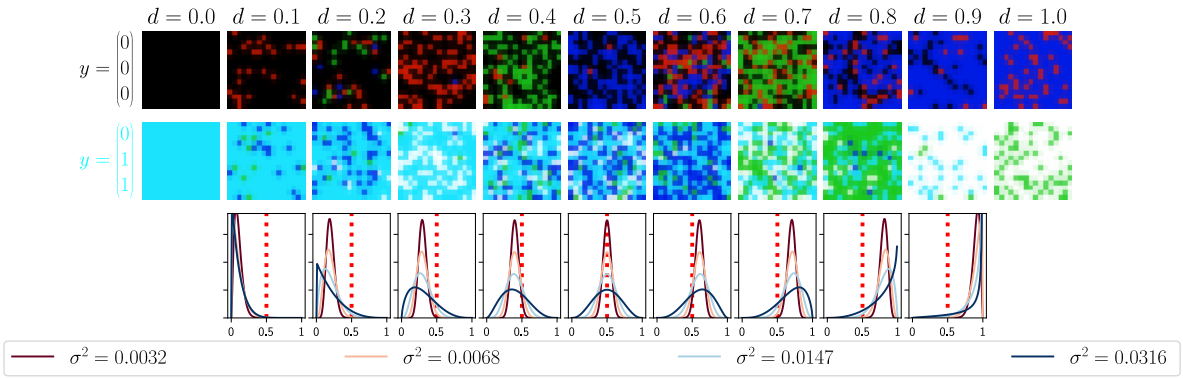


Figure 2: Outputs from Algorithm 1 with gradually increasing levels of difficulty d for true labels $(0, 0, 0)$ and $(0, 1, 1)$. The closer d is to 0, the purer the image and the simpler the labelling. The third row represents Beta confusion distributions (defined in Equation (1)) and used in Algorithm 2 for four levels of workers' abilities σ^2. The decision limit for workers between choosing y or a label in $\mathcal{N}_y$ is at 0.5.

$\Delta^n \subset \mathbb{R}^{n+1}$ be the simplex of dimension n, $\mathbf{1}_n \in \mathbb{R}^n$ the vector of ones and $A \otimes B$ the Kronecker product between two matrices A and B.

We generate n_t tasks following Algorithm 1, that provides an image x, a ground truth label y and a difficulty level d as follows. Each task/image has pixels whose color is a combination of the true color y and alternative ones from $\mathcal{N}_y$. For a difficulty level $d \in [0, 1]$, if a sample drawn from a Bernoulli distribution $\mathcal{B}er(d)$ is 0, then the pixel value remains y. Otherwise, a Dirichlet distribution with parameter $(1/3, 1/3, 1/3)$ provides a sampled distribution over the neighborhood $\mathcal{N}_y$.

Algorithm 1: Generate one RGB image with visual label corruption.

Data: $d \in [0, 1]$: task difficulty level; y: task label; $W \in \mathbb{N}^*$: width of image.
Result: x: image; $\mathcal{N}_y$: neighborhood; $\nu_{\cdot|y}$: confusion distribution.

$x \leftarrow \mathbb{1}_{W^2} \otimes y \in [0, 1]^{W^2 \times 3}$ `// Vectorized image of constant` y
$\nu_{\cdot|y} \in \Delta^2 \sim \mathcal{D}ir\,(1/3, 1/3, 1/3)$
for $k = 1, \ldots, W^2$ **do** `// For each pixel`
  $switch \sim \mathcal{B}er(d)$
  if $switch = 1$ **then**
    $(q_{y'})_{y' \in \mathcal{N}_y} \sim \mathcal{D}ir\,(\nu_{\cdot|y})$ `// Corruption distribution`
    $x[k] \leftarrow \arg\max_{y' \in \mathcal{N}_y}(q_{y'})$

The pixel color is then chosen as the most likely one in $\mathcal{N}_y$ with respect to the former distribution. To sample balanced difficulties and control the maximum level, a candidate difficulty distribution is $\mathcal{U}(0, d_{\max})$.

The workers' answers are implemented using Algorithm 2. For each image x_i, $i \in [n_t]$ to label, the worker must decide whether the true label y_i is the dominant color in x_i. To model this decision, a random variable with support $[0, 1]$ (the confusion-difficulty

Algorithm 2: Simulate one worker answer for a fixed task

Data: y: label; d: difficulty; $\nu_{\cdot|y}$: confusion distribution; $\{(\sigma^{(j)}_{y\leftrightarrow y'})^2\}_{y,y'}$ worker abilities to distinguish y from $y' \in \mathcal{N}_y$.

Result: Worker answer $y^{(j)}$

$confu^{(j)} \sim \mathcal{B}eta\left(\mu = d, \sigma^2 = \max_{y\in\mathcal{N}_{y'}} \left(\sigma^{(j)}_{y\leftrightarrow y'}\right)^2\right)$ // Equation (1)

if $confu^{(j)} < \frac{1}{2}$ **then** // Worker answers true color

  $y^{(j)} = y$

else // The worker is wrong

  $(alt_{y'})_{y'\in\mathcal{N}_y} \sim \left(\mathcal{B}eta\left(\mu = (1-d)\nu_{y'|y}, \sigma^2 = \max_{\ell\in\mathcal{N}_y\setminus\{y'\}} \left(\sigma^{(j)}_{y'\leftrightarrow\ell}\right)^2\right)\right)_{y'\in\mathcal{N}_y}$

  $y^{(j)} = \arg\max_{y'\in\mathcal{N}_y} alt_{y'}$ // Choose worker's most present color in $\mathcal{N}_y$

range) reflects the prior expectations on the worker, *i.e.,* their abilities to distinguish the different colors (third row of Figure 2). We have considered Beta distributions with mean $\mu = d_i$ and variance σ^2 depending on their abilities. If their confusion (the draw from the Beta distribution) is above ½ (Figure 3) then a majority of pixels are considered from $\mathcal{N}_{y_i}$. Then we model the worker's decision among the colors from $\mathcal{N}_{y_i}$ with another Beta distribution, the only changes are the mean, now at $(1-d_i)\nu_{y'|y_i}$ to reflect the proportion of each alternative color, and the variance that no longer takes into account y_i but only its possible confusions.

The parametrization of the Beta distribution is done as follows. Let $\alpha, \beta > 0$, then for $\sigma^2 < \mu(1-\mu)$:

$$X \sim \mathcal{B}eta(\mu, \sigma^2) \quad (1)$$
$$\Leftrightarrow X \sim \mathcal{B}eta\left(\alpha = \mu\left(\tfrac{\mu(1-\mu)}{\sigma^2} - 1\right), \beta = \tfrac{1-\mu}{\mu}\alpha\right).$$

In practice, given parameters (μ, σ^2), we use $\left(\mu, \min(\sigma^2, \sigma^2_{\max} - \epsilon)\right)$ with $\epsilon > 0$ small and $\sigma^2_{\max}$ *s.t.* $\alpha > 1$ to avoid U-shaped or degenerated distributions that would not reflect a confusion behavior.

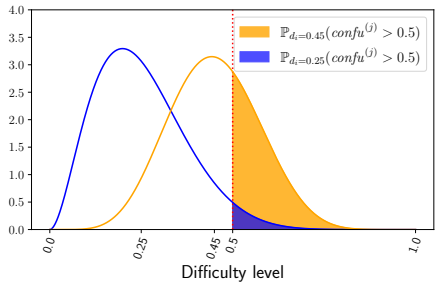
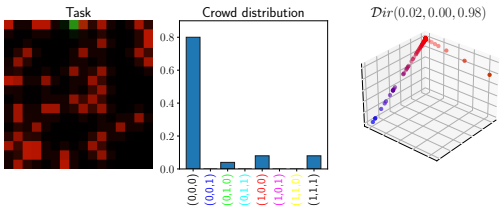


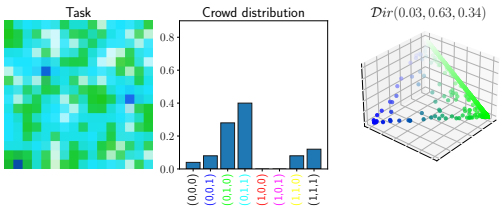
Figure 3: Probability of confusion, *i.e.,* to answer a label in $\mathcal{N}_{y_i}$ depending on the difficulty of the task ($d_i = 0.25$ and 0.45) for a fixed confusion level $\sigma^2 = 0.015$.

2.2 From crowd to learnable labels

To learn human decisions from crowd labels we use soft labels (*i.e.,* the distribution of votes) known to provide better calibrated architectures (Guo et al., 2017). Besides, we have access to the most confusing color: $\arg\max_{y'\in\mathcal{N}_y} \nu_{y'|y}$ that we call the second-best choice. In replicating a human prediction, we thus consider the Second-Best Accuracy (SBA), that is "the second label predicted is the second-best



(a) $y_i = (0, 0, 0)$, $d_i = 0.2$, $\nu_{\cdot|y_i}$ is concentrated near $(1, 0, 0)$, not $(0, 1, 0)$ nor $(0, 0, 1)$. Except spammers, there were no votes in $\mathcal{N}_{y_i}$.



(b) $y_i = (0, 1, 1)$, $d_i = 0.45$, $\nu_{\cdot|y_i}$ is concentrated near $(0, 1, 0)$ mostly. We see the crowd hesitated with the second-best choice $(0, 1, 0)$.

Figure 4: Labelling for $n_w = 25$ workers: 6 spammers and 19 workers from Algorithm 2. We display 1000 draws from $\nu_{\cdot|y_i}$ the confusion distribution.

choice". This evaluates the model capacity to take into account the confusion in its prediction the way humans would when presented a task.

2.3 Identifying spammers in crowd

Here, we call spammer a worker that answers any label randomly, without looking at the task given i.e., $\mathbb{P}(y_i^{(j)} = c | y_i = c') = \mathbb{P}(y_i^{(j)} = c)$ for $c, c' \in [K]$. In our simulations we use $\mathbb{P}(y_i^{(j)} = c) = K^{-1}$. A strategy proposed by Raykar et al. (2012) to detect spammers is to compute a *spam score* $s^{(j)} \in [0, 1]$ from the confusion matrix $\pi^{(j)}$ of each worker. Using $\hat{u}_j = \arg\min_{u_j} \left\| \pi^{(j)} - \mathbf{1}_K u_j^\top \right\|_F^2 \; s.t \; u_j^\top \mathbf{1}_K = 1$:

$$s^{(j)} = \left\| \pi^{(j)} - \mathbf{1}_K \hat{u}_j^\top \right\|_F^2 = \frac{1}{K(K-1)} \sum_{c<c'} \sum_{k \in [K]} \left(\pi_{ck}^{(j)} - \pi_{c'k}^{(j)} \right)^2 \quad . \tag{2}$$

This is simply the sum of the unbiased variance of each column of $\pi^{(j)}$. The closer $s^{(j)}$ is to 0, the more likely worker j is a spammer. However, $\pi^{(j)}$ in Equation (2) can not be computed without knowing y_i. The DS model, proposed by Dawid et al. (1979), estimates the confusion matrices without access to y_i by maximizing the likelihood with unknowns $\pi^{(j)}$'s, indicators and prevalence $\rho \in \Delta^7$: $\prod_{i \in [n_t]} \prod_{k \in [K]} \left\{ \rho_k \prod_{j \in [n_w]} \prod_{\ell \in [K]} (\pi_{k\ell}^{(j)}) \right\}^{\mathbb{1}_{\{y_i = k\}}}$.

3 Experiments

We compare the impact of the workers abilities in the labelling aggregation using two settings with a different number of workers. First, we estimate $\hat{\pi}^{(j)}$ with the DS model. Then, from Equation (2), we get $\hat{s}^{(j)}$ the worker's estimated spam score, and train a k-means ($k = 2$) on the scores to separate the spammers (simulated from Section 2.3) from workers (simulated with Algorithm 2). From Table 1, when only a few workers can be trusted, removing spammers drastically improves the learning procedure.

Table 1: Top-1, Top-2 and SBA accuracies with logistic regression on simulated crowd data taking $n_t = 500$, $W = 64$ and $d_{\max} = 0.6$ using soft labels with and without spam removal. Images are smoothed using a Gaussian filter with standard deviation 0.5. The levels of workers abilities are those from Figure 2. Test size is 30%.

	$n_w = 25$ with 17 spammers			$n_w = 100$ with 88 spammers		
	Top-1	Top-2	SBA	Top-1	Top-2	SBA
soft labels	0.35	0.53	0.19	0.19	0.39	0.17
remove spams + soft labels	**0.71**	**0.88**	**0.38**	**0.80**	**0.92**	**0.42**

Figure 5: CIFAR-10H dataset: example of tasks labelled by workers 1098, 2160 and 2156 identified as spammers. Answered labels do not match the obvious true label.

Running the spammer detection on the CIFAR-10H dataset, we obtain that there are 19 spammers (out of 2571) that could be removed from the crowd (*e.g.,* Figure 5).
Acknowledgment: Work supported by the Chaire CaMeLOt ANR-20-CHIA-0001-01.

References

Dawid, AP and AM Skene (1979). Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm. *J. R. Stat. Soc. Ser. C. Appl. Stat.* 28.1, pp. 20–28.
Whitehill, J et al. (2009). Whose Vote Should Count More: Optimal Integration of Labels from Labelers of Unknown Expertise. *NIPS*. Vol. 22. Curran Associates, Inc.
Raykar, VC and S Yu (2012). Eliminating Spammers and Ranking Annotators for Crowdsourced Labeling Tasks. *J. Mach. Learn. Res.* 13, pp. 491–518.
Rodrigues, F, F Pereira, and B Ribeiro (2013). Learning from multiple annotators: distinguishing good from random labelers. *Pattern Recognition Letters* 34.12, pp. 1428–1436.
Guo, C et al. (2017). On calibration of modern neural networks. *ICML*. PMLR, p. 1321.
Patrini, G et al. (2017). Making deep neural networks robust to label noise: A loss correction approach. *CVPR*, pp. 1944–1952.
Khetan, A, ZC Lipton, and A Anandkumar (2018). Learning From Noisy Singly-labeled Data. *ICLR*.
Peterson, JC, RM Battleday, and TL Griffithsand O Russakovsky (2019). Human Uncertainty Makes Classification More Robust. ICCV, pp. 9617–9626.
Qin, Y et al. (2019). A Multi-class Classification Algorithm Based on Hypercube. 2019 IEEE DDCLS, pp. 406–409.
Aitchison, L (2020). A statistical theory of cold posteriors in deep neural networks. *ICLR*.

Evaluating the Impact of Parenthood on Career Progression in STEM jobs, across Gender

Mira Rahal[1], Adeline Samson[2], Elise Arnaud[3], Isabel Torres[4]

[1] Laboratoire d'École d'Économie de Paris, *mira.rahal@psemail.eu*

[2] Laboratoire Jean Kuntzman, *Université Grenobles Alpes*, *adeline.leclercq-samson@univ-grenoble-alpes.fr*

[3] Laboratoire Jean Kuntzman, *Université Grenobles Alpes*, *elise.arnaud@univ-grenoble-alpes.fr*

[4] Mothers in Science, *isabel.torres@mothersinscience.com*

Résumé. Nous avons étudié l'impact de la parentalité sur les carrières scientifiques de Science, Technology, Engineering, Mathematics et Médecine (STEMM) à partir des réponses à une enquête réalisée par Mothers in Science en France en 2020. Dans les secteurs académiques, nous avons trouvé que les femmes connaissent souvent une baisse de taux de publications après être devenues parents, ce qui est souvent moins le cas pour les pères. En utilisant un modèle à effets mixtes, nous avons montré que les femmes travaillant dans le secteur des sciences physiques connaissent une baisse significative de leur productivité après être devenues parents. Nous avons également mis en évidence un effet inégal de la parentalité sur la réputation professionnelle selon le sexe, à travers un modèle de régression logistique ordinale.

Mots-clés. Parentalité, inégalités de genre, discrimination, carrières scientifiques

Abstract. We investigated the impact of parenthood on career progression in the Science, Technology, Engineering, Mathematics, and Medicine (STEMM) sectors based on responses to a survey conducted by Mothers in Science in France in 2020. We found that in academic sectors, mothers publish at lower rates than fathers, and this effect persists throughout their careers. Using a Mixed effects model, we showed that women working in Physical Sciences experience a significant reduction in their scientific productivity after becoming parents. Finally, we also showed an unequal effect of parenthood on professional reputation according to gender, through an ordinal logistic regression model.

Keywords. Parenthood, gender inequality, discrimination in the workplace, academia

1. Background

During their working-age lives, a high proportion of individuals experience parenthood by either adopting or having a child. This important event adds more responsibilities to workers. Parenthood was demonstrated to be a crucial turning point in the career trajectories of men and women, especially those engaged in the fields of STEM (*Science, Technology, Engineering,* and *Mathematics*). Various studies have related this aspect of parenthood to the high demanding nature of many STEM jobs (Fox, 2010, Fox et.al, 2011, Cech & Blair-Loy, 2017).

A recent US-based study focused on assessing the gap in scientific productivity in academia across genders, both prior to experiencing parenthood and following it (Morgan et. al., 2021). This study demonstrated a "gendered effect of parenthood" on productivity, measured by parents' rates

of publications. Before experiencing parenthood, the annual productivity rates of men and women were established to be similar. However, after becoming parents, mothers were found to produce a significantly lower number of papers than fathers. In addition, statistical results suggested that the annual productivity of mothers decreased significantly after childbirth in comparison to female non-mothers. In contrast, for the period preceding children, faculty from both gender who went on becoming parents were found to be slightly more productive than those who did not.

An international survey led by *Mothers in Science* in partnership with other organizations further investigated the challenges faced by parents in STEM careers globally. The current study, conducted by *Mothers in Sciences* and *Femmes et* Sciences, analyses the data set collected in France, with the aiming of measuring the impact of parenthood on the career progression of students and professionals in STEMM in France. The study tests two hypotheses:

(H1) Productivity: *Work productivity in mothers decreases after having or adopting a child, but fathers' productivity remains unchanged and is similar to that of childless people*

(H2) Maternal Wall*: Mothers, but not fathers, face bias and discrimination in the workplace*

1. Data and Methods

1.1. *Data*

The survey under study, entitled "Parenthood and career progression in STEMM" was distributed globally in early 2020 and was targeted for people of any gender, with or without children, who are working or studying in any professional sector in STEMM. It included a variety of questions ranging from general (*Age, sex, profession, sector of activity, number of children if any*, …), to specifics that identify career characteristics, personal responsibilities, and work-life balance. Mainly, respondents working in the "Academic" sectors were further asked to estimate the number of publications for each time phase relative to their Highest Degree (HD) completion. That is, they provided the number of peer- reviewed scientific publications produced during Undergrad, Masters, PhD, the 3 years following HD, and for each later 3-year period[1]. Moreover, parents were asked to answer a set of questions aimed at identifying obstacles faced in their careers (if any) after experiencing parenthood, or upon their return from parental leave.

The French sample included 1800 responses (1308 females vs 492 male responses) and researchers who were either PhD candidates (184, 11%) or had at least completed a PhD degree (933, 52%), working in the sectors of Physical Sciences (585, 32%), Math and Computer Sciences (492, 27%), Life Sciences (373, 20%) and Social Sciences (94, 5%). As expected, since parents were more concerned with the topic of the survey, 72% of respondents were parents including 980 mothers and 328 fathers.

1.2. *Methods*

1.2.1. *Assessing trends in Scientific productivity* (**H1**)

The effect of parenthood on the scientific productivity of mothers and fathers was measured as their publication rates with respect to the childbirth timing. We thus mapped for each respondent

[1] The time modalities available where: Undergrad, Masters, PhD, 3 years following highest degree completion, [4-6], [7-9], [10-12], or [13-15] years after highest degree completion

the phases related to Highest Degree completion with phases connected to the event of parenthood. For a respondent having his first child "*During PhD"*, for instance, the phase "*[0-3] years after HD"* was mapped to a phase called "t_1" which translates to "1 period after parenthood" (given that this person's highest completed degree is PhD). Consequently, if this respondent reported 3 publications during his Masters and 6 publications between "*4-6 years after HD completion"*, the numbers assigned for his "t_{-1}" and "t_2" were respectively 3 and 6.

As one of the goals of the study was to compare the rates of productivity across parents and childless respondents, we have assigned "counterfactual" birth events to the non-parents. For this, we used "Age" and "Sector of Activity" to assign the counterfactual **period** of experiencing parenthood. We analyzed the age distribution among gender for each sector, to be able to match non-parents with parents having the same characteristics (same age, gender and sector). We assigned to non-parents the childbirth timing corresponding to the majority of parents having those characteristics. For example, for a childless woman who is 31 and working in the sector of Physical Sciences, the counterfactual period of parenthood that she was assigned corresponded to the period during which **most** of 31-year-old mothers, working in the same sector, have reported having their first child ("*During the first 3 years following HD*" in this case). However, the described strategy was not applicable to the cases when there was no specific majority in the possible options to assign. To circumvent this problem, we proceeded as follows: when we could not apply a "majority rule", we have assigned the childless respondents with the counterfactual birth period that corresponded to the "Latest"[2] period during which parents of similar characteristics had their child. That is, having two 35-year-old fathers with one experiencing parenthood "*During PhD*" and the other experiencing it "*[4-6] years after HD*" has led us to assign 35-year-old childless males, working in the same academic sector, with the latter period as counterfactual birth time.

To test hypothesis (**H1**), a Poisson Generalized Linear Mixed Effects Model (PGLMM) was applied on the number of publications reported by each respondent, denoted *nbofpub*, to account for global fixed effects and inter-individual heterogeneity. The effect of covariates is thus estimated using the following model (1):

$$\begin{aligned} log\big(E\big[nbofpub_{i,t}\big]\big|X_i\big) = {} & \big(\beta_0 + \delta_0 Gender_i + \gamma_0 Parent_i + \alpha_{i,0}\big) \\ & + t * (\beta_1 + \delta_1 Gender_i + \gamma_1 Parent_i) \\ & + I[Parenthood]_{i,t} * \big(\beta_2 + \delta_2 Gender_i + \alpha_{i,1}\big) \quad (1) \end{aligned}$$

where: $Gender_i$and $Parent_i$ are dummy variables equal to 1 when respondent *i* is a female or a parent respectively ; t is a numeric variable assigned to the period of parenthood ranging from -3 to 3 ; $I[Parenthood]_{i,t}$ is an indicator function equals to 1 when $observation_{i,t}$ belongs to a **parent** respondent *i* at time $t > 0$; $\alpha_{i,0}$ accounts for a random effect on the intercept ; $\alpha_{i,1}$ accounts for a random effect on parenthood timing; X_i is the vector including the chosen variables associated with respondent *i*.

Model (1) was run on each sector separately, with the exclusion of the subset related to Social Sciences due to the small number of responses.

[2] We have also tried a method of assigning the "Earliest" periods as counterfactual periods of experiencing parenthood yet we had quiet similar results.

1.2.2. *Evaluating differences in link between parental leave and professional reputation* (**H2**)

In the aim of testing whether mothers and fathers faced significantly different career obstacles upon becoming parents, we explored the reactions to the statements: "*The time I stayed away from work for parental leave (with my first child) affected my (a) Roles taken at work, (b) Application Rates of publications/grants, (c) Professional Reputation, (d) Field work*". However, we only performed an advanced statistical analysis on the statement relating "Parental leave" with "Professional Reputation" as it was the best proxy of discrimination in the workplace.

As the answer was ordered categorically (from "*Strongly disagree*", *to* "*Strongly agree*"), we fitted an Ordered Logistic Model with the rating of the statement *(c)* as our main dependent variable, and *Gender* as a main regressor. As covariates, the model accounted for age, sector of activity, length of parental leave (short or long), overall number of publications, sector of activity, number of children (one, two, 3 or more), and an interaction term between number of children and gender.

2. Results

2.1. *Mothers working in the sector of Physical Sciences experience a decrease in their rate of productivity after parenthood*

Results of (1) are summarized in *Figure 1* which displays the estimated risk ratio along with their confidence intervals, for the different fitted models.

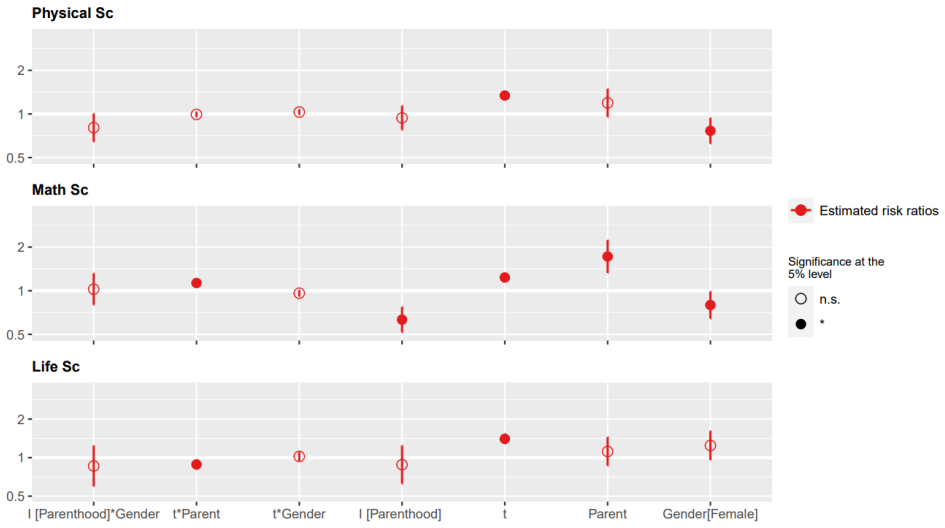


Figure 1: Estimated risk ratios of model (1) fitted on the data of respondents working in Physical Sciences, Math Sciences, and Life Sciences.

Model (1) estimated a significant effect of gender on scientific productivity, for the sectors of Physical and Math Sciences. Mainly, among respondents working in these sectors, women were found to be more likely to have less publications than men, regardless the status and time of

parenthood. In regards to the estimates of the coefficient investigating the "gendered effect" of parenthood $\delta_2\big(I[Parenthood]_{i,t} * Gender_i\big)$, *Figure 1* indicates that it was not significant in any case. However, for the Physical Sciences data set, the p-value was 0.055. Then, we studied the estimated trends of publications with respect to the timing of parenthood for mothers and fathers and their childless counterparts. Displayed in *Figure 2*, our estimates predicted that after t_0, mothers working in the academic sector of Physical Sciences report the lowest productivity with respect to the three other groups.

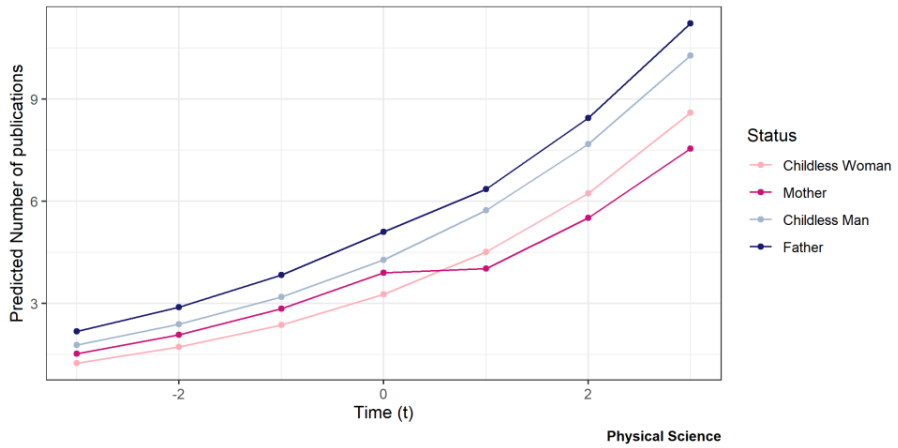


Figure 2: Estimated trends in rates of publications from (1) with respect to the timing of parenthood (real or counterfactual) for Childless Women, Mothers, Childless Men and Fathers.

2.2. *Mothers were more likely to agree that Parental leave had negatively affected their Professional Reputation*

The results of our fitted Ordinal Logistic Model described in Section **1.2.2.** suggested that among females, the odds of agreeing more with the statement that "*Parenthood has had an effect on their professional reputation*" is 3.64[3] than this odds among father respondents (with all other variables being fixed). The described result provided evidence that mothers were more likely to face bias and discrimination in the workplace than fathers.

Bibliography

- Blair-Loy M, Cech EA (2017) Demands and devotion: Cultural meanings of work and overload among women researchers and professionals in science and technology industries. *Social Forces* 32:5–27

- Breslow, N. E., & Clayton, D. G. (1993). Approximate Inference in Generalized Linear Mixed Models. *Journal of the American Statistical Association*, *88*(421), 9–25.

[3] The estimated Odd Ratio associated with Gender was 3.64 with the [1.65-8.34] confidence interval, and a p-value=0.002

- Dean CB, Nielsen JD. (2007) Generalized linear mixed models: a review and some extensions. Lifetime Data Anal. Dec;13(4):497-512.

- Fox MF (2010) Women and men faculty in academic science and engineering: Social organizational indicators and implications. *American Behavioral Scientist* 53:997–101

- Fox MF, Fonseca C, Bao J (2011) Work and family conflict in academic science: Patterns and predictors among women and men in research universities. *Social Studies of Science* 41:715–735

- Harell, F.E (2015). Ordinal Logistic Regression. In: *Regression Modeling Strategies: With applications to linear models, logistic and ordinal regression, and survival analysis*. Springer Series in Statistics

- Morgan A. C., Way S. F., Hoefer M. J. D., Larremore D. B., Galesic M., Clauset A., (2021) The unequal impact of parenthood in academia. *Sci. Adv. 7*, eabd1996

Profondeur à noyau multivariée et mesure de risque extrême associée

Sara Armaut[1], Roland Diel[2] & Thomas Laloë[3]

[1] *Université de Nice Côte d'Azur, armaut@unice.fr*
[2] *Université de Nice Côte d'Azur, Roland.Diel@unice.fr*
[3] *Université de Nice Côte d'Azur, Thomas.Laloe@unice.fr*

Résumé. De nos jours, dans pratiquement tous les domaines tels que la finance, la médecine, l'écologie, l'industrie..., il est impossible d'éviter des risques ! Par exemple, en finance, le "risque" signifie souvent la possibilité de perdre de l'argent. En hydrologie, le risque peut par ailleurs représenter la quantité d'eau dépassant le niveau de remplissage maximum d'un barrage. La *Conditional Covariate Tail Expectation* (ou CCTE) est une mesure de risque qui quantifie un coût moyen associé à $d \geq 1$ facteurs de risque non nécessairement homogènes. Dans notre cadre d'étude, la zone de risque est représentée par un ensemble de niveau inférieur associé à une fonction de profondeur statistique multivariée. Nous proposons un estimateur consistant de la CCTE pour des niveaux extrêmes : cet estimateur fait intervenir une estimation de l'ensemble de niveau associé à la profondeur en question via une méthode *plug-in*. Dans le modèle gaussien, on obtient une vitesse de convergence de la CCTE basée sur la profondeur dite à noyau.

Mots-clés. Estimation *plug-in*, fonction de profondeur multivariée, profondeur à noyau, théorie du risque, modèle gaussien.

Abstract. Nowadays, in almost all fields such as finance, medicine, ecology, industry.., it is not possible to avoid risks. For instance, in finance, risk often means that there is potential for money loss. In hydrology, risk could represent the amount of water which exceeds the maximum storage level of a dam. The *Conditional Covariate Tail Expectation* (CCTE) is a risk measure that quantifies an expected cost associated to $d \geq 1$ risk factors which are heterogeneous in nature. In our setting, the risk region in the problem at hand is represented by a depth-based lower level set. We provide a consistent estimator of the CCTE for extreme levels : this estimator involves a *plug-in* approach when estimating the lower level set. Under the gaussian model, we derive a rate of convergence for the CCTE based on the so-called kernel depth.

Key-Words. *Plug-in* estimation, multivariate depth function, kernel depth, risk theory, gaussian model.

1 Introduction

La théorie du risque est une branche de la statistique qui s'intéresse aux évènements peu probables. Le but principal recherché est de pouvoir gérer la part d'incertude de certains évènements, ce qui permet la mise en place de mesures de prévention. La théorie du risque est appliquée dans divers domaines, par

exemple, en hydrologie pour prévoir les crues, en finance pour protéger la valeur du portefeuille après un investissement, etc...
Usuellement, une mesure de risque est une application définie sur un ensemble de variables aléatoires à valeurs réelles. Il est à noter que, pour un phénomène, la prise en compte d'un seul facteur de risque pourrait considérablement affecter l'exactitude du modèle étudié. Modéliser la structure de dépendance de données multivariées permet alors d'obtenir des résultats significatifs et réprésentatifs dans l'analyse du risque. Par exemple en hydrologie, plusieurs phénomènes sont décrits par le biais de deux ou plusieurs variables corrélées. Ces dernières sont considérées jointement afin de représenter efficacement ces phénomènes hydrologiques. Ainsi, la probabilité de réalisation d'un risque ne peut pas être estimée sur la base de l'analyse univariée. La littérature de risques hydrologiques a été largement étudiée dans un cadre multivarié et traite principalement un ou plusieurs des éléments suivants : (1) montrer l'importance et l'utilité du cadre multivarié, (2) trouver un modèle de distributions multivariées approprié afin de modéliser les risques, et (3) définir et étudier des temps de retour miltivariés (cf. Chebana and Ouarda (2011)). L'étude de risques amène à l'étude de "régions" quantiles. Celle des risques univariés via des quantiles univariés a été largement traitée dans la littérature. Quant aux risques multivariés, l'étude de quantiles multivariés a gagné beaucoup d'attention ces dernières décénies, notamment les quantiles basés sur une loi de probabilité multivariée (cf. Belzunce et al. (2007), Dehaan and Huang (1995), Cousin and Di Bernardino (2013)), ou encore les quantiles basés sur une fonction de profondeur (cf. Zuo and Serfling (2000)).

Il est pertinent d'analyser le comportement d'un coût modélisé par une variable aléatoire réelle Y (par exemple, la somme d'argent gagnée ou perdue dans un investissement sur une certaine période), et ce par rapport à $d \geq 1$ facteurs de risque différents $\boldsymbol{X} \in \mathbb{R}^d$. Intuitivement, la variable de coût Y va dépendre des facteurs de risques du phénomène étudié. Dans un cadre d'étude général, Laloë et al. (2015) proposent d'analyser le comportement d'une variable aléatoire réelle Y qui dépend d'un vecteur aléatoire de risques $\boldsymbol{X} \in \mathbb{R}^d$. Plus précisément, la *Covariate Conditionnal Tail Expectation* suivante définie une mesure de risque :

$$\mathrm{CCTE}_{F,\alpha}(Y, \boldsymbol{X}) := \mathbb{E}[Y | \boldsymbol{X} \in \mathcal{L}_{F_{\boldsymbol{X}}}(\alpha)], \alpha \in (0,1), \tag{1.1}$$

où,

$$\mathcal{L}_{F_{\boldsymbol{X}}}(\alpha) := \left\{ x \in \mathbb{R}^d : F_{\boldsymbol{X}}(x) \geq \alpha \right\}, \tag{1.2}$$

est l'ensemble de niveau (supérieur) associé à la fonction de répartion $F_{\boldsymbol{X}} := F$ du v.a $\boldsymbol{X} \in \mathbb{R}^d$.

2 *Depth-based Covariate Conditionnal Tail Expectation*

L'utilisation d'une mesure de risque basée sur les ensembles de niveau d'une fonction de répartition nous restreint à considérer uniquement des orientations particulières du risque (cf. Figure 1, graphe gauche). Par exemple, on peut considérer des températures très basses ou très élévées, mais pas les deux à la fois. Afin de contourner ce problème de dépendance en une direction canonique, au lieu de considérer des ensembles de niveau de la forme $\mathcal{L}_{F_{\boldsymbol{X}}}(\alpha)$, des ensembles de

niveau de la forme $\mathcal{L}_{F_{R\boldsymbol{X}}}(\alpha)$ ont été étudiés dans la littérature, où R est une matrice de rotation dans $\mathbb{R}^d$ (cf. Torres et al. (2020)). Une telle matrice de rotation R permettrait de mettre en évidence d'autres zones de risque dans $\mathbb{R}^d$; cependant, dans ce cas, la CCTE dépendrait encore d'une orientation via la rotation R. Pour cela, afin d'étudier une mesure de risque complètement indépendante de l'orientation, nous proposons une approche nouvelle en considérant la CCTE pour des ensembles de niveau associés à une fonction de profondeur statistique multivariée (au sens de la Définiton 2.1 dans Zuo and Serfling (2000)), et qui est notée

$$D : \mathbb{R}^d \times \mathscr{P}(\mathbb{R}^d) \to \mathbb{R}_+,$$

avec $\mathscr{P} := \mathscr{P}(\mathbb{R}^d)$ désignant l'ensemble des mesures de probabilités sur $\mathbb{R}^d$. En efffet, une fonction de profondeur ordonne des données selon leur *degré de centralité*, et fournit un ordre statistique du *centre* vers l'extérieur (cf. Zuo and Serfling (2000)).

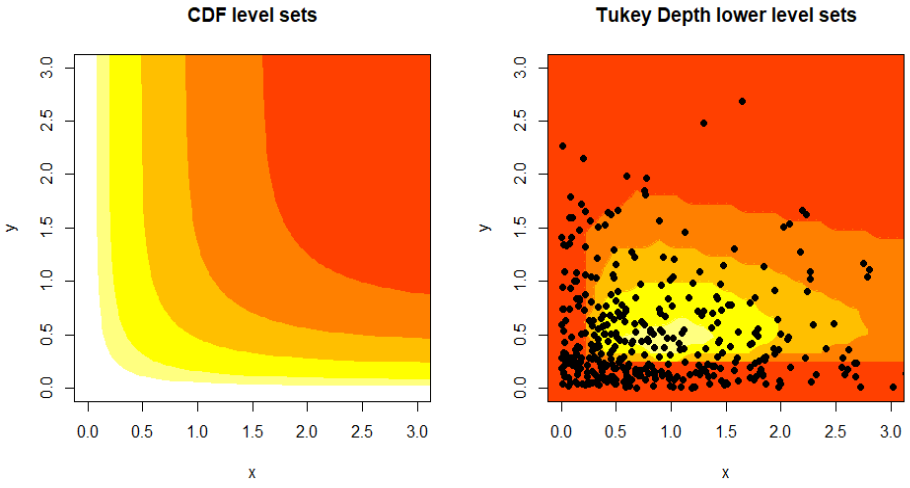


FIGURE 1 – Ensembles de niveaux (supérieurs) pour une fonction de répartition et ensembles de niveaux inférieurs pour une profondeur respectivement (de gauche à droite respectivement).

Afin d'étudier les zones de risque via une fonction de profondeur, il faut s'intéresser aux régions de faible profondeur sur lesquelles nous pouvont étudier le comportement de la covariable de coût Y. En d'autres termes, considérant un niveau $\alpha > 0$, une mesure de probabilité P sur $\mathbb{R}^d$, nous considérons l'ensemble de niveau

$$\boldsymbol{\mathcal{L}_D(\alpha) := \left\{ x \in \mathbb{R}^d : D(x, P) \leq \alpha \right\}.} \tag{2.1}$$

Étant donné une suite d'estimateurs consistants $(\tilde{P}_n)_{n\geq 1}$ de P, nous proposons d'estimer $\boldsymbol{\mathcal{L}_D(\alpha)}$ par l'ensemble

$$\boldsymbol{\mathcal{L}_{n,D}(\alpha) := \mathcal{L}_n(\alpha) = \left\{ x \in \mathbb{R}^d : D_n(x) := D(x, \tilde{P}_n) \leq \alpha \right\}, n \geq 1.} \tag{2.2}$$

3 Estimation et convergence à des niveaux extrêmes

Soit $\boldsymbol{X}$ un vecteur aléatoire de loi $P \in \mathscr{P}(\mathbb{R}^d)$, Y une va réelle, et $\alpha > 0$. Dans notre cadre d'étude, on s'interesse à l'estimation de la CCTE basée sur les ensembles de niveau d'une profondeur, définie par :

$$\mathrm{CCTE}_{D,\alpha}(Y, \boldsymbol{X}) := \mathbb{E}[Y | \boldsymbol{X} \in \mathcal{L}_D(\alpha)]. \tag{3.1}$$

Soient $n \geq 1$. Soient deux échantillons,

$$\begin{aligned} \tilde{S}_n &:= (\tilde{\boldsymbol{X}}_i)_{i=1,..,n} \text{ de même loi que } \boldsymbol{X}, \text{ et} \\ S_n &:= ((Y_i, \boldsymbol{X}_i))_{i=1,..,n} \text{ de même loi que } (Y, \boldsymbol{X}), \end{aligned} \tag{3.2}$$

tels que les $(Y_i, \boldsymbol{X}_i)_i$ et les $(\tilde{\boldsymbol{X}}_i)_i$ soient indépendants. Considérons $\mathcal{L}_n(\alpha)$ un estimateur (calculé à partir de $\tilde{S}_n$) de $\mathcal{L}_D(\alpha)$. On peut alors construire à partir de l'échantillon S_n un estimateur pour la CCTE (3.1) :

$$\widehat{\mathrm{CCTE}}^n_{D,\alpha}(Y, \boldsymbol{X}) := \mathbb{E}_{S_n}[Y | \boldsymbol{X} \in \mathcal{L}_n(\alpha)] = \frac{\sum_{i=1}^n Y_i \mathbb{1}_{\boldsymbol{X}_i \in \mathcal{L}_n(\alpha)}}{\sum_{i=1}^n \mathbb{1}_{\boldsymbol{X}_i \in \mathcal{L}_n(\alpha)}}, \tag{3.3}$$

avec la convention $0/0 = 0$.

Ici, on s'intéresse à étudier des régions de plus en plus extrêmes, autrement dit, on étudie le comportement de la CCTE lorsque le paramètre $\alpha := \alpha_n$ dépend de n et tend vers zéro lorsque $n \to \infty$. A titre non exhaustif, on peut citer les travaux de Gardes and Girard (2010) qui étudient des quantiles extrêmes de lois à queues lourdes, et ceux de Methni et al. (2014) qui analysent des pertes extrêmes via des mesures de risques basées sur des moments conditionnels.
Notre premier résultat, à savoir le Théorème 1, lie la vitesse de convergence de la $\widehat{\mathrm{CCTE}}_{D,\alpha_n}$ à celle de la probabilité de la différence symmétrique entre $\mathcal{L}_D(\alpha_n)$ et $\mathcal{L}_n(\alpha_n)$ sous P. On introduit tout d'abord l'hypothèse suivante.

(H0) : pour un certain réel $\gamma \in [\frac{1}{2}, 1]$,

(i) la suite $u_n := \sqrt{n} P[\mathcal{L}_D(\alpha_n)]^{2-\gamma} \to \infty$, ou de façon équivalente,

$$n^{-\frac{1}{2}\left(\frac{1}{2-\gamma}\right)} = o_{n\to\infty}\left(P[\mathcal{L}_D(\alpha_n)]\right),$$

et

$$Y \in \mathbb{L}^{\frac{1}{1-\gamma}},$$

(ii) il existe une suite croissante de réels strictement positifs $(w_n)_n$ t.q. :

$$\frac{P[\mathcal{L}_D(\alpha_n) \Delta \mathcal{L}_n(\alpha_n)]^\gamma}{P[\mathcal{L}_D(\alpha_n)]} = \mathcal{O}_{P,n}\left(w_n^{-1}\right).$$

Dans l'esprit de Armaut et al. (2021), le Théorème 1 assure que l'estimateur de la CCTE est consistant avec une vitesse $w_n \wedge u_n$.

Théorème 1. *Soit $P \in \mathscr{P}$ et (α_n) une suite décroissante de réels dans $(0, \alpha_{\max}(P))$. Supposons que l'hypothèse **(H0)** est vérifiée. Alors,*

$$\left|\widehat{\mathrm{CCTE}}^{n}_{D,\alpha_n}(Y, \boldsymbol{X}) - \mathrm{CCTE}_{D,\alpha_n}(Y, \boldsymbol{X})\right| = \mathcal{O}_{P,n}(w_n^{-1} \vee u_n^{-1}).$$

Cas particulier : la profondeur à noyau gaussien.

La profondeur D_h à noyau gaussien est définie par (cf. Cuevas et al. (2007)) :

$$D_h(x, P) = \mathbb{E}_P\left[K_h(\|x - X\|)\right], \tag{3.4}$$

où $K_h(\cdot) = \frac{1}{h}K\left(\frac{\cdot}{h}\right), h > 0$, et $K(t) = \frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}$ est le noyau gaussien standard, et $\|\cdot\|$ est la norme euclidienne sur $\mathbb{R}^d$.
Un estimateur de D_h est alors donné par :

$$D_{h,n}(x) := D_h(x, P_n) = \frac{1}{n}\sum_{i=1}^{n} K_h(\|x - X_i\|). \tag{3.5}$$

Dans ce qui suit, on désigne par

$$\mathscr{P}_{m,\Sigma} := \left\{P \in \mathscr{P} | P \overset{law}{=} \mathbb{P}_{\mathcal{N}(m,\Sigma)} \text{ s.t. } m \in \mathbb{R} \text{ and } \Sigma \text{ is invertible}\right\} \subset \mathscr{P}$$

le modèle gaussien. Dans le cas de la profondeur D_h, on montre, par des raisonnements assez techniques, que l'hypothèse **(H0)** précédente est satisfaite avec les vitesses de convergence énoncées dans **(H1)** ci-dessous. D'après le Théorème 2, ceci nous permet de conclure quant à la vitesse de convergence de la CCTE dans ce modèle paramétrique. Pour un v.a. $\boldsymbol{X}$ de loi $P \in \mathscr{P}_{m,\Sigma}$, on désignera par $\lambda_1, \cdots, \lambda_d > 0$ les valeurs propres de la matrice de covariance Σ de $\boldsymbol{X}$.

(H1) Pour un certain $\gamma \in \left[\frac{1}{2}, 1\right]$, $Y \in \mathbb{L}^{\frac{1}{1-\gamma}}$. On considère

$$\delta_1 := \frac{1}{2}\frac{(2-\gamma)\left(1 + \frac{h^2}{\min_{1\le i\le d}\lambda_i}\right) - 1}{(2-\gamma)\left(1 + \frac{h^2}{\min_{1\le i\le d}\lambda_i}\right)} \in (0, 1/2),$$

$$\delta_2 := \frac{1}{2}\frac{1 + h^2\left(\frac{1}{\min_{1\le i\le d}\lambda_i} - \frac{\gamma}{\max_{1\le i\le d}\lambda_i}\right) - \gamma}{1 + h^2\left(\frac{1}{\min_{1\le i\le d}\lambda_i} - \frac{\gamma}{\max_{1\le i\le d}\lambda_i}\right)} \in (0, 1/2)$$

et $\delta \in (\max(\delta_1, \delta_2), 1/2)$. On introduit les deux suites suivantes

$$\tilde{u}_n := n^{\frac{1}{2} - \left(\frac{1}{2} - \delta\right)(2-\gamma)\left(1 + \frac{h^2}{\min_{1\le i\le d}\lambda_i}\right)} \log(n)^{\frac{(d-2)(2-\gamma)}{2}},$$

et

$$\tilde{w}_n := n^{\frac{\gamma}{2} - \left(\frac{1}{2} - \delta\right)\left(1 + \frac{h^2}{\min_{1\le i\le d}\lambda_i} - \gamma\frac{h^2}{\max_{1\le i\le d}\lambda_i}\right)} \log(n)^{\frac{d-2}{2}\left(1 - \frac{\gamma}{2}\right)}.$$

Théorème 2. *Soit $P \in \mathscr{P}_{m,\Sigma}$ et $h > 0$. On considère la suite $\alpha_n = z^*/n^{\frac{1}{2}-\delta}$ à valeurs dans $(0, \tilde{C}_{d,h}) = (0, \sup_u D_h(u, P))$ pour une certaine constante $z^* \in (0, \tilde{C}_{d,h})$. Alors, sous les notations de **(H1)***

$$\left|\widehat{\mathrm{CCTE}}^{n}_{D_h,\alpha_n}(Y, \boldsymbol{X}) - \mathrm{CCTE}_{D_h,\alpha_n}(Y, \boldsymbol{X})\right| = \mathcal{O}_{P,n}(\tilde{u}_n^{-1} \vee \tilde{w}_n^{-1}).$$

Références

E. Armaut, R. Diel, and T. Laloë. Estimation of the covariate conditional tail expectation : a depth-based level set approach. Working paper or preprint, September 2021. URL `https://hal.archives-ouvertes.fr/hal-03336080`.

F. Belzunce, A. Castaño, A. Olvera-Cervantes, and A. Suárez-Llorens. Quantile curves and dependence structure for bivariate distributions. *Computational Statistics & Data Analysis*, 51(10) :5112–5129, 2007.

F. Chebana and T. BMJ Ouarda. Multivariate quantiles in hydrological frequency analysis. *Environmetrics*, 22(1) :63–78, 2011.

A. Cousin and E. Di Bernardino. On multivariate extensions of value-at-risk. *Journal of multivariate analysis*, 119 :32–46, 2013.

A. Cuevas, M. Febrero-Bande, and R. Fraiman. Robust estimation and classification for functional data via projection-based depth notions. *Computational Statistics*, 22 :481–496, 02 2007.

L. Dehaan and X. Huang. Large quantile estimation in a multivariate setting. *Journal of Multivariate Analysis*, 53(2) :247–263, 1995.

Laurent Gardes and Stéphane Girard. Conditional extremes from heavy-tailed distributions : An application to the estimation of extreme rainfall return levels. *Extremes*, 13(2) :177–204, 2010.

T. Laloë, R. Servien, and E. Di Bernardino. Estimating covariate functions associated to multivariate risks : a level set approach. *Metrika, Springer Verlag*, pages 497–526, 2015.

Jonathan El Methni, Laurent Gardes, and Stéphane Girard. Non-parametric estimation of extreme risk measures from conditional heavy-tailed distributions. *Scandinavian Journal of Statistics*, 41(4) :988–1012, 2014.

R. Torres, E. Di Bernardino, H. Laniado, and R. Lillo. On the estimation of extreme directional multivariate quantiles. *Communications in Statistics-Theory and Methods*, 49(22) :5504–5534, 2020.

Y. Zuo and R. Serfling. General notions of statistical depth function. *Annals of statistics*, pages 461–482, 2000.

Kalman Recursions Aggregated Online

Eric Adjakossa [1] & Yannig Goude [2] & Olivier Wintenberger [3]

[1] *AgroParisTech, UMR 518 MIA-Paris, F-75005 Paris, France - eric.adjakossa@agroparistech.fr*

[2] *EDF Lab, OSIRIS Department 7 Boulevard Gaspard Monge, 91120 Palaiseau, France - yannig.goude@edf.fr*

[3] *Laboratoire de Probabilités, Statistique et Modélisation (LPSM, UMR 8001) Sorbonne Université 4 place Jussieu, 75005 Paris, France - olivier.wintenberger@upmc.fr*

Résumé. Dans ce travail, nous améliorons la qualité de la prédiction par agrégation d'experts en utilisant les propriétés sous-jacentes des modèles qui fournissent ces experts. Nous nous limitons au cas où les prédictions d'experts sont issues de récursions de Kalman par ajustement de modèles espace-état. En utilisant des poids exponentiels, nous avons construit différents algorithmes d'agrégation de récursions de Kalman en ligne (KAO) qui compétissent avec le meilleur expert ou la meilleure combinaison convexe des experts de façon adaptative ou non. Nous améliorons les résultats existants de la littérature sur l'agrégation d'experts lorsque les experts sont des récursions de Kalman en utilisant leurs propriétés de second ordre. Nous appliquons notre approche aux récursions de Kalman et l'étendons au contexte général d'experts en ajustant un modèle espace-état aux erreurs d'experts fournis.

Mots-clés. Agrégation en ligne, Filtre de Kalman

Abstract. In this work, we aim at improving the prediction of experts aggregation by using the underlying properties of the models that provide expert predictions. We restrict ourselves to the case where expert predictions come from Kalman recursions, fitting state-space models. By using exponential weights, we construct different algorithms of Kalman recursions Aggregated Online (KAO) that compete with the best expert or the best convex combination of experts in a more or less adaptive way. We improve the existing results on experts aggregation literature when the experts are Kalman recursions by taking advantage of the second-order properties of the Kalman recursions. We apply our approach to Kalman recursions and extend it to the general adversarial expert setting by state-space modeling the errors of the experts. We apply these new algorithms to a real data set of electricity consumption and show how it can improve forecast performances compared to other exponentially weighted average procedures.

Keywords. Online aggregation, Kalman filter

Motivation and results

Kalman filter (KF), also known as Linear Quadratic Estimator (LQE) is a powerful mathematical and algorithmic tool proposed by Kalman [1960]. In a context where observations of a random signal are observed in a stream fashion, it aims at tracking a target by recursively learning the parameters of an unobserved structure called the state of the process. Its recursive implementation makes it particularly useful to solve a wide range of real life and industrial problems. Among many industrial applications, one can cite the guidance of systems in aerospace embedded systems where the Markovian structure of Kalman allows memory savings McGee and Schmidt [1985], robotic Chen [2011], industrial electronics Auger et al. [2013] and energy management Dordonnat et al. [2012]. KF is also expansively used in econometric and more generally statistical time series analysis and forecasting where Kalman filter stands as a generalization of popular exponential smoothing with a time-dependent smoothing factor [Durbin and Koopman, 2012, Schneider, 1988, Hyndman et al., 2008]. In the field of atmospheric and oceanic meteorological forecasting, KF has been used this last 20 years in a data assimilation way to correct forecasts of high-dimensional numerical models for wind speed and wind energy forecasting Cassola and Burlando [2012], ensemble weather forecasting Mackenzie [2003], Evensen [2009], and air pollution forecasting De Ridder et al. [2012]. Ensemble Kalman Filters are stated Carrassi et al. [2018] to be among the most popular approaches for data assimilation in that field.

With the development of machine learning prediction methods, the access to more and more data and the diffusion of data science knowledge in industries, it is more and more common to have access to many alternatives for a single forecasting task. However KF is not applicable in a centralized model due to inherent drawbacks such as complexity, fault tolerance and cascaded processing that are well known in the federated KF literature initiated by Carlson and co-authors, see Carlson and Berarducci [1994] for instance. One solution is the distributed Kalman recursion proposed by Govaers and Koch [2012] but the methodology still relies deeply on well-specified state-space models. We depart from this federated KF literature as we instead use the unified theoretical framework for aggregating experts from Cesa-Bianchi and Lugosi [2006]. On the contrary to the KF federated learning, the individual sequence learning of Cesa-Bianchi and Lugosi [2006] investigates aggregation under general hypothesis that are set on the data generative process, allowing experts coming from very diverse methods, and observations y are only supposed to be bounded and sequentially observed: $y_1, \dots, y_n \in [-B, B]$ (B being an unknown constant). To forecast y according to its past values, we suppose to have access to a set of N experts producing forecasts of the sequence at each instant t - usually these experts are also based on some exogenous information X also observed online. After that, aggregation is computed step by step: $\hat{y}_t = \sum_{j=1}^{N} \hat{p}_t^{(j)} \hat{y}_t^j$ where the weights are updated according to past performances of each experts, supposing bounded losses. Many algorithms have been derived for that purpose since the seminal works of Littlestone and Warmuth [1994],

Freund and Schapire [1997] and Vovk [1998], among others. One can cite Herbster and Warmuth [1998] where the fixed-share algorithm is proposed to track the best expert changing with time, and Devaine et al. [2013] extend it to the case of specialized experts (experts activate during periods for which they are supposed to be designed). Recently, Gaillard et al. [2014], Wintenberger [2017] proposed online aggregation algorithms fully automatized in the sense that the multiple learning rates are sequentially updated using past data. In parallel with this theoretical development, aggregation of experts has been used in many real data application. Without being exhaustive we can cite applications in e-commerce Huard et al. [2020], photo-voltaic power forecasting Thorey et al. [2018], air quality forecasting Thorey et al. [2018], finance Amat et al. [2018], De Franco et al. [2020], electricity load forecasting Gaillard et al. [2016], Goehry et al. [2019] and district heating networks optimization Geysen et al. [2018].

We propose here to conciliate these two important fields of research by considering the question of online aggregation of experts coming from Kalman recursions to increase the accuracy of the prediction. We observe (y_t) sequentially through time $t \geq 1$ and predictions $\hat{y}_t^{(m)}$, $1 \leq m \leq M$, issued from Kalman recursions at time $t \geq 1$. Here $M \geq 0$ denotes the number of different Kalman recursions used as experts. The Kalman recursions are embedded into a state-space model for a formal definition). We introduce different Kalman recursions Aggregated Online (KAO) procedures that compute recursively weights $\rho_t^{(m)}$, $t \geq 1$, $1 \leq m \leq M$. We provide theoretical guarantees on the average prediction $\hat{y}_t = \sum_{m=1}^{M} \rho_t^{(m)} \hat{y}_t^{(m)}$.

We obtain bounds on the regret of KAO algorithms that are similar to the ones encountered in the individual sequence literature, see the book from Cesa-Bianchi and Lugosi [2006]. The novelty of our approach is to derive regret bounds directly on the cumulative quadratic prediction risk as defined by Wintenberger [2017]. The risk of prediction of $\hat{y} \in \mathcal{F}_{t-1}$ is defined as

$$L_t(\hat{y}) = E\left[(\hat{y} - y_t)^2 \mid \mathcal{F}_{t-1}\right], \qquad a.s., \qquad t \geq 1, \tag{1}$$

where $(\mathcal{F}_t)$ is the filtration including the past response variables $\sigma(y_s, X_s;\, 0 \leq s \leq t) \subset \mathcal{F}_t$, $t \geq 0$. The risk of prediction arises naturally when dealing with Kalman recursions. Indeed, Kalman recursions are online algorithms that provide the best linear predictions in Gaussian state-space models. We refer to the classical monograph from Durbin and Koopman [2012] for details. The cumulative prediction risk of a recursive algorithm predicting $\hat{y}_t$ at each time $t \geq 1$ is the sum $\sum_{t=1}^{T} L_t(\hat{y}_t)$ up to the horizon $T \geq 1$. Our regret bounds on KAO algorithm predictions $(\hat{y}_t)$ are a.s. deterministic bound called

respectively model selection regret or aggregation regret and defined as

$$R_T^S(m) \geq \sum_{t=1}^{T} L_t(\hat{y}_t) - L_t(\hat{y}_t^{(m)}) \qquad 1 \leq m \leq M\,, \tag{2}$$

$$R_T^A(\pi) \geq \sum_{t=1}^{T} L_t(\hat{y}_t) - L_t\left(\sum_{m=1}^{M} \pi^{(m)} \hat{y}_s^{(m)}\right)\,, \tag{3}$$

for any vector of weights $\pi := (\pi^{(m)})_{1\leq m\leq M}$ in the simplex. We suppress the dependence in m and π in R_t^S and R_t^A when the regret bounds are uniform in m and π, respectively. Bounds on the regrets, especially the model selection one, are very important since if the best state-space model $1 \leq m \leq M$ is correctly specified, and then the KF predictions $\hat{y}_t^{(m)}$ provides the smallest risk of prediction L_t among all online algorithms for any $t \geq 1$. Regret bounds on the cumulative prediction risk have attracted some interest since Audibert [2007] showed that the classical Exponentiated Weighted Aggregation (EWA) algorithm from Vovk [1990] does not achieve a fast rate model selection regret with high probability even in the most favorable iid case. The fast rate model selection regret is achieved for instance by the Bernstein Online Aggregation (BOA) algorithm in the iid setting with strongly convex loss in Wintenberger [2017] and then extended to any stochastic setting and exp-concave risk in Gaillard and Wintenberger [2016]. For adaptive procedures we improve their optimal regret

$$R_T^S = O(\log M + \log\log T + x)\,, \qquad T \geq 1$$

with probability $1 - e^{-x}$, $x > 0$, to the a.s. bound

$$R_T^S = O(\log M + \log\log T) \qquad T \geq 1\,,$$

with a leading constant divided by 2. This optimal regret bound holds when the observations satisfy some classical Gaussian state-space models. That the observations satisfy such a model is very unlikely in practice; It is the price to pay to guarantee a.s. optimal regret bounds on the cumulative prediction risk for unbounded responses. Existing regret bounds such as the one of Gaillard and Wintenberger [2016] requires the boundedness of the response, hypothesis which in turn is not satisfied by Gaussian state space models. In practice both BOA and KAO regret increase similarly but we see the improvement of a 1/2 factor of KAO regret w.r.t. BOA regret. We interpret this factor as the necessity to learn the variance of each expert for BOA whereas KAO uses the estimation of their own variance provided by each expert.

References

Christophe Amat, Tomasz Michalski, and Gilles Stoltz. Fundamentals and exchange rate forecastability with simple machine learning methods. *Journal of International Money and Finance*, 88:1–24, 2018.

Jean-Yves Audibert. No fast exponential deviation inequalities for the progressive mixture rule. *arXiv preprint math/0703848*, 2007.

François Auger, Mickael Hilairet, Josep M Guerrero, Eric Monmasson, Teresa Orlowska-Kowalska, and Seiichiro Katsura. Industrial applications of the kalman filter: A review. *IEEE Transactions on Industrial Electronics*, 60(12):5458–5471, 2013.

Neal A Carlson and Michael P Berarducci. Federated kalman filter simulation results. *Navigation, Journal of the institute of navigation*, 41(3):297–322, 1994.

Alberto Carrassi, Marc Bocquet, Laurent Bertino, and Geir Evensen. Data assimilation in the geosciences: An overview of methods, issues, and perspectives. *WIREs Climate Change*, 9(5):e535, 2018. doi: 10.1002/wcc.535. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/wcc.535`.

Federico Cassola and Massimiliano Burlando. Wind speed and wind energy forecast through kalman filtering of numerical weather prediction model output. *Applied Energy*, 99:154 – 166, 2012. ISSN 0306-2619. doi: https://doi.org/10.1016/j.apenergy.2012.03.054. URL `http://www.sciencedirect.com/science/article/pii/S0306261912002747`.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

SY Chen. Kalman filter for robot vision: a survey. *IEEE Transactions on Industrial Electronics*, 59(11):4409–4420, 2011.

Carmine De Franco, Christophe Geissler, Vincent Margot, and Bruno Monnier. Esg investments: Filtering versus machine learning approaches. *arXiv preprint arXiv:2002.07477*, 2020.

Koen De Ridder, Ujjwal Kumar, Dirk Lauwaet, Lisa Blyth, and Wouter Lefebvre. Kalman filter-based air quality forecast adjustment. *Atmospheric Environment*, 50:381 – 384, 2012. ISSN 1352-2310. doi: https://doi.org/10.1016/j.atmosenv.2012.01.032. URL `http://www.sciencedirect.com/science/article/pii/S1352231012000532`.

Marie Devaine, Pierre Gaillard, Yannig Goude, and Gilles Stoltz. Forecasting electricity consumption by aggregating specialized experts. *Machine Learning*, 90(2):231–260, 2013.

Virginie Dordonnat, Siem Jan Koopman, and Marius Ooms. Dynamic factors in periodic time-varying regressions with an application to hourly electricity load modelling. *Computational Statistics & Data Analysis*, 56(11):3134–3152, 2012.

James Durbin and Siem Jan Koopman. *Time series analysis by state space methods*. Oxford university press, 2012.

Geir Evensen. *Data assimilation: the ensemble Kalman filter*. Springer Science & Business Media, 2009.

Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

Pierre Gaillard and Olivier Wintenberger. Sparse accelerated exponential weights. *arXiv preprint arXiv:1610.05022*, 2016.

Pierre Gaillard, Gilles Stoltz, and Tim Van Erven. A second-order bound with excess losses. In *Conference on Learning Theory*, pages 176–196, 2014.

Pierre Gaillard, Yannig Goude, and Raphaël Nedellec. Additive models and robust aggregation for gefcom2014 probabilistic electric load and electricity price forecasting. *International Journal of forecasting*, 32(3):1038–1050, 2016.

Davy Geysen, Oscar De Somer, Christian Johansson, Jens Brage, and Dirk Vanhoudt. Operational thermal load forecasting in district heating networks using machine learning and expert advice. *Energy and Buildings*, 162:144–153, 2018.

Benjamin Goehry, Yannig Goude, Pascal Massart, and Jean-Michel Poggi. Aggregation of multi-scale experts for bottom-up load forecasting. *IEEE Transactions on Smart Grid*, 11(3):1895–1904, 2019.

Felix Govaers and Wolfgang Koch. An exact solution to track-to-track-fusion at arbitrary communication rates. *IEEE Transactions on Aerospace and Electronic Systems*, 48(3): 2718–2729, 2012.

Mark Herbster and Manfred K Warmuth. Tracking the best expert. *Machine learning*, 32(2):151–178, 1998.

Malo Huard, Rémy Garnier, and Gilles Stoltz. Hierarchical robust aggregation of sales forecasts at aggregated levels in e-commerce, based on exponential smoothing and Holt's linear trend method. working paper or preprint, June 2020. URL `https://hal.archives-ouvertes.fr/hal-02794320`.

Rob Hyndman, Anne B Koehler, J Keith Ord, and Ralph D Snyder. *Forecasting with exponential smoothing: the state space approach*. Springer Science & Business Media, 2008.

R. E. Kalman. A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering*, 82(1):35–45, 03 1960. ISSN 0021-9223. doi: 10.1115/1.3662552. URL https://doi.org/10.1115/1.3662552.

Nick Littlestone and Manfred K Warmuth. The weighted majority algorithm. *Information and computation*, 108(2):212–261, 1994.

Dana Mackenzie. Ensemble kalman filters bring weather models up to date. *Siam news*, 36(8):10–03, 2003.

Leonard A McGee and Stanley F Schmidt. Discovery of the kalman filter as a practical tool for aerospace and industry. 1985.

Wolfgang Schneider. Analytical uses of kalman filtering in econometrics—a survey. *Statistical Papers*, 29(1):3–33, 1988.

Jean Thorey, Christophe Chaussin, and Vivien Mallet. Ensemble forecast of photovoltaic power with online crps learning. *International Journal of Forecasting*, 34(4):762–773, 2018.

Vladimir Vovk. A game of prediction with expert advice. *Journal of Computer and System Sciences*, 56(2):153–173, 1998.

Vladimir G Vovk. Aggregating strategies. *Proc. of Computational Learning Theory, 1990*, 1990.

Olivier Wintenberger. Optimal learning with bernstein online aggregation. *Machine Learning*, 106(1):119–141, 2017.

Méthodes Génératives - ML - Deep

Étude du comportement de l'algorithme DbAS pour l'optimisation de propriétés de séquences nucléiques

Teddy Ardouin [1], Laurent Drazek [2], Pierre Mahé [3] & Adeline Samson [4]

[1] *bioMérieux, Université Grenoble-Alpes, teddy.ardouin@biomerieux.com*
[2] *bioMérieux , laurent.drazek@biomerieux.com*
[3] *bioMérieux , pierre.mahe@biomerieux.com*
[4] *Université Grenoble-Alpes, adeline.leclercq-samson@univ-grenoble-alpes.fr*

Résumé. Cette communication porte sur l'optimisation des propriétés d'une séquence de nucléotides résolue sur un espace discret. Dans cette perspective, nous étudions les performances de DbAS (Design by Adaptive Sampling), algorithme prometteur pour l'optimisation de propriétés de protéines telles que la conformation 3D ou la fluorescence. L'objectif est de générer un ensemble de séquences à la fois varié et possédant les meilleures valeurs d'une fonction f à optimiser. Un Auto-encodeur Variationnel (VAE) est utilisé comme modèle génératif associé à DbAS et deux types de fonctions f, unimodale et multimodale, sont étudiées. Bien que l'algorithme fonctionne bien pour des séquences nucléiques, l'augmentation de la longueur des séquences et l'optimisation d'une fonction multimodale dégradent les performances. Ces limitations proviennent à la fois du modèle génératif et de l'objectif de DbAS qui ne pénalise pas le manque de diversité.

Mots-clés. Optimisation, espace discret, séquences nucléiques, DbAS, Auto-Encodeur Variationnel, fonction multimodale, diversité

Abstract. We are interested in nucleotides sequences property optimisation. To solve this discrete space optimisation problem, we asses the DbAS (Design by Adaptive Sampling) algorithm and its level of performances. DbAS is a state of the art algorithm used for protein optimisation. Our purpose is to generate a set of sequences with high quality and diversity thanks to the function to optimize f. A Variational Autoencoder (VAE) is used as a generative model with DbAS and two different functions f are compared, a unimodal and a multimodal. Although the algorithm well performs for our problem, performances decrease with multimodal function and the increase of the sequence length. Previous limitations come from the generative model and the DbAS objective which does not enforce a diversity of solutions.

Keywords. Optimisation, discrete space, DbAS, nucleic sequences, Variational Autoencoder, multimodal function, diversity

1 Introduction

1.1 Contexte

L'optimisation de séquences discrètes est un domaine actif de recherche et présente des grands intérêts à la fois dans le monde académique et de l'industrie. Elle consiste à résoudre $\hat{x} = \text{argmax}_{x \in \mathcal{V}^L} f(x)$, $\mathcal{V}^L$ étant l'ensemble des séquences de longueur $L \in \mathbb{N}$ et de vocabulaire discret $\mathcal{V}$ et $f : \mathcal{V}^L \to \mathbb{R}$ la fonction à optimiser. La difficulté principale vient de la nature discrète de $\mathcal{V}$, qui empêche l'utilisation d'algorithmes d'optimisation classiques de type Newton, descente de gradient ou autres.

De nombreux cas d'applications existent dans la médecine et la biologie. Par exemple l'optimisation des propriétés d'une protéine dans l'article de Brookes et Listgarten [2] avec pour objectif de maximiser la fluorescence de ces molécules. La génération de nouvelles molécules pharmaceutiques plus efficaces ou moins toxiques entre aussi dans cette catégorie de problèmes avec l'étude de Gómez-Bombarelli *et al.* [3]. Enfin, il est possible d'optimiser les propriétés de séquences nucléiques pour des applications dans le diagnostic *in vitro*, ce qui est la motivation principale de notre travail. On peut prendre l'exemple de la détection du variant Omicron du SARS-CoV-2 par Lopez-Rincon *et al.* [5], .

Des algorithmes adaptés à la problématique d'optimisation d'une fonction f dans un espace discret existent dans la littérature, comme l'exemple des algorithmes génétiques. Nous nous intéressons à DbAS (Design by Adaptive Sampling) [2], un algorithme inspiré de l'évolution dirigée (méthode d'optimisation de protéines en laboratoire ayant valu le prix Nobel de chimie à Arnold [1] en 2018). Cette méthode adapte *in silico* l'évolution dirigée en utilisant un formalisme probabiliste et obtient des performances parmi les meilleures de l'état de l'art pour le design de protéines.

1.2 Problématique

Notre contribution est l'étude du comportement de l'algorithme DbAS pour l'optimisation de séquences nucléiques de petites tailles, pour différents types de fonctions f à optimiser, unimodale et multimodale. Notre objectif est de générer un ensemble diversifié de séquences parmi les meilleures pour la fonction f.

2 Méthodologie

2.1 Les fonctions f à optimiser

Nous étudions le comportement de DbAS pour deux fonctions f différentes appliquées à des séquences nucléiques de longueur L noté $x \in \{A, T, C, G\}^L$ [1].

[1] A, T, C et G en référence aux noms usuels des nucléotides constituant l'ADN (Adenine, Thymine, Cytosine et Guanine

La première, unimodale, est la température de fusion associée à une séquence ADN (notée Tm pour *melting temperature*). C'est une mesure d'intérêt en biologie moléculaire qui quantifie l'énergie de cohésion entre deux séquences d'un duplex ADN.

La seconde, multimodale, se base sur deux ensembles de séquences distincts. f mesure l'appartenance de la séquence nucléique d'intérêt à l'un des groupes tout en vérifiant l'exclusion de l'autre. Cette mesure possède aussi un intérêt dans de nombreuses applications en biologie moléculaire. Nous l'appelons pour la suite le critère inclus/exclus.

2.2 L'algorithme d'optimisation DbAS

DbAS est un algorithme itératif composé de deux éléments principaux: une fonction f à optimiser (qui dépend de l'application, exemple de la fluorescence d'une protéine avec Listgarten *et al.* [2]) et un modèle génératif noté $p(x|\theta)$ (paramétrisé par θ) capable de générer des nouvelles séquences à partir d'un jeu de données $X_N = \{x_i\}_{i=1}^N$. L'objectif de DbAS est de trouver le vecteur des paramètres θ^* du modèle génératif tel que: $\theta^* = \text{argmax}_\theta \, \mathbb{E}_{p(x|\theta)}[P(S|x)]$, S étant l'ensemble des séquences ayant une image par la fonction f supérieure à un seuil $\gamma \in \mathbb{R}$: $S = \{y = f(x) \geq \gamma\}$. Littéralement, l'objectif est de trouver les paramètres du modèle génératif qui maximisent la probabilité de générer des séquences dans l'ensemble S. Lors d'une phase d'apprentissage, les paramètres θ sont modifiés de sorte qu'une fonction de coût $\mathcal{L}(x, \theta)$, quantifiant par exemple l'erreur de reconstruction pour un Auto-encodeur Variationnel, soit minimale par rapport au jeu de données X_N. Le pseudo code de DbAS est présenté avec ci-dessous.

Algorithme 1 Algorithme DbAS

Paramètres $T \in \mathbb{N}$ nombre d'itérations, $M \in \mathbb{N}$ nombre fixé de séquences par itération, $\mathcal{L}(x, \theta)$ fonction de coût associée à $p(x|\theta)$, $f(X^{(t)}) = Y^{(t)} \in \mathbb{R}^M$ vecteur des valeurs de f, $S^{(t)}$ l'ensemble des composantes de $Y^{(t)}$ supérieures à $\gamma_Q^{(t)}$ avec $Q \in [0, 1]$, $W^{(t)} \in \mathbb{R}^M$ vecteur de poids et $CDF(y, \gamma)$ fonction de répartition de la distribution y.

$X^{(0)} \leftarrow \{x_1^{(0)}, ..., x_M^{(0)}\}$ ▷ jeu de données initial

$W^{(0)} \leftarrow \{w_1^{(0)}, ..., w_M^{(0)}\} = \{1, ..., 1\}$

$\theta^{(1)} \leftarrow argmin_\theta \sum_{j=1}^M w_j^{(0)} \mathcal{L}(x_j^{(0)}, \theta)$ ▷ entraînement initial du modèle génératif

Pour $t \leftarrow 1, T$ **faire**

  $X^{(t)} \leftarrow \{x_1^{(t)}, ..., x_M^{(t)}\}$ avec $x_j^{(t)} \sim p(x|\theta^{(t)})$

  $Y^{(t)} \leftarrow f(X^{(t)})$

  $S^{(t)} \leftarrow \{Y^{(t)} \geq \gamma_Q^{(t)}\}$ avec $\gamma_Q^{(t)} = Q^{ieme}$ centile des composantes de $Y^{(t)}$

  $W^{(t)} \leftarrow \{w_1^{(t)}, ..., w_M^{(t)}\}$ avec $w_j^{(t)} = P(S^{(t)}) = 1 - CDF(y_j^{(t)}, \gamma_Q^{(t)})$ et $\sum_j w_j^{(t)} = 1$

  $\theta^{(t+1)} \leftarrow argmin_\theta \sum_{j=1}^M w_j^{(t)} \mathcal{L}(x_j^{(t)}, \theta)$

Fin Pour

Retourne $X^{(T)}, \theta^{(T)}$

Au fil des itérations, le modèle génératif va proposer des meilleures séquences. Un avantage de cette méthode est qu'elle est capable de tirer partie d'une incertitude liée à la valeur de f (du bruit par exemple), qui se répercute directement dans le calcul des poids.

2.3 Un Auto-encodeur Variationnel comme modèle génératif

Nous utilisons à la manière de l'article initial un Auto-encodeur Variationel (VAE) [4] comme modèle génératif. Cependant, DbAS peut s'utiliser avec d'autres modèles génératifs (modèle de Markov, modèle de Markov caché, ...) comme le suggère Gane *et al.* [6].

Le VAE est composé de deux parties: un encodeur noté $q_\phi(z|x)$ et un décodeur $p_\theta(x|z)$, ϕ et θ les paramètres. L'encodeur a pour but de transformer x en une distribution dans laquelle on peut tirer une représentation continue z dans un espace $\mathbb{R}^d$ dit latent, de petite dimension. Le décodeur a le rôle inverse en retournant la séquence initiale à partir du vecteur latent. La réduction de dimension imposée par la dimension d de l'espace latent force le modèle à capturer l'information essentielle de x dans z. Dans notre application, l'encodeur et le décodeur seront des réseaux de neurones classiques multi-couches. Les séquences de taille L sont transformées au format one-hot (matrice de 0 et de 1 de taille $L \times 4$) puis passe dans l'encodeur pour retourner un vecteur de dimension d. Le décodeur prend en entrée un vecteur latent et retourne une matrice de taille $L \times 4$, sur laquelle on applique un opérateur Softmax qui calcule pour chaque position de la séquence une distribution de probabilité sur les différents nucléotides. Il est ainsi possible de déduire la séquence la plus probable décodée.

Lors de l'entraînement, le VAE a deux objectifs: reconstruire les données après projection dans $\mathbb{R}^d$ et organiser de manière régulière l'espace latent. Cela se répercute dans la fonction de coût. Pour une séquence $x \in \{A, T, C, G\}^L$:

$$\mathcal{L}(x, \theta, \phi) = \mathbb{E}_{q_\phi(z|x)}[log\ p_\theta(x|z)] - \beta * D_{KL}(q_\phi(z|x)\ ||\ p(z)) \quad \text{avec } p(z) = \mathcal{N}(0, I_d)$$

D_{KL} la divergence de Kullback-Liebler et I_d matrice identité de taille $d \times d$. Le premier terme quantifie la qualité de reconstruction et le second la régularité de l'espace latent. L'objectif étant que cet espace soit le plus proche de $p(z) = \mathcal{N}(0, I_d)$. L'hyperparamètre $\beta \in \mathbb{R}$ permet de gérer la balance entre les deux termes. Une fois entraîné et grâce aux propriétés de l'espace latent, des nouvelles séquences peuvent être générées avec le VAE à partir de vecteurs de $\mathbb{R}^d$ tirés dans $\mathcal{N}(0, I_d)$.

3 Résultats

Deux analyses ont été effectuées. La première compare les performances de DbAS appliqué à l'optimisation de Tm pour différentes tailles de séquences. L'objectif est d'analyser l'évolution des performances avec ce paramètre. Les tailles de séquences vont de 8 à 13 nucléotides, ce qui permet de rester dans des espaces de séquences de taille raisonnable: pour un temps de calcul correct, nous pouvons calculer analytiquement les maxima globaux. La deuxième analyse est à taille de séquences fixe égale à 8 et compare les performances pour les fonctions introduites précédemment. Il est ainsi possible d'évaluer si l'algorithme DbAS remplit nos objectifs de qualité et diversité pour une fonction multimodale et unimodale. Les deux études comptabilisent 50 lancements de DbAS, en utilisant

20 itérations et 2000 séquences générées pour chacune d'entre elles, conduisant à un budget de séquences de 40000. Les résultats obtenus sont présentés dans la Figure 1 et Table 1.

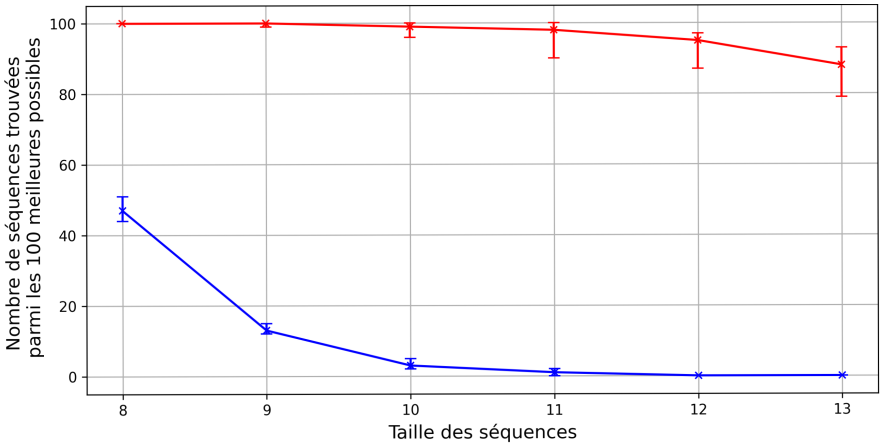


Figure 1: Comparaison des performances de DbAS (en rouge) et d'un tirage uniforme dans l'ensemble des séquences possibles (en bleu) pour différentes tailles de séquences nucléiques pour l'optimisation de la fonction Tm. En abscisse: taille de la séquence, en ordonnée: nombre médian de séquences trouvées parmi les 100 meilleures possibles. Les barres d'erreurs correspondent au 1^{er} et 3^{eme} quartile.

Fonction	Tm	Critère inclus/exclus
Moyenne (erreur type)	97.6 (1.1)	48.3 (1.0)

Table 1: Comparaison des performances de DbAS pour l'optimisation de Tm et du critère inclus/exclus pour des séquences de taille 8 nucléotides. Nombre moyen de séquences trouvées parmi les 100 meilleures.

4 Discussion

Nous pouvons remarquer avec la Figure 1 que l'optimisation via DbAS fonctionne, les performances sont meilleures qu'un tirage aléatoire uniforme dans l'espace des séquences. DbAS obtient une très bonne diversité de solutions parmi les meilleures possibles pour la plupart des tailles ce qui satisfait notre problématique pour une fonction f unimodale à optimiser. Cependant, cette diversité diminue légèrement avec l'augmentation de la taille des séquences. Elle reste modérée comparée à l'optimisation aléatoire qui ne trouve rapidement plus de solutions. Ainsi, les performances de DbAS pour notre problématique de diversité et qualité pour l'optimisation d'une fonction f unimodale sont très satisfaisantes.

Au niveau du type de fonction à optimiser sur la Table 1, il y a une différence claire

de performances. En moyenne, la quasi totalité des séquences parmi les 100 meilleures sont trouvées pour Tm et un peu moins de la moitié pour le critère d'inclus/exclus, ce qui n'est pas satisfaisant. DbAS ne parvient pas à explorer efficacement l'ensemble des meilleures solutions pour une fonction multimodale.

Nous faisons globalement face à un problème de diversité et d'exploration pour DbAS. Il provient de notre modèle génératif qui ne permet pas une exploration efficace et a tendance à rester sur des zones de l'espace des séquences précises. C'est un problème connu de la littérature lié au VAE, le flou de génération. En effet, il peut arriver au cours de DbAS que des séquences éloignées en terme de distance d'édition possèdent des poids élevés (souvent pour f multimodale), ce qui fournit au VAE des données d'apprentissage hétérogènes. Dans cette configuration, il est difficile de capturer l'hétérogénéité des données pour la génération. Le VAE a tendance à réaliser une génération moyennée par rapport aux données d'apprentissage ou à se bloquer sur un mode. Ces problèmes apparaissent moins fréquemment dans la littérature de l'optimisation de protéines: les objectifs y sont plus simples et les jeux de données plus structurés.

Une seconde explication aux faibles performances vis à vis de notre problématique est que DbAS ne force en aucun cas à la diversité en visant seulement à maximiser la probabilité de générer les meilleures séquences. L'ajout d'un terme dans l'objectif de DbAS lié à la diversité permettrait d'améliorer ce point, comme récemment suggéré par Listgarten *el al.* [7]. Cependant dans cet article, les auteurs doivent faire un compromis entre la qualité des séquences générées et leur diversité.

References

[1] Frances H. Arnold. Design by directed evolution. *Accounts of Chemical Research*, 31(3):125–131, 1998.

[2] David H. Brookes and Jennifer Listgarten. Design by adaptive sampling. *CoRR*, abs/1810.03714, 2018.

[3] Rafael Gòmez-Bombarelli *et al.* Automatic chemical design using a data-driven continuous representation of molecules. *ACS Central Science*, 4(2):268–276, 2018. PMID: 29532027.

[4] Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[5] Carmina A. Perez-Romero *et al.* Sars-cov-2 omicron variant ai-based primers. *bioRxiv*, 2022.

[6] Andreea Gane *et al.* . A comparison of generative models for sequence design. 2019.

[7] Danqing Zhu *et al.* Machine learning-based library design improves packaging and diversity of adeno-associated virus (aav) libraries. *bioRxiv*, 2021.

Limite d'échelle pour les réseaux de neurones résiduels

Pierre Marion[1*]

[1] *Sorbonne Université, CNRS, Laboratoire de Probabilités, Statistique et Modélisation, LPSM, F-75005 Paris, France*

Résumé. Les réseaux de neurones résiduels, introduits par He *et al.* (2016), ont permis des avancées majeures dans plusieurs domaines de recherche en apprentissage automatique, en permettant l'entraînement de réseaux avec plusieurs milliers de couches. Néanmoins, l'entraînement de ces réseaux reste complexe et instable, en particulier à cause d'un mauvais conditionnement du réseau lorsque la profondeur augmente. Nous nous intéresserons dans cet exposé à l'étude du comportement de ce type de réseaux lorsqu'on ajoute un facteur d'échelle qui dépend de la profondeur, dans la limite où la profondeur L est grande. Nous montrerons en particulier qu'à l'initialisation, le facteur d'échelle doit évoluer comme $1/\sqrt{L}$ pour obtenir une limite non dégénérée. Nous illustrerons ce résultat par des expériences simples. Nous ferons également le lien avec les modèles de réseaux de neurones en temps continu, et en particulier les modèles d'équations différentielles ordinaires neuronales introduites par Chen *et al.* (2018).

Mots-clés. Apprentissage statistique, théorie des réseaux de neurones, limite d'échelle

Abstract. Residual neural networks, introduced by He *et al.* (2016), lead to breakthroughs in several domains of machine learning, by allowing training of networks with several thousand layers. Nevertheless, training of these networks remains complex and instable, in particular due to a bad conditionning of the network when the depth increases. In this talk, we are interested in studying the behavior of this family of models when adding a scale factor depending on the depth L, in the infinite depth limit. We show that, at initialization, this factor must scale like $1/\sqrt{L}$ to obtain a non-degenerated limit. We will illustrate this result by simple experiments. We will also draw a link with continuous-time neural networks, in particular neural ordinary differential equations introduced by Chen *et al.* (2018).

Keywords. Statistical learning, theory of deep learning, scaling limits

1 Réseaux de neurones résiduels

Les modèles de réseaux de neurones résiduels (ResNets), introduits par He *et al.* (2016), ont permis des percées majeures, initialement en vision par ordinateur, puis plus récemment en compréhension du langage naturel (Vaswani *et al.*, 2017).

*L'auteur est soutenu par une bourse de la région Île-de-France et par MINES ParisTech.

Nous considérons ici le modèle suivant, dans un cadre d'apprentissage supervisé, où les données consistent en n couples i.i.d. $(x_i, y_i)_{1 \leq i \leq n}$, où $x_i \in \mathbb{R}^{n_{\text{in}}}$ et $y_i \in \mathbb{R}^{n_{\text{out}}}$. Ce cadre comprend à la fois la régression et la classification multi-classes, en encodant les classes sous forme de vecteurs de $\mathbb{R}^{n_{\text{out}}}$. Soit $f(\cdot, \theta_k) : \mathbb{R}^d \to \mathbb{R}^d$ une fonction paramétrée par θ_k, qui encode la k-ième couche du réseau, et $L \in \mathbb{N}^*$ la profondeur du réseau. Alors, pour un vecteur d'entrée x, la sortie du modèle $F_\theta(x)$ est donnée par la suite d'opérations

$$\begin{aligned} h_0 &= Ax, \\ h_{k+1} &= h_k + f(h_k, \theta_{k+1}) \quad \text{pour} \quad 0 \leq k \leq L-1, \\ F_\theta(x) &= Bh_L. \end{aligned} \tag{1}$$

Le modèle est paramétrisé par $\theta = (A, B, (\theta_k)_{1 \leq k \leq L}))$, où $A \in \mathbb{R}^{d \times n_{\text{in}}}$, $B \in \mathbb{R}^{n_{\text{out}} \times d}$. Pour fixer les idées, on peut avoir en tête l'exemple

$$f(h_k, \theta_{k+1}) = V_{k+1}\sigma(W_{k+1}h_k + b_{k+1}),$$

où σ est une fonction réelle appliquée composante par composante, et $\theta_k = (W_k, V_k, b_k)$, $W_k \in \mathbb{R}^{d \times d}$, $V_k \in \mathbb{R}^{d \times d}, b_k \in \mathbb{R}^d$. Ce choix de paramétrisation est très proche de celui proposé dans l'article originel sur les ResNets (He *et al.*, 2016). La perte sur les données d'entraînement s'écrit finalement $\mathcal{L}(\theta) = \frac{1}{n}\sum_{i=1}^n L(F_\theta(x_i), y_i)$, où L est une fonction de perte, comme la perte quadratique dans le cas régression ou l'entropie croisée dans le cas classification.

La caractéristique essentielle des réseaux de neurones résiduels est la présence de connexions résiduelles, c'est-à-dire que la fonction $f(\cdot, \theta_{k+1})$ ne paramétrise pas le nouvel état caché h_{k+1}, mais la différence entre deux états cachés successifs $h_{k+1} - h_k$. Cette idée a permis d'entraîner avec succès des réseaux très profonds, avec L de l'ordre de plusieurs centaines voire plusieurs milliers, ce qui n'était pas possible avec des réseaux de neurones classiques.

Néanmoins, même avec les connexions résiduelles, l'entraînement reste difficile à cause de problèmes d'instabilités. Ces instabilités sont liées à une évolution des moments de h_k en fonction de k (Shao *et al.*, 2020). La façon la plus courante de remédier à ce problème en pratique est d'ajouter une étape de normalisation, comme la normalisation par batch (Ioffe et Szegedy, 2015), qui consiste à effectuer une transformation affine après chaque couche pour que la sortie de la couche soit d'espérance nulle et de variance unitaire. Néanmoins, ce choix de normalisation introduit des difficultés pratiques et théoriques (Brock *et al.*, 2021), ce qui pousse à chercher des alternatives.

Une alternative populaire est d'introduire un facteur d'échelle devant le terme résiduel $f(\cdot, \theta_k)$, c'est-à-dire de remplacer le modèle (1) par

$$\begin{aligned} h_0 &= Ax, \\ h_{k+1} &= h_k + \frac{1}{L^\beta} f(h_k, \theta_{k+1}) \quad \text{pour} \quad 0 \leq k \leq L-1, \\ F_\theta(x) &= Bh_L. \end{aligned} \tag{2}$$

Il n'y a pas de consensus sur la forme que devrait prendre ce facteur d'échelle. Par exemple, Hanin et Rolnick (2018) affirment qu'il faut $\beta > 1$ pour garantir la stabilité du réseau, alors que Zhang *et al.* (2019) montrent qu'il suffit d'avoir $\beta > 1/2$. Cohen *et al.* (2021) suggèrent que le facteur d'échelle observé en pratique est $\beta \approx 0.7$, tandis que De et Smith (2020) observent empiriquement que la normalisation par batch est équivalente à prendre $\beta \approx 1/2$, qui est aussi le facteur conseillé par Arpit *et al.* (2019).

Cela pousse à s'intéresser au comportement du modèle (2) lorsque $L \to +\infty$, en fonction de $1/L^\beta$, afin de trouver quelles valeurs de β amènent de bonnes propriétés pour le modèle.

2 Comportement à l'initialisation dans la limite de la grande profondeur

A l'initialisation du modèle, c'est-à-dire avant l'apprentissage, les paramètres θ sont choisis de manière aléatoire. Nous supposons donc que $\theta_1, \dots, \theta_L$ est une suite de vecteurs i.i.d., ce qui est la situation typique. Dans ce cas, les états cachés $h_1, \dots, h_L$ sont aussi des vecteurs aléatoires, mais pas i.i.d.—sous nos hypothèses, cette suite est en fait une martingale.

Nous faisons les hypothèses suivantes, relativement peu contraignantes, qui portent sur le choix à la fois de l'architecture du réseau, c'est-à-dire de la fonction f, et de la stratégie d'initialisation, c'est-à-dire de la loi des θ_k.

Hypothèses Pour tout $0 \leq k \leq L-1$,

(A_1) conditionnellement à h_k, les coordonnées de $f(h_k, \theta_{k+1})$ sont des variables identiquement distribuées, décorrélées, et centrées.

(A_2) $\frac{\|h_k\|^2}{2} \leq \mathbb{E}\big(\|f(h_k, \theta_{k+1})\|^2 \big| h_k\big) \leq \|h_k\|^2$.

(A_3) il existe une constante $C \geq 1$ (qui dépend uniquement du choix de f et de la dimension d) telle que $\|f(h_k, \theta_{k+1})\|^2 \leq C\|h_k\|^2$ presque sûrement.

L'hypothèse (A_1) est naturelle et implique en particulier que $(h_k)_{1 \leq k \leq L}$ est une martingale. L'hypothèse (A_2) assure que $f(\cdot, \theta_{k+1})$ est proche d'être une isométrie en espérance. Enfin, l'hypothèse (A_3) assure que la norme d'opérateur de $f(\cdot, \theta_{k+1})$ est bornée presque sûrement. Ces hypothèses sont vérifiées par un grand nombre de choix d'architectures de réseaux de neurones et de stratégies d'initialisation. La proposition suivante donne deux exemples standards qui vérifient les hypothèses. Notons que nous n'avons pas inclus de terme de biais comme le biais est communément initialisé à zéro.

Proposition 1 *Les hypothèses (A_1) à (A_3) sont vérifiées dans les cas suivants :*

1. *si* $f(h_k, \theta_{k+1}) = V_{k+1}\max((W_{k+1}h_k), 0)$*, où les matrices* V_{k+1} *et* W_{k+1} *sont indépendantes et leurs coefficients sont des variables i.i.d. suivant la loi uniforme* $\mathcal{U}(-\sqrt{3/d}, \sqrt{3/d})$*.*

2. *si* $f(h_k, \theta_{k+1}) = V_{k+1}\sigma(h_k)$*, où les coefficients de* V_{k+1} *sont des variables i.i.d. suivant la loi uniforme* $\mathcal{U}(-\sqrt{3/d}, \sqrt{3/d})$*, et* $\sigma : \mathbb{R} \to \mathbb{R}$ *est telle que* $ax \leq |\sigma(x)| \leq bx$ *pour un certain* $a \geq 1/\sqrt{2}$ *et* $b \leq 1$*.*

Dans la proposition ci-dessus, les fonctions $\max(\cdot, 0)$ et σ sont appliquées composante par composante.

Taille de la dernière couche cachée Le résultat suivant s'intéresse au comportement de la taille de la dernière couche cachée en fonction de β. Nous le présentons ici de manière relativement informelle pour ne pas alourdir le propos. Par exemple, nous avons montré des bornes non-asymptotiques pour $\|h_L - h_0\|$ et $\|h_L\|$, que nous n'exposons pas ici, et qui impliquent en particulier les convergences dans les cas $\beta < 1/2$ et $\beta > 1/2$.

Théorème 1 *Considérons le modèle* (2) *sous les hypothèses* (A_1) *à* (A_3)*. Nous pouvons distinguer trois régimes :*

- *si* $\beta < 1/2$*,*
$$\|h_L\| \xrightarrow{\mathbb{P}} +\infty \text{ lorsque } L \to \infty.$$
- *si* $\beta > 1/2$*,*
$$\|h_L - h_0\| \xrightarrow{\mathbb{P}} 0 \text{ lorsque } L \to \infty.$$
- *si* $\beta = 1/2$*, il existe une constante* c *tel que pour* L *assez grand et pour tout* $\lambda > 0$*,*
$$\exp(-\lambda)\exp\left(\frac{1}{4}\right)\|h_0\|^2 \leq \|h_L\|^2 \leq \exp(\lambda)\|h_0\|^2$$
avec probabilité au moins $1 - 2\exp\big(-c\lambda^2\big)$*.*

Dans le premier régime, le modèle est instable au sens où on observe une explosion de la sortie dans la limite où L est grand. La stabilité est retrouvée lorsque $\beta \leq 1/2$. Néanmoins, lorsque $\beta < 1/2$, on montre que le modèle dégénère vers l'identité dans la limite où L est grand, au sens où la dernière couche cachée h_L est asymptotiquement égale à la première couche cachée h_0. Ce comportement n'est pas souhaitable, car il signifie que la sortie du modèle ne dépend pas de la valeur des paramètres θ_k. Enfin, dans le cas limite $\beta = 1/2$, la sortie h_L admet des fluctuations dont la taille est indépendante de L. En résumé, ce résultat montre que le seul cas où le comportement asymptotique de h_L à l'initialisation n'est pas trivial est $\beta = 1/2$.

Ce résultat est vérifié par des expériences numériques simples.

Gradients Le Théorème 1 s'intéresse au comportement limite de la dernière couche cachée h_L. Une autre quantité intéressante à regarder est la norme des gradients $p_k := \frac{\partial L}{\partial h_k}$, pour $0 \leq k \leq L$. En effet, les réseaux de neurones étant entraînés par descente de gradient, il est essentiel de s'assurer que les gradients ne sont ni trop grands, afin de ne pas faire diverger la descente de gradient, ni trop petits, afin de permettre l'apprentissage. Les gradients p_k vérifient une équation de récurrence, cette fois pour k décroissant, connue sous le nom de méthode de l'adjoint ou de *backpropagation*. Par exemple, dans le cas 2. de la Proposition 1, la récurrence sur p_k s'écrit

$$p_k = p_{k+1} + \frac{1}{L^\beta} V_{k+1}^T (\sigma'(h_k) \odot p_{k+1}),$$

où $\odot$ désigne le produit terme à terme de deux vecteurs. La ressemblance qualitative entre cette équation et la récurrence (2) indique qu'il est raisonnable d'imaginer que les gradients suivent un comportement similaire au Théorème 1 : si $\beta < 1/2$, p_0 explose par rapport à p_L. Si $\beta > 1/2$, p_0 est asymptotiquement égal à p_L. Cela signifie qu'à l'initialisation, les gradients ne retiennent pas d'informations liées aux poids sur les couches intermédiaires, ce qui n'est pas souhaitable et pourrait expliquer le ralentissement de l'apprentissage observé dans ce cas par Zhang *et al.* (2019).

Ce comportement est vérifié par des expériences numériques simples.

3 Lien avec les modèles en temps continu

Pour terminer le travail, nous faisons le lien avec les modèles de réseaux de neurones en temps continu. Lorsque $\beta = 1$ et que les poids sont suffisamment réguliers, le modèle (1) peut être vu comme la discrétisation d'Euler d'un modèle continu, connu sous le nom d'équation différentielle ordinaire (EDO) neuronale (Chen *et al.*, 2018), qui s'écrit

$$\frac{dh_t}{dt} = f(h_t, \theta_t). \tag{3}$$

Ce type de modèle continu a reçu un écho important, du fait de son attrait conceptuel et pratique. Néanmoins, l'étude précédente montre que ce n'est pas le bon modèle limite à l'initialisation, puisqu'on arrive à la conclusion que le seul facteur d'échelle qui donne un résultat non trivial à l'initialisation est $\beta = 1/2$, et non $\beta = 1$. D'autre part, en prenant des poids i.i.d. sur chaque couche, la limite continue des poids n'est pas assez régulière pour donner un sens à l'EDO (3), mais correspond plutôt à une diffusion. Nous discuterons de ces aspects et proposerons un modèle limite plus pertinent que (3) pour rendre compte du comportement du réseau à l'initialisation dans la limite $L \to +\infty$.

Bibliographie

Devansh Arpit, Víctor Campos et Yoshua Bengio : How to initialize your network? robust initialization for weightnorm & resnets. *In Advances in Neural Information Processing Systems*, volume 32, pages 10902–10911. Curran Associates, Inc., 2019.

Andrew Brock, Soham De et Samuel L Smith : Characterizing signal propagation to close the performance gap in unnormalized resnets. *In Proceedings of the 9th International Conference on Learning Representations, ICLR*, 2021.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt et David K Duvenaud : Neural ordinary differential equations. *In Advances in Neural Information Processing Systems*, volume 31, pages 6572–6583. Curran Associates, Inc., 2018.

Alain-Sam Cohen, Rama Cont, Alain Rossier et Renyuan Xu : Scaling properties of deep residual networks. *In Proceedings of the 38nd International Conference on International Conference on Machine Learning*, pages 2039–2048. JMLR.org, 2021.

Soham De et Sam Smith : Batch normalization biases residual blocks towards the identity function in deep networks. *In Advances in Neural Information Processing Systems*, volume 33, pages 19964–19975. Curran Associates, Inc., 2020.

Boris Hanin et David Rolnick : How to start training: The effect of initialization and architecture. *In Advances in Neural Information Processing Systems*, volume 31, pages 571–581. Curran Associates, Inc., 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren et Jian Sun : Deep residual learning for image recognition. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

Sergey Ioffe et Christian Szegedy : Batch normalization: Accelerating deep network training by reducing internal covariate shift. *In Proceedings of the 32nd International Conference on International Conference on Machine Learning*, page 448–456. JMLR.org, 2015.

Jie Shao, Kai Hu, Changhu Wang, Xiangyang Xue et Bhiksha Raj : Is normalization indispensable for training deep neural network? *In Advances in Neural Information Processing Systems*, volume 33, pages 13434–13444. Curran Associates, Inc., 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser et Illia Polosukhin : Attention is all you need. *In Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Huishuai Zhang, Da Yu, Mingyang Yi, Wei Chen et Tie-Yan Liu : Convergence theory of learning over-parameterized resnet: A full characterization. *arXiv:1903.07120*, 2019.

Can Push-forward Generative Models Fit Multimodal Distributions?

Antoine Salmona [1] & Julie Delon [2] & Agnès Desolneux [3] & Valentin De Bortoli [4]

[1] *CNRS, ENS Paris-Saclay, Centre Borelli UMR 9010, antoinesalmona2@gmail.com*
[2] *Université de Paris, CNRS, MAP5 UMR 8145, julie.delon@parisdescartes.fr*
[3] *CNRS, ENS Paris-Saclay, Centre Borelli UMR 9010, agnes.desolneux@ens-paris-saclay.fr*
[4] *CNRS, DI/ENS Ulm UMR 8548, valentin.debortoli@gmail.com*

Résumé. Les modèles génératifs sont aujourd'hui l'un des sujets de recherche les plus populaires en apprentissage automatique, notamment grâce à leur impressionnante capacité à générer des images synthétiques réalistes. Cependant, il reste souvent difficile de savoir si ces modèles s'approchent correctement de la distribution sous-jacente des données ou s'ils génèrent uniquement des échantillons qui semblent similaires aux données. Dans ce travail, nous nous concentrons sur la classe particulière des modèles génératifs *push-forward*, qui inclut les *Variational Autoencoders*, les *Generative Adversarial Networks* et les *Normalizing Flows*. Nous montrons que ces modèles doivent avoir de grandes constantes de Lipschitz afin de bien approcher les distributions multimodales. Ainsi, il existe pour ces modèles une compétition entre la stabilité de leur apprentissage et leur capacité à générer des distributions multimodales.

Mots-clés. Modèles Génératifs, Réseaux de Neurones, Apprentissage Profond, Constante de Lipschitz, Mélange de Gaussiennes.

Abstract. Generative models are today one of the most popular research topic in machine learning thanks to their impressive ability to generate realistic synthetic images. However, it remains unclear whether these models approach correctly the underlying data-distribution or only generate samples that seem similar to the data. In this work, we focus on the particular class of the push-forward generative models, which includes Variational Autoencoders, Generative Adversarial Networks and Normalizing Flows. We show that these models must have large Lipschitz constants in order to approximate multimodal distributions. Thus, there is for these models a contradiction between their training stability and their ability to generate multimodal distributions.

Keywords. Generative Models, Neural Networks, Deep Learning, Lipschitz Constant, Gaussian Mixture Model.

Generative modeling has become over the last past years one of the most popular research topic in machine learning and computer vision. From a mathematical point of view, the goal of generative modeling can be seen as predicting new synthetic samples from an unknown probability distribution ν on $\mathbb{R}^n$ given the information of m "true" samples

x_i drawn from ν (the data distribution). A general approach to solve this problem is to define a parametric family of probability distributions $(\nu_\theta)_{\theta\in\Theta}$ and solve the problem

$$\min_{\theta\in\Theta} d\left(\frac{1}{m}\sum_i \delta_{x_i}, \nu_\theta\right), \tag{1}$$

where d is a measure of similarity between probability distributions (d can either be a ϕ-divergence or a distance) and δ_{x_i} is the dirac distribution in x_i. Beside their direct application (generate new data) [15, 1], generative models have been used in numerous applications in various machine learning subfields, such as for instance solving inverse problems [13, 9] or machine translation [7, 19]. However, most of generative modeling methods still lack theoretical understanding and it remains often unclear whether the method approaches correctly the probability distribution ν or only generates samples that appear to have been drawn from ν without really recovering the underlying structure of the distribution. In this work, we focus on a particular class of generative models that we call *push-forward models*. Those models have in common that for any $\theta \in \Theta$, the parametric distribution ν_θ approaching ν is of the form

$$\nu_\theta = g_\theta \# \mu_d, \tag{2}$$

where $\mu_d = \mathcal{N}(0, I_d)$ is the Gaussian standard distribution in dimension d, $\#$ is the push-forward operator[1], and $g_\theta : \mathbb{R}^d \to \mathbb{R}^n$ is a deterministic neural network of parameter θ. This class includes three of the most popular generative models: the *Variational Auto-Encoders* (VAEs) [8], the *Generative Adversarial Networks* (GANs) [6], and the *Normalizing Flows* (NFs) [14].

Deep neural networks are most of the time Lipschitz functions by design, since their activation functions are generally Lipschitz. This is mainly due to the fact that deep neural networks have to be differentiable almost everywhere in order to be trained. Moreover, it has been observed that constraining the Lipschitz constant of a neural network to be small increases its robustness [18, 5], and stabilizes its training [11]. For this reason, several state-of-the-art GANs normalize the spectral norms of the weight matrices of the generative network, such as for instance BigGAN (Brock et al. [3]) or SAGAN (Zhang et al. [20]).

Yet, by studying the properties of the push-forward measure $g\#\mu_d$, with $g : \mathbb{R}^d \to \mathbb{R}^n$ being a Lipschitz function, we show that when the data-distribution ν is multimodal, the Lipschitz constant of g has to be large in order to well approximate ν. This is even more critical when the support of ν and $\mathbb{R}^d$ have distinct topologies, which occurs for instance when the support of ν is disconnected, or when the support of ν has "holes": in that case, we show that there doesn't exist any Lipschitz function g such that $\nu = g\#\mu_d$. Thus,

[1] If μ is a measure on $\mathbb{R}^d$ and f is a mapping from $\mathbb{R}^d$ to $\mathbb{R}^n$, the push-forward measure $f\#\mu$ is the measure on $\mathbb{R}^n$ such that for all Borel set A of $\mathbb{R}^n$, $f\#\mu(\mathsf{A}) = \mu(f^{-1}(\mathsf{A}))$.

for push-forward generative models, there is a competition between the stability of the training which requires a small Lipschitz constant, and the ability to generate multimodal distributions which requires a large Lipschitz constant. As an outcome, when trained on multimodal distributions, the push-forward generative models generally interpolates between the modes of the distribution, or generates points from only one mode, which is well known as *mode collapsing* [4, 17] in the GAN literature.

Our results are mainly consequences of the Gaussian Isoperimetric Inequality (Borell [2], Sudakov and Tsirelson [16], or see Ledoux [10] for a short proof) which states, in its differentiate form, that for any Borel set A of $\mathbb{R}^d$

$$\mu_d^+(\partial \mathsf{A}) \geq \varphi(\Phi^{-1}(\mu_d(\mathsf{A}))), \tag{3}$$

where $\varphi(x) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{x^2}{2})$ is the density of μ_1, $\Phi(x) = \int_{-\infty}^{x} \varphi(t)dt$ is the cumulative distribution function of μ_1 and where for any measure γ on $\mathbb{R}^d$ and any subset A of $\mathbb{R}^d$, $\gamma^+(\partial \mathsf{A})$ is the γ-surface area of A, defined by

$$\gamma^+(\partial \mathsf{A}) = \liminf_{\varepsilon \to 0^+} \frac{\gamma(\mathsf{A}_\varepsilon) - \gamma(\mathsf{A})}{\varepsilon}, \tag{4}$$

where $\mathsf{A}_\varepsilon = \{x \in \mathbb{R}^d \mid \inf_{a \in \mathsf{A}} \|x - a\| \leq \varepsilon\}$ is the ε-extension of A.

From (3), one can deduce the following property on the push-forward measure $g\#\mu_d$.

Theorem 1. *Let $g : \mathbb{R}^d \to \mathbb{R}^n$ be a Lipschitz function with Lipschitz constant* $\mathrm{Lip}(g)$*. Then for any Borel set A of $\mathbb{R}^n$,*

$$\mathrm{Lip}(g)(g\#\mu_d)^+(\partial \mathsf{A}) \geq \varphi(\Phi^{-1}(g_\#\mu_d(\mathsf{A}))), \tag{5}$$

In addition, we have that for any $r \geq 0$

$$g_\#\mu_p(A_r) \geq \Phi(r/\mathrm{Lip}(g) + \Phi^{-1}(g_\#\mu_d(\mathsf{A}))). \tag{6}$$

This theorem implies in particular that the Lipschitz constant of g has to be large in order for $g\#\mu_d$ to be multimodal. For instance, when $\nu = \frac{1}{2}(\mathcal{N}(-m, \sigma^2 I_n) + \mathcal{N}(m, \sigma^2 I_n))$ is an equilibrated mixture of two isotropic Gaussian distributions, Theorem 1 allows us to derive that if g is a Lipschitz function such that $g\#\mu_d = \nu$, then one must have

$$\mathrm{Lip}(g) \geq \sigma \exp\left(\frac{\|m\|^2}{2\sigma^2}\right). \tag{7}$$

Moreover, if the support of the target distribution ν and the support of μ_d have distinct topologies, there doesn't exist any Lipschitz function g such that $g\#\mu_d = \nu$. This typically occurs when the support of ν is disconnected, or when the support of ν has "holes". A last consequence of Theorem 1 is that when ν is a probability measure on $\mathbb{R}$ with a continuous density with respect to the Lesbegue measure and with support equal to $\mathbb{R}$, the Monge

map [12] between μ_1 and ν is the mapping with smallest Lipschitz constant which achieves $g\#\mu_1 = \nu$.

Another consequence of the Gaussian Isoperimetric Inequality is that the total variation distance between $g\#\mu_d$ and ν is bounded from below when ν is multimodal and the Lipschitz constant of g is not large enough.

Theorem 2. *Let ν be a probability measure on $\mathbb{R}^n$ and let $g : \mathbb{R}^d \to \mathbb{R}^n$ be a Lipschitz function. Let*

$$d_{TV}(g\#\mu_d, \nu) = \sup_{\mathsf{A} \in \mathcal{B}(\mathbb{R}^n)} |g\#\mu(\mathsf{A}) - \nu(\mathsf{A})|, \tag{8}$$

be the total variation distance between $g\#\mu_d$ and ν, and where $\mathcal{B}(\mathbb{R}^n)$ is the Borel algebra on $\mathbb{R}^n$. Then

$$d_{TV}(g\#\mu_d, \nu) \geq \sup_{\mathsf{A} \in \mathcal{B}(\mathbb{R}^n), r>0} \Phi\left(r/\mathrm{Lip}(g) + \Phi^{-1}(g\#\mu_d(\mathsf{A}))\right) - g\#\mu_d(\mathsf{A}) - \nu(\mathsf{A}_r \setminus \mathsf{A}). \tag{9}$$

One can derive a similar lower bound for the Kullback-Leibler divergence between $g\#\mu_d$ and ν. This general bound is not necessarily meaningful in itself since the right-hand term in (9) can be negative. Yet it allows to derive explicit lower bounds in particular cases. For instance, when $\nu = \frac{1}{2}(\mathcal{N}(-m, I_n) + \mathcal{N}(m, I_n))$ with $m \in \mathbb{R}^n$, we can derive an explicit lower bound on the total variation distance between $g\#\mu_d$ and ν:

$$d_{TV}(g\#\mu_d, \nu) \geq \int_0^{\frac{\|m\|}{2\mathrm{Lip}(g)}} \varphi(t)dt - \frac{1}{2}\int_{\frac{\|m\|}{2}}^{\frac{3\|m\|}{2}} \varphi(t)dt. \tag{10}$$

When $\|m\|$ goes to infinity, the first term of the lower bound grows to $1/2$, and the second term decreases to 0.

References

[1] Antreas Antoniou, Amos Storkey, and Harrison Edwards. Augmenting image classifiers using data augmentation generative adversarial networks. *International Conference on Artificial Neural Networks*, 27:594–603, 2018.

[2] Christer Borell. The Brunn-Minkowski inequality in Gauss space. *Inventiones Mathematicae*, 30(2):207–216, 1975.

[3] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. *International Conference on Learning Representations*, 6, 2018.

[4] Ricard Durall, Avraam Chatzimichailidis, Peter Labus, and Janis Keuper. Combating mode collapse in GAN training: An empirical analysis using Hessian eigenvalues. *International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 16:211–218, 2021.

[5] Mahyar Fazlyab, Alexander Robey, Hamed Hassani, Manfred Morari, and George Pappas. Efficient and accurate estimation of Lipschitz constants for deep neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

[6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27, 2014.

[7] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *Conference On Computer Vision and Pattern Recognition*, 31:1125–1134, 2017.

[8] Diederik P Kingma and Max Welling. Auto-Encoding Variational Bayes. *Stat*, 1050:1, 2014.

[9] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. *Conference on Computer Vision and Pattern Recognition*, 31:4681–4690, 2017.

[10] Michel Ledoux. A short proof of the Gaussian isoperimetric inequality. *High Dimensional Probability*, 43:229–232, 1998.

[11] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *International Conference on Learning Representations*, 6, 2018.

[12] Gabriel Peyré and Marco Cuturi. Computational optimal transport: with applications to data science. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019.

[13] Suman Ravuri, Karel Lenc, Matthew Willson, Dmitry Kangin, Remi Lam, Piotr Mirowski, Megan Fitzsimons, Maria Athanassiadou, Sheleem Kashem, Sam Madge, et al. Skilful precipitation nowcasting using deep generative models of radar. *Nature*, 597(7878):672–677, 2021.

[14] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. *International Conference on Machine Learning*, 32:1530–1538, 2015.

[15] Veit Sandfort, Ke Yan, Perry J Pickhardt, and Ronald M Summers. Data augmentation using generative adversarial networks (CycleGAN) to improve generalizability in ct segmentation tasks. *Scientific Reports*, 9(1):1–9, 2019.

[16] Vladimir N Sudakov and Boris S Tsirelson. Extremal properties of half-spaces for spherically invariant measures. *Journal of Soviet Mathematics*, 9(1):9–18, 1978.

[17] Hoang Thanh-Tung and Truyen Tran. On catastrophic forgetting and mode collapse in generative adversarial networks. *International Joint Conference on Neural Networks*, 32, 2020.

[18] Aladin Virmaux and Kevin Scaman. Lipschitz regularity of deep neural networks: analysis and efficient estimation. *Advances in Neural Information Processing Systems*, 31, 2018.

[19] Zhen Yang, Wei Chen, Feng Wang, and Bo Xu. Improving neural machine translation with conditional sequence generative adversarial nets. *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1:1346–1355, 2018.

[20] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. *International Conference on Machine Learning*, 36:7354–7363, 2019.

Generating new crystal structures with statistical methods

Arsen Sultanov[1] & Jean-Claude Crivello[2] & Tabea Rebafka[3]& Nataliya Sokolovska[4]

[1] *Univ Paris Est Creteil, CNRS, ICMPE, UMR 7182, 2 rue Henri Dunant, 94320 Thiais, France, arsen.sultanov@cnrs.fr*
[2] *Univ Paris Est Creteil, CNRS, ICMPE, UMR 7182, 2 rue Henri Dunant, 94320 Thiais, France, crivello@cnrs.fr*
[3] *LPSM, Sorbonne Université, Université de Paris & CNRS, 4, Place Jussieu, 75252, Paris cedex 05, France, tabea.rebafka@sorbonne-universite.fr*
[4] *Sorbonne University, INSERM, NutriOmics, Paris, France, Nataliya.Sokolovska@sorbonne-universite.fr*

Résumé. Dans cette étude, nous proposons une méthode basée sur un modèle probabiliste de *denoising diffusion* pour générer des matériaux hypothétiques définis par leurs structures cristallines. Nous expliquons le modèle statistique sous-jacent, comment il peut être appliqué à la génération de structure cristalline et pourquoi cette approche est bien justifiée à ce problème.

Mots-clés. Modèle génératif, modèle de diffusion, réseau de neurons, dynamique de langevin, science des matériaux

Abstract. In this study, we propose a method for generating hypothetical crystal material structures based on a denoising diffusion probabilistic model. We explain the problem and the underlying statistical model, how it can be applied to crystal structure generation and why this is approach well-justified.

Keywords. Diffusion model, generative model, neural networks, langevin dynamics, material science

1 Introduction

1.1 Background

One of the great challenges in material sciences is the discovery of new materials with desired properties. For example, having a cheap and safe material that could easily absorb and release hydrogen might lead to a breakthrough in the energy storage applications.

In practice it is hard to discover a new material with given properties (or any material at all) and much hope is placed in generative statistical algorithms.

Traditionally, the new materials are found through laborious experimental research. In the last decades with the advent of powerful computers and the density functional theory (DFT) it became possible to estimate the stability of hypothetical crystal structures via numerical simulations. In turn, this allows to find new materials with extensive brute-force screening. This approach is however still far from efficient, as the computation is extremely slow.

More recently, various deep learning generative models, such as variational autoencoders, generative adversarial networks were adapted to this task [10, 9, 5, 8, 6, 4, 2]. The idea is to

train a DL model on a database of existing structures and use it to draw new structures from the data distribution.

There are several major hurdles for the generative models in this domain. Firstly, the available data is very limited: in the popular material database Materials Project, there are only slightly above 33 thousand stable structures as of now, which is a lot less than used by most generative models for images. Secondly, the nature of the data is very specific: the structure should ideally be represented in a way that is invariant to the rigid transformations of the space such as rotations and translations, and the periodic nature of the structure should also be taken into account.

When using a model with a latent representation, such as GAN and VAE, another particularity is that the desired (i.e. stable) structures define a discrete subset inside the latent space, the structures "in between" are not stable, which might make the sampling quite challenging.

2 Crystal representation

A crystallographic structure can be defined by its:

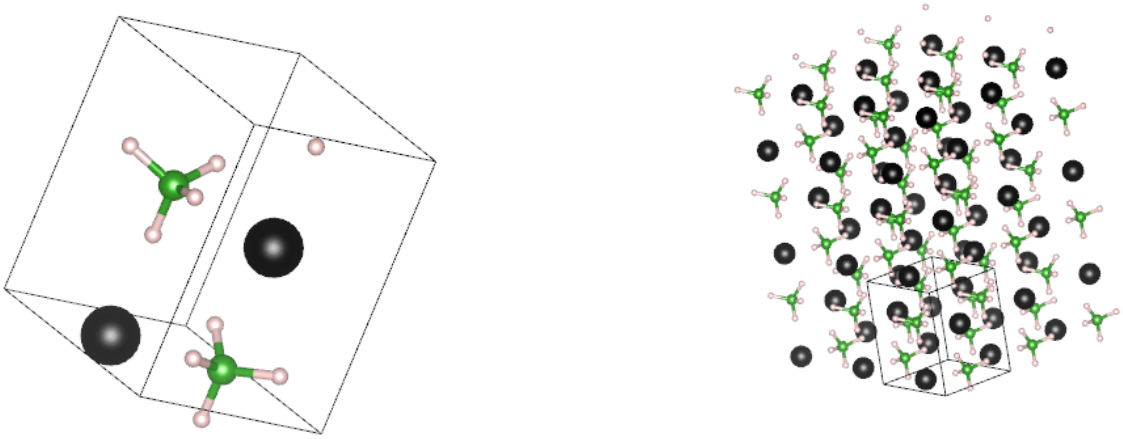
- Unit cell: the whole crystal structure is obtained by tiling the space with a unit cell (a parallelepiped). The unit cell is not unique, and is defined by three vectors in 3-dimensional space. A smallest (once again, not unique) unit cell is called a primitive cell.
- Composition: The elements that are found inside the structure and their quantities inside a unit cell (naturally, the quantity will depend on the choice of the unit cell). I.e. one can distinguish the number (the quantity) and the nature (the set of elements) of a composition.
- Positions: The coordinates of those atoms inside the unit cell. Either relative to the basis formed by the unit cell or relative to an orthonormal basis.

As noted above, if one aims at generating new stable crystal structures choosing a good representation is a crucial step. It is noted by many researchers that a representation that would be invariant to the isometries of the space containing the crystal, such as rotations and translations, might greatly help in this endeavor [5, 10].

For example, if we simply write down all the coordinates of the atoms inside a unit cell, the result will heavily depend on the choice of the origin: the same two atoms may have coordinates that are close to or vastly different from one another depending on whether they are inside the same unit cell or in the neighbouring ones.

3 The approach

Our approach is similar to the denoising diffusion probabilistic model [3]. This type of model is usually used for image generation or super-resolution. In this model, a noise is sequentially added to the image. At each following step the distribution of the pixels steadily approaches



(a) A single unit cell

(b) The same unit cell repeated 3 times in all directions

Figure 1: Illustration of a unit cell inducing the crystal structure of a Li-B-H compound. Images were obtained using VESTA software [7]

the standard Gaussian distribution. A DNN is trained to estimate the reverse process and the sampling is conducted by sequential denoising of an image sampled with a Gaussian distribution.

In our adaptation, the noise is added to the atomic positions (which define a $n \times 3$ matrix, with n being the number of atoms in the structure). The new structures are obtained by sequential denoising of atomic positions in a structure sampled randomly with a uniform distribution.

Similarly, the number of atoms, lattice cell parameters and the nature of the atoms might be estimated using this denoising approach.

While atomic positions are defined by the coordinates relative to the unit cell, the denoising network uses exclusively the relative information: distances between the atoms, and angles formed by the triplets of atoms, ensuring the invariance to rigid transformations of the space. This is achieved by translating a structure into a graph representation before passing to the network.

The denoising network architecture is based on a edge-gated graph convolutional neural network applied to two graphs corresponding to the crystal structure [1]. For each pair of atoms, the network outputs the movement of one atom relative to another, the movements of one atoms relative to all others are then summed and are treated as the predicted movement of the atom.

During the presentation we will explain the model in more detail, discuss the choices that were made and show the numeric results that could validate this approach.

References

[1] Kamal Choudhary and Brian DeCost. "Atomistic Line Graph Neural Network for improved materials property predictions". In: *npj Computational Materials* 7.1 (Nov. 2021). ISSN:

2057-3960. DOI: 10.1038/s41524-021-00650-1. URL: http://dx.doi.org/10.1038/s41524-021-00650-1.

[2] Court CJ et al. "3-D Inorganic Crystal Structure Generation and Property Prediction via Representation Learning". In: *Journal of Chemical Information and Modeling* (2020).

[3] Jonathan Ho, Ajay Jain, and Pieter Abbeel. *Denoising Diffusion Probabilistic Models*. 2020. arXiv: 2006.11239 [cs.LG].

[4] Jordan Hoffmann et al. *Data-Driven Approach to Encoding and Decoding 3-D Crystal Structures*. 2019. arXiv: 1909.00949 [cs.LG].

[5] Sungwon Kim et al. "Generative Adversarial Networks for Crystal Structure Prediction". In: *American Chemical Society* (2020).

[6] Teng Long et al. "Constrained crystals deep convolutional generative adversarial network for the inverse design of crystal structures". In: *npj Computational Materials* 7 (May 2021), p. 66. DOI: 10.1038/s41524-021-00526-4.

[7] Koichi Momma and Fujio Izumi. "*VESTA3* for three-dimensional visualization of crystal, volumetric and morphology data". In: *Journal of Applied Crystallography* 44.6 (Dec. 2011), pp. 1272–1276. DOI: 10.1107/S0021889811038970. URL: https://doi.org/10.1107/S0021889811038970.

[8] Juhwan Noh et al. "Inverse Design of Solid-State Materials via a Continuous Representation". In: *Matter* 1.5 (2019), pp. 1370–1384. ISSN: 2590-2385. DOI: https://doi.org/10.1016/j.matt.2019.08.017. URL: https://www.sciencedirect.com/science/article/pii/S2590238519301754.

[9] Asma Nouira, Nataliya Sokolovska, and Jean-Claude Crivello. *CrystalGAN: Learning to Discover Crystallographic Structures with Generative Adversarial Networks*. 2019. arXiv: 1810.11203 [cs.LG].

[10] Tian Xie et al. *Crystal Diffusion Variational Autoencoder for Periodic Material Generation*. 2021. arXiv: 2110.06197 [cs.LG].

Comment imposer des pré-activations gaussiennes dans un réseau de neurones ?

Pierre Wolinski [1] & Julyan Arbel [2]

[1] *Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP, LJK, 38000 Grenoble, France, pierre.wolinski@inria.fr*

[2] *Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP, LJK, 38000 Grenoble, France, julyan.arbel@inria.fr*

Résumé. Le but de ce travail est de proposer un moyen de modifier la loi d'initialisation des poids d'un réseau de neurones ainsi que sa fonction d'activation, de façon à assurer que toutes les pré-activations soient gaussiennes. Nous proposons une famille de couples initialisation/activation, où les fonctions d'activation couvrent un continuum allant des fonctions bornées de type Heaviside ou tanh, jusqu'à la fonction identité.

Ce travail est motivé par la contradiction entre des travaux existants quant au caractère gaussien des pré-activations : d'un côté, les travaux dans la lignée des Neural Tangent Kernels et de l'Edge of Chaos en font abondamment usage ; de l'autre, des résultats théoriques et expérimentaux mettent à mal cette hypothèse.

La famille de couples initialisation/activation que nous proposons permettra d'avancer sur la question qui traverse des travaux actuels : est-il souhaitable d'avoir des pré-activations gaussiennes dans un réseau de neurones ?

Mots-clés. réseaux de neurones, initialisation, transformée de Mellin.

Abstract. The goal of the present work is to propose a way to modify both the initialization distribution of the weights of a neural network and its activation function, such that all pre-activations are Gaussian. We propose a family of pairs initialization/activation, where the activation functions span a continuum from bounded functions (s.a. Heaviside or tanh) to the identity function.

This work is motivated by the contradiction between existing works dealing with Gaussian pre-activations: on one side, the works in the line of the Neural Tangent Kernels and the Edge of Chaos are assuming it, while on the other side, theoretical and experimental results challenge this hypothesis.

The family of pairs initialization/activation we are proposing will help us to answer this hot question: is it desirable to have Gaussian pre-activations in a neural network?

Keywords. neural networks, initialization, Mellin transform.

1 Introduction

Prenons un réseau de neurones de profondeur L à son initialisation. L'action opérée par sa l-ième couche est :

$$\forall l \in [1, L] : \quad X^{(l+1)} = \phi(Z^{(l+1)}) \tag{1}$$

$$\text{où} : \quad Z^{(l+1)} = \frac{1}{\sqrt{n_l}} \left(W^{(l)} X^{(l)} + b^{(l)} \right), \tag{2}$$

où $X^{(l+1)} \in \mathbb{R}^{n_{l+1}}$ est son *activation*, $Z^{(l+1)} \in \mathbb{R}^{n_{l+1}}$ est sa *pré-activation*, ϕ est la *fonction d'activation* agissant coordonnée par coordonnée, $W^{(l)} \in \mathbb{R}^{n_{l+1} \times n_l}$ est la *matrice de poids* de la couche, $b^{(l)} \in \mathbb{R}^{n_{l+1}}$ son *vecteur de biais*, et $X^{(l)} \in \mathbb{R}^{n_l}$ son *entrée*.

Nous nous intéressons à la loi des pré-activations $Z^{(l)}$, pour une entrée fixe $X^{(0)}$, et des poids $W^{(l)}$ et biais $b^{(l)}$ initialisés aléatoirement selon des distributions données.

Hypothèse gaussienne pour les pré-activations. Dans l'analyse théorique des propriétés des réseaux de neurones à leur initialisation, l'hypothèse d'une distribution gaussienne des pré-activations est courante. On la retrouve par exemple dans le cadre des « Neural Tangent Kernels » [4] ou de l'« Edge of Chaos » [6], deux axes de recherche ayant respectivement pour but d'étudier la dynamique d'entraînement des réseaux de neurones et la propagation de l'information au travers de ceux-ci. D'une part, cette *hypothèse gaussienne* est justifiée dans le cas des réseaux de neurones « infiniment larges » (c'est-à-dire lorsque les largeurs n_l des couches tendent vers l'infini), par application du théorème central limite (voir Eqn. (2) pour $n_l \to \infty$) [5]. D'autre part, elle est apparemment indispensable pour obtenir les résultats de [4, 6]. Cependant, l'hypothèse gaussienne reste discutée pour des raisons théoriques et pratiques.

Premièrement, d'un point de vue strictement théorique, il a été démontré que, dans le cas des réseaux de neurones de largeur finie (n_l fini), la loi de $Z^{(l)}$ tend à avoir des queues de plus en plus lourdes au fur et à mesure que l augmente, c'est-à-dire au cours de la propagation de l'information dans le réseau de neurones [9]. Deuxièmement, une série d'expériences tend à montrer que pousser la loi des pré-activations vers une gaussienne (via un *a priori* bayésien, par exemple) conduit à des performances moins bonnes qu'en la poussant vers une loi aux queues plus lourdes, comme une loi de Laplace [2].

Contributions. Le but de ce travail est de permettre de tester l'efficacité de réseaux de neurones à l'architecture *réaliste*, c'est-à-dire de largeur finie (n_l potentiellement petit) et couramment utilisée (« fully connected », ou « convolutional »), *ajustés de façon à s'assurer que toutes leurs pré-activations soient de loi gaussienne*. À cette fin, nous proposons une famille de couples $(\mathrm{P}_\theta, \phi_\theta)$, où P_θ est la loi d'initialisation des poids, ϕ_θ est la fonction d'activation, et $\theta \in]2, \infty[$ est un paramètre, tel que :

— pour une couche l donnée :

$$\left.\begin{array}{rcll} Z_i^{(l)} & \sim & \mathcal{N}(0,1) & \text{i.i.d.} \\ W_{ij}^{(l)} & \sim & \mathrm{P}_\theta & \text{i.i.d.} \\ b_i^{(l)} & = & 0 & \end{array}\right\} \Rightarrow Z_i^{(l+1)} := \frac{1}{\sqrt{n_l}}\left(W^{(l)}\phi_\theta(Z^{(l)}) + b^{(l)}\right) \sim \mathcal{N}(0,1),$$

autrement dit, les pré-activations $Z_i^{(l+1)}$ restent gaussiennes pour tout l ;

— P_θ est la distribution de Weibull symétrique $\mathcal{W}(\theta, 1)$, de fonction de répartition :

$$F_W(t) = \frac{1}{2} + \frac{1}{2}\mathrm{sgn}(t)\exp\left(-|t|^\theta\right); \tag{3}$$

— ϕ_θ définie de façon à ce que $Z_i^{(l+1)} := \frac{1}{\sqrt{n_l}}\left(W^{(l)}\phi_\theta(Z^{(l)}) + b^{(l)}\right)$ soit gaussienne ;

— cas limites :

$$\begin{array}{rcccccccc} \theta & \to & \infty & \Rightarrow & \mathrm{P}_\theta & \xrightarrow{d} & \mathcal{R} & \text{et} & \phi_\theta \xrightarrow{p.p.} \mathrm{Id}, \\ \theta & \to & 2 & \Rightarrow & \mathrm{P}_\theta & \xrightarrow{d} & \mathcal{W}(2,1) & \text{et} & \phi_\theta \xrightarrow{p.p.} \mathrm{sgn}, \end{array}$$

où $\mathcal{R}$ est la loi de Rademacher (si $\xi \sim \mathcal{R}$, alors $\mathbb{P}(\xi = \pm 1) = 1/2$) et Id est la fonction identité.

Dans le cas limite $\theta \to \infty$, on initialise les poids à ± 1, ce qui correspond à de la « quantification » binaire, utilisée dans la compression de réseaux de neurones [7], avec une fonction d'activation linéaire, régulièrement utilisée dans certaines analyses théoriques sur les réseaux de neurones [1].

Dans le cas limite $\theta \to 2$, on retrouve des poids dont la distribution possède des queues gaussiennes, avec une fonction d'activation de type « Heaviside », utilisée dès les débuts des réseaux de neurones [8].

2 Obtention de la famille $\{(\mathrm{P}_\theta, \phi_\theta) : \theta \in]2, \infty[\}$

Décomposition du problème. Pour évaluer la distribution des pré-activations à la sortie d'une couche, nous devons être capables de traiter une combinaison linéaire de produits de variables aléatoires :

$$Z^{(l+1)} = \frac{1}{\sqrt{n_l}}\left(W^{(l)}\phi(Z^{(l)}) + b^{(l)}\right). \tag{4}$$

Nous voulons que, pour des pré-activations $Z_i^{(l)} \sim \mathcal{N}(0,1)$ i.i.d., les composantes de $Z^{(l+1)}$ soient $\mathcal{N}(0,1)$. Sans perte de généralité, nous ne considérons qu'une seule composante de $Z^{(l+1)}$, nous omettons les numéros de couche, nous remplaçons $Z^{(l)}$ par X, et le biais $b^{(l)}$ par 0 (voir Sec. 3 pour une discussion) : $Z = \frac{1}{\sqrt{n}}\sum_{i=1}^n W_i\phi(X_i)$.

Comme nous voulons aboutir à $Z \sim \mathcal{N}(0,1)$, et que Z est une combinaison linéaire de v.a. $(W_i\phi(X_i))_i$ i.i.d., nous voulons construire une loi d'initialisation des W_i et une fonction ϕ de façon à ce que tous les $(W_i\phi(X_i))_i$ soient distribués selon la loi $\mathcal{N}(0,1)$.

Problème principal. Soit $X \sim \mathcal{N}(0,1)$. Trouver une loi de probabilité P et une fonction ϕ tels que : $W \sim \mathrm{P} \;\Rightarrow\; W\phi(X) \sim \mathcal{N}(0,1)$.

Nous décomposons ce problème en deux parties, en introduisant une variable aléatoire intermédiaire $Y = \phi(X)$:

— trouver des lois P et Q telles que : $W \sim \mathrm{P}, Y \sim \mathrm{Q} \;\Rightarrow\; WY =: G \sim \mathcal{N}(0,1)$;

— pour une loi Q, trouver une fonction ϕ telle que : $X \sim \mathcal{N}(0,1) \;\Rightarrow\; Y = \phi(X) \sim \mathrm{Q}$.

Multiplication de deux variables aléatoires. Rappelons que $G = WY \sim \mathcal{N}(0,1)$. Considérons les variables aléatoires $|W|$, $|Y|$ et $|G| = |W||Y|$. Notons $f_{|W|}$, $f_{|Y|}$ et $f_{|G|}$ leur densité. Sous réserve d'intégrabilité, nous pouvons utiliser la propriété suivante de la transformée de Mellin $\mathcal{M}$:

$$\mathcal{M}f_{|G|} = (\mathcal{M}f_{|W|}) \cdot (\mathcal{M}f_{|Y|}), \qquad \text{où} \quad (\mathcal{M}f)(t) = \int_0^\infty x^{t-1} f(x)\,\mathrm{d}x.$$

$$\text{D'où : } f_{|Y|}(y) := \mathcal{M}^{-1}\left[\frac{\mathcal{M}f_{|G|}}{\mathcal{M}f_{|W|}}\right](y).$$

Ainsi, par symétrie, nous pouvons obtenir f_Y en partant de $f_{|G|}$ (la densité de la valeur absolue d'une gaussienne $\mathcal{N}(0,1)$) et de $f_{|W|}$. Nous choisissons $W \sim \mathcal{W}(\theta,1)$, la loi de Weibull symétrique de paramètres $(\theta,1)$ (voir Eqn. 3). En résumé :

$$W \sim \mathrm{P}_\theta = \mathcal{W}(\theta,1), \qquad Y \sim \mathrm{Q}_\theta : f_{|Y|} = \mathcal{M}^{-1}\left[\frac{\mathcal{M}f_{|G|}}{\mathcal{M}f_{|W|}}\right](y).$$

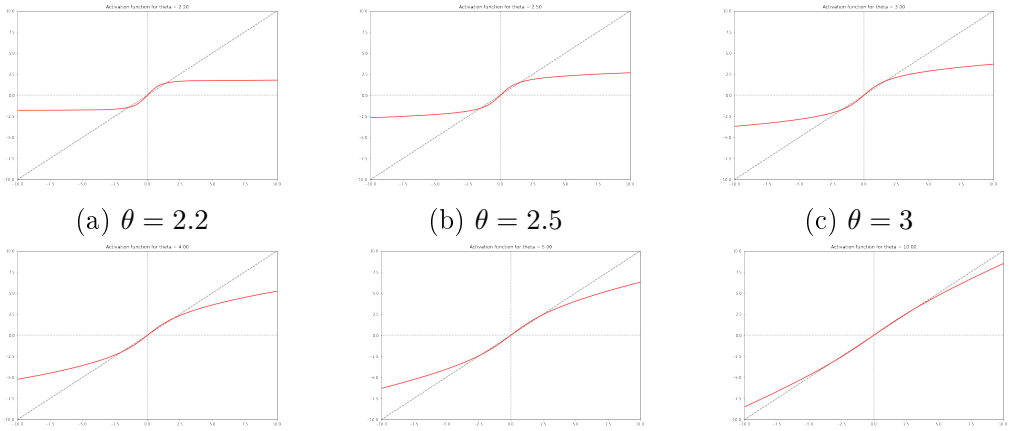
Obtention de la fonction d'activation. Pour obtenir ϕ, il suffit d'effectuer le calcul :

$$\phi_\theta(z) := F_Y^{-1}(F_G(t)), \qquad F_Y(t) := \frac{1}{2} + \frac{1}{2}\mathrm{sgn}(t)\int_0^{|t|} f_{|Y|}(y)\,\mathrm{d}y. \tag{5}$$

2.1 Résultats

Point technique. Une difficulté technique consiste à inverser la transformée de Mellin (voir Eqn. (5)). Cela est possible numériquement en utilisant la technique proposée dans [3], moyennant quelques astuces. L'inversion est instable lorsque $\theta \to 2$.

Exemples de fonctions d'activation. Ces fonctions d'activations ϕ_θ nous assurent, lorsqu'elle sont combinées avec l'initialisation des poids adéquate de type $\mathcal{W}(\theta,1)$, d'avoir des pré-activations gaussiennes. La famille des ϕ_θ constitue un continuum entre les fonctions bornées, de type Heaviside ou tanh, jusqu'au fonctions linéaires.



(a) $\theta = 2.2$ (b) $\theta = 2.5$ (c) $\theta = 3$

(d) $\theta = 4$ (e) $\theta = 5$ (f) $\theta = 10$

FIGURE 1 – Fonctions d'activation ϕ_θ pour $\theta \in \{2.2; 2.5; 3; 4; 5; 10\}$.

3 Discussion

Hypothèse de l'indépendance des composantes $Z_i^{(l)}$ des pré-activations. Dans ce travail, nous avons supposé que les composantes $Z_i^{(l)}$ des pré-activations étaient indépendantes. C'est une hypothèse simplificatrice qui est en partie justifiée par le résultat obtenu dans [5]. Dans cette variante du théorème central limite, l'échangeabilité des composantes $Z_i^{(l)}$ peut se substituer à l'hypothèse d'indépendance pour assurer le caractère gaussien des pra-activations de la couche suivante $Z^{(l+1)}$. Cependant, si nous restons dans un cadre non asymptotique, où les couches sont de largeur n_l faible, ce résultat ne s'applique pas et il est nécessaire d'évaluer l'influence des dépendances entre les composantes de $Z^{(l)}$. Nous explorerons cette piste dans la suite de ces travaux.

Initialisation des biais. Nous n'avons pas tenu compte de l'initialisation des biais, étant donné qu'ils ne font que s'ajouter à une somme qui est déjà gaussienne (voir Eqn. (4)). On peut les initialiser à 0 ou de manière gaussienne, à condition de corriger le premier terme de l'équation (4) de façon à ce que le résultat reste $\mathcal{N}(0, 1)$.

Retour aux Neural Tangent Kernels et à l'Edge of Chaos. Dans ces travaux, l'hypothèse des pré-activations gaussiennes provient de la limite infinie de la taille des couches. Il est maintenant possible de reprendre le même travail, mais avec les couples initialisation/activation que nous proposons, pour des réseaux aux couches peu larges, tout en conservant l'hypothèse gaussienne (nécessaire aux calculs).

Loi de Weibull pour l'initialisation des poids. La loi de Weibull comporte deux avantages indispensables ici : 1) elle a une densité nulle en 0, condition nécessaire pour

que le produit WY ait une densité finie en 0 lorsque $f_Y(0) > 0$; 2) un résultat connu permet de trouver le paramètre de queue de Y à partir du paramètre θ de la loi de W, de façon à ce que WY ait une queue gaussienne.

Tester des réseaux avec plusieurs couples $(\mathrm{P}_\theta, \phi_\theta)$**.** Nous pouvons maintenant tester l'hypothèse des pré-activations gaussiennes en entraînant une famille de réseaux de neurones ayant chacun leur propre couple $(\mathrm{P}_\theta, \phi_\theta)$ de loi d'initialisation et de fonction d'activation. Nous pourrons comparer leurs performances avec celles des réseaux de même architecture, mais conçus avec des lois d'initialisation et des fonctions d'activation usuelles. Le résultat de ces expériences contribuera à répondre à la question actuellement débattue : est-il souhaitable d'avoir des pré-activations gaussiennes dans un réseau de neurones ?

Références

[1] S. Arora, N. Cohen, W. Hu, and Y. Luo. Implicit regularization in deep matrix factorization. *Advances in Neural Information Processing Systems*, 32, 2019.

[2] V. Fortuin, A. Garriga-Alonso, F. Wenzel, G. Rätsch, R. Turner, M. van der Wilk, and L. Aitchison. Bayesian neural network priors revisited. *arXiv preprint arXiv:2102.06571*, 2021.

[3] B. Gabutti and L. Sacripante. Numerical inversion of the Mellin transform by accelerated series of Laguerre polynomials. *Journal of Computational and Applied Mathematics*, 34(2):191–200, 1991.

[4] A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in Neural Information Processing Systems*, 31, 2018.

[5] A. G. d. G. Matthews, J. Hron, M. Rowland, R. E. Turner, and Z. Ghahramani. Gaussian process behaviour in wide deep neural networks. In *International Conference on Learning Representations*, 2018.

[6] B. Poole, S. Lahiri, M. Raghu, J. Sohl-Dickstein, and S. Ganguli. Exponential expressivity in deep neural networks through transient chaos. *Advances in Neural Information Processing Systems*, 29, 2016.

[7] H. Pouransari, Z. Tu, and O. Tuzel. Least squares binary quantization of neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 698–699, 2020.

[8] F. Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386, 1958.

[9] M. Vladimirova, S. Girard, H. Nguyen, and J. Arbel. Sub-Weibull distributions: Generalizing sub-Gaussian and sub-exponential properties to heavier tailed distributions. *Stat*, 9(1):e318, 2020.

Processus - Statistique mathématique

Vitesses de contraction du posterior pour les processus gaussiens profonds contraints en classification et estimation de densité

François Bachoc [1] & Agnès Lagnoux [2]

[1] *Institut de Mathématiques de Toulouse; UMR5219. Université de Toulouse; CNRS. UT3, F-31062 Toulouse, France.* francois.bachoc@math.univ-toulouse.fr

[2] *Institut de Mathématiques de Toulouse; UMR5219. Université de Toulouse; CNRS. UT2J, F-31058 Toulouse, France.* lagnoux@univ-tlse2.fr

Résumé. Nous fournissons des vitesses de contraction du posterior pour les processus gaussiens contraints profonds en estimation non paramétrique de densité et en classification. Les contraintes sont des bornes sur les valeurs et dérivées des processus gaussiens dans les couches de la structure de composition. Les vitesses de contraction sont d'abord données dans un cadre général, sous la forme d'une nouvelle fonction de concentration que l'on introduit et qui prend les contraintes en compte. Ensuite, le cadre général est appliqué au mouvement Brownien intégré, au processus de Riemann Liouville et au processus de Matérn, avec des classes de fonctions standard. Dans chacun des exemples, on retrouve des vitesses minimax classiques.

Mots-clés. Prior gaussien, prior gaussien profond, inférence bayésienne, estimation non paramétrique de densité, classification, contraction du posterior, classes de fonctions régulières, fonctions de covariance de Matérn.

Abstract. We provide posterior contraction rates for constrained deep Gaussian processes in non-parametric density estimation and classification. The constraints are in the form of bounds on the values and on the derivatives of the Gaussian processes in the layers of the composition structure. The contraction rates are first given in a general framework, in terms of a new concentration function that we introduce and that takes the constraints into account. Then, the general framework is applied to integrated Brownian motions, Riemann-Liouville processes, and Matérn processes and to standard smoothness classes of functions. In each of these examples, we can recover known minimax rates.

Keywords. Gaussian priors, deep Gaussian priors, Bayesian inference, nonparametric density estimation, classification, posterior contraction, smoothness classes, Matérn covariance functions.

1 Introduction

1.1 Vitesse de concentration du posterior

On considère une fonction densité de probabilité fixée, continue et inconnue $p_0 : [-1,1]^d \to [0,\infty)$ et l'on observe des vecteurs aléatoires $X_1, \dots, X_n$ i.i.d. de densité p_0. Dans cette communication, on se limite au cas de l'estimation d'une densité, mais le manuscrit long Bachoc et Lagnoux (2021), que résume cette communication, considère aussi un problème de classification. On considère un prior Bayésien sur la densité p_0 sous la forme d'une densité de probabilité aléatoire P_0. La règle de Bayes fournit alors un posterior qui prend la forme d'une distribution de probabilité sur l'espace $\mathcal{D}$ des densités de probabilité continues sur $[-1,1]^d$. On exprime ce posterior sous la forme

$$\mathbb{P}\left(P_0 \in \cdot | \, X_1, \dots, X_n\right),$$

et on le voit comme une distribution dépendant des observations $X_1, \dots, X_n$. On note que la loi de ce posterior aléatoire dépend de la densité inconnue p_0. On s'intéresse alors à montrer une vitesse de contraction du posterior aléatoire vers la vraie densité p_0. C'est-à-dire que l'on cherche une suite $(\varepsilon_n)_{n\in\mathbb{N}}$, tendant vers zéro aussi vite que possible, telle que

$$\mathbb{P}\left(h\left(P_0, p_0\right) \geq M_n \varepsilon_n | \, X_1, \dots, X_n\right) \xrightarrow[n\to\infty]{p} 0, \tag{1}$$

pour toute suite $(M_n)_{n\in\mathbb{N}}$ qui tend vers l'infini. Ici h est la distance de Hellinger. Dans cette convergence, la probabilité porte sur les observations $X_1, \dots, X_n$ et dépend donc de p_0. Un référence générale sur les vitesses de contraction du posterior est Ghosal, Ghosh et van der Vaart (2000).

1.2 Processus Gaussiens

On considère ici que la densité aléatoire P_0 du prior Bayésien s'écrit, pour $x \in [-1,1]^d$,

$$P_0(x) = \frac{e^{Z(x)}}{\int_{[-1,1]^d} e^{Z(t)} dt},$$

où $Z : [-1,1]^d \to \mathbb{R}$ est un processus gaussien continu. Notons $w_0 = \log(p_0)$. Notons $\mathbb{H}_Z$ l'espace de Hilbert à noyau reproduisant (RKHS) de Z, avec la norme associée $||\cdot||_{\mathbb{H}_Z}$. Dans van der Vaart et van Zanten (2008), il est introduit la fonction $\phi_{w_0} : (0,\infty) \to [0,\infty)$ donnée par, pour $\varepsilon > 0$,

$$\phi_{w_0}(\varepsilon) = \inf_{\substack{h\in\mathbb{H}_Z \\ ||h-w_0||_\infty < \varepsilon}} ||h||^2_{\mathbb{H}_Z} + \mathbb{P}\left(||Z||_\infty < \varepsilon\right). \tag{2}$$

Cette fonction est interprétée comme mesurant la concentration du processus gaussien Z autour de la fonction w_0 et est donc appelée fonction de concentration. Dans van der

Vaart et van Zanten (2008), il est montré qu'il y aura contraction du posterior à vitesse ε_n, au sens de (1), for toute suite ε_n satisfaisant

$$\phi_{w_0}(\varepsilon_n) \leq n\varepsilon_n^2.$$

Ainsi on voit que la fonction de concentration ϕ_{w_0} mesure la vitesse de contraction : plus cette fonction tend vite vers l'infini en zéro, plus la vitesse de contraction sera lente. De plus, on voit que ϕ_{w_0} est plus petite lorsque w_0 appartient au RKHS $\mathbb{H}_Z$, ce qui montre que l'on a une vitesse plus rapide lorsque la densité inconnue est en accord avec le prior (en terme de régularité). Le second terme dans (2) est plus petit lorsque le processus Z est régulier, ce qui laisse voir que la vitesse de contraction sera plus rapide pour une fonction régulière et un prior régulier.

1.3 Processus gaussiens profonds

Les processus gaussiens profonds sont fournis par des compositions de processus gaussiens et on étés proposés initialement par Damianou et Lawrence (2013). Ici, on va considérer des processus profonds de la forme suivante. On prend $H \in \mathbb{N}^* \setminus \{1\}$ et $d_1 = d$, d_2,...,$d_H \in \mathbb{N}^*$, et $d_{H+1} = 1$. Pour $h = 1, \ldots, H$, on considère un processus gaussien centré multivarié $Z_h = (Z_{h,1}, \ldots Z_{h,d_{h+1}}) : \mathbb{R}^{d_h} \to \mathbb{R}^{d_{h+1}}$. On suppose que $Z_1, \ldots, Z_H$ sont indépendants, à composantes indépendantes et continues. On considère alors un processus gaussien profond donné par

$$Z_H \circ \cdots \circ Z_1,$$

qui fournit un prior (non Gaussien) pour les fonctions de $[-1, 1]^d$ dans $\mathbb{R}$. Les processus profonds peuvent être vus comme des réseaux de neurones profonds dans lesquels les fonctions d'activation ont un prior Bayésien donné par des processus gaussiens, voir la Figure 1. Les processus gaussiens profonds ont peu de garanties théoriques à notre connaissance, à l'exception de Bachoc et Lagnoux (2021) et de Finocchio et Schmidt-Hieber (2021).

2 Vitesses de contraction générales

2.1 Prior avec contraintes

Pour les preuves que nous fournissons dans Bachoc et Lagnoux (2021), nous devons conditionner les processus $Z_1, \ldots, Z_H$ par des contraintes sur les valeurs, pour $h = 1, \ldots, H-1$, $i = 1, \ldots, d_{h+1}$,

$$||Z_{h,i}||_\infty \leqslant 1, \tag{3}$$

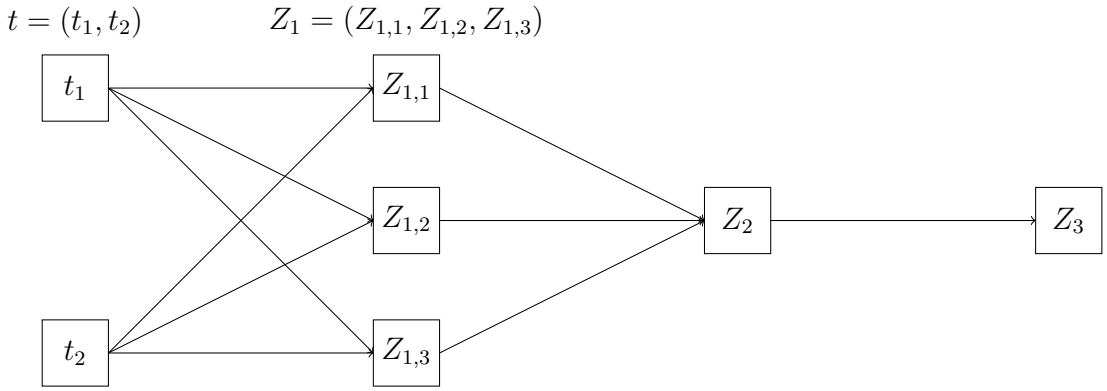


Figure 1: Exemple d'un processus gaussien profond de $[-1,1]^2$ dans $\mathbb{R}$. Ici $H = 3$, $d_1 = d = 2$, $d_2 = 3$, $d_3 = 1$, et $d_{H+1} = d_4 = 1$.

et des contraintes sur les dérivées, pour $h = 2, \ldots, H$, pour $i = 1, \ldots, d_{h+1}$, et pour $j = 1, \ldots, d_h$,

$$\left|\left|\frac{\partial Z_{h,i}}{\partial x_j}\right|\right|_\infty \leqslant K_{h,i,j}, \tag{4}$$

avec des constantes $K_{h,i,j} > 0$. Dans (3) et (4), les normes infinies sont prises sur le domaine $[-1,1]^d$. On utilisera alors le prior gaussien profond $Z_{c,H} \circ \cdots \circ Z_{c,1}$, où $Z_{c,1}, \ldots, Z_{c,H}$ sont des processus qui suivent la loi de $Z_1, \ldots, Z_H$ conditionnée par (3) et (4). Le prior utilisé sur p_0 est finalement donné par, pour $x \in [-1,1]^d$,

$$P_0(x) = \frac{e^{Z_{c,H} \circ \cdots \circ Z_{c,1}(x)}}{\int_{[-1,1]^d} e^{Z_{c,H} \circ \cdots \circ Z_{c,1}(t)} dt}. \tag{5}$$

On note que les contraintes (3) (resp. (4)) sont satisfaites avec probabilité non nulle pour tout les processus gaussiens à trajectoires continues (resp. continuement différentiables), ce qui autorise la plupart des fonctions de covariance continues (resp. dérivables) usuelles.

2.2 Fonction de concentration adaptée à la composition et aux contraintes

Pour $h = 1, \ldots, H$, $i = 1, \ldots, d_{h+1}$, on note $\mathbb{H}_{h,i}$ le RKHS de $Z_{h,i}$ avec la norme $||\cdot||_{\mathbb{H}_{h,i}}$. On écrit également la log densité $w_0 = \log(p_0)$ avec la même structure de composition que $\log(P_0)$: $w_0 = z_{0,H} \circ \cdots \circ z_{0,1}$ avec, pour $h = 1, \ldots, H$, $z_{0,h} = (z_{0,h,1}, \ldots, z_{0,h,d_{h+1}})$. On suppose que les fonctions $z_{0,1}, \ldots, z_{0,H}$ satisfont les mêmes contraintes que $Z_{c,1}, \ldots, Z_{c,H}$ ci-dessus et on remarque que Bachoc et Lagnoux (2021) argumentent qu'il s'agit en fait d'une hypothèse peu restrictive. Nous introduisons alors la fonction de concentration

$\Phi_{c,z_0} : (0,\infty) \to [0,\infty)$ suivante, prenant en compte la composition et les contraintes. On écrit $\mathcal{I} = \{(h,i);\, h \in \{1,\dots,H\}, i \in \{1,\dots,d_{h+1}\}\}$. Pour $\varepsilon > 0$, on définit

$$\begin{aligned}\Phi_{c,z_0}(\varepsilon) = &\sum_{i=1}^{d_2} \left(\frac{3}{2} \inf_{\substack{g \in \mathbb{H}_{1,i} \\ \|g - z_{0,1,i}\|_\infty < \varepsilon}} \|g\|^2_{\mathbb{H}_{1,i}} - 2\log \mathbb{P}\left(\|Z_{1,i}\|_\infty < \varepsilon\right) \right) \\ &+ \sum_{\substack{(h,i)\in\mathcal{I} \\ h \geqslant 2}} \left(\frac{3}{2} \inf_{\substack{g \in \mathbb{H}_{h,i} \\ \|g - z_{0,h,i}\|_\infty < \frac{\varepsilon}{2} \\ \|\partial g/\partial x_j - \partial z_{0,h,i}/\partial x_j\|_\infty < \frac{K_{\min}}{4}, \\ j=1,\dots,d_h}} \|g\|^2_{\mathbb{H}_{h,i}} \right. \\ &\left. - 2\log \mathbb{P}\Big(\|Z_{h,i}\|_\infty \leqslant \frac{\varepsilon}{2}\Big) - 2\sum_{j=1}^{d_h} \log \mathbb{P}\Big(\|\partial Z_{h,i}/\partial x_j\|_\infty \leqslant \frac{K_{\min}}{4}\Big) \right),\end{aligned} \tag{6}$$

avec

$$K_{\min} = \min_{h=2,\dots,H} \min_{i=1,\dots,d_{h+1}} \min_{j=1,\dots,d_h} K_{h,i,j}.$$

2.3 Vitesses de contraction

Nous montrons alors le résultat de contraction général suivant.

Théorème 1 *Considérons une suite $(\varepsilon_n)_{n\in\mathbb{N}}$ satisfaisant $\Phi_{c,z_0}(\varepsilon_n) \leq n\varepsilon_n^2$. Alors, le prior par processus Gaussien profond contraint* (5) *présente une contraction du posterior à vitesse ε_n comme dans* (1).

Un avantage de ce résultat est que les termes de la fonction de concentration (6), correspondants à chaque fonction $z_{0,h,i}$, ont généralement le même ordre de grandeur que dans (2) dans le cas non contraint. De plus, on est libre de choisir la décomposition de $\log(p_0)$ qui fournit la plus petite fonction Φ_{c,z_0}.

3 Exemple du prior de Matérn

On considère que la densité p_0 appartient à une classe de régularité $\mathcal{F}^\beta([-1,1]^d,\mathbb{R}) \cap \mathcal{H}^\beta([-1,1]^d,\mathbb{R})$, ou $\beta > 0$ est le paramètre de régularité, ce qui correspond à avoir $\lfloor\beta\rfloor$ dérivées qui sont $\beta - \lfloor\beta\rfloor$ Hölder, avec une autre condition similaire portant sur la transformée de Fourier.

On prend alors pour chaque processus gaussien $Z_{h,i}$ une fonction de covariance de Matérn de paramètre de régularité $\alpha_{h,i} \geq \beta$. On a alors la vitesse de contraction du posterior suivante.

Théorème 2 *On a une contraction du posterior comme dans* (1) *à vitesse*

$$n^{-\beta/(2\alpha_{1,\min}+d)},$$

où $\alpha_{1,\min} = \min(\alpha_{1,1}, \ldots, \alpha_{1,d_2})$.

Lorsque $\alpha_{1,\min} = \beta$ on obtient la vitesse minimax $n^{-\beta/\beta+d}$, ce qui corrrespond à un prior bien adapté.

4 Idées des preuves

L'idée de la preuve du Théorème 1 est d'utiliser les résultats de van der Vaart et van Zanten (2008) qui montrent que la vitesse contraction est régie par ϕ_{w_0} dans (2). Ces résultats peuvent en fait s'appliquer à n'importe quel processus gaussien défini sur un compact. On voit alors les composants du processus profond $(Z_{h,i}; (h,i) \in \mathcal{I})$ comme un unique processus gaussien W indexé par le compact $\mathcal{X} = [-1,1]^{d_{\max}} \times \mathcal{I}$, avec $d_{\max} = \max(d_1, \ldots, d_H)$. On construit son RKHS $\mathbb{H}$ et on introduit sa fonction de concentration ϕ_{c,w_0}, comme dans (2) autour d'une fonction $w_0 : \mathcal{X} \to \mathbb{R}$. On montre alors que pour $\varepsilon > 0$

$$\phi_{c,w_0}(\varepsilon) \leq \Phi_{c,w_0}(\varepsilon),$$

où Φ_{c,w_0} est comme dans (6). On peut alors conclure en utilisant les mêmes techniques que dans van der Vaart et van Zanten (2008).

La preuve du Théorème 2 consiste à contrôler les termes dans (6), avec des méthodes similaires à van der Vaart et van Zanten (2008).

Bibliographie

Bachoc, F. and Lagnoux, A. (2021). Posterior contraction rates for constrained deep Gaussian processes in density estimation and classication, *Posterior contraction rates for constrained deep Gaussian processes in density estimation and classication*.

Damianou, A. and Lawrence, N. (2013). Deep Gaussian Processes, *Proceedings of the Sixteenth International Workshop on Artificial Intelligence and Statistics (AISTATS)*, JMLR W&CP 31, pp. 207–215.

Finocchio, G. and Schmidt-Hieber, J. (2021). Posterior contraction for deep Gaussian process priors, *arXiv:2105.07410*s.

Ghosal, S., Ghosh, J. K. and van der Vaart, A. W. (2000). Convergence rates of posterior distributions, *Annals of Statistics*, 28, pp. 500–531.

van der Vaart, A. W. and van Zanten, J. H. (2008). Rates of contraction of posterior distributions based on Gaussian process priors, *Annals of Statistics*, 36, pp. 1435–1463.

Sélection de variables en régression SIR par seuillage doux ou seuillage dur de la matrice d'intérêt

Hadrien Lorenzo [1,2,3] & Jérôme Saracco [1,2,4]

[1] *ASTRAL, Inria, 200 Avenue de la Vieille Tour, 33405 Talence, France*
[2] *Institut de mathématiques de Bordeaux, UMR 5251 CNRS, 33400 Talence, France*
[3] *Corporate Research Advanced Data Analytics, Sartorius, Zone Industrielle les Paluds, Avenue de Jouques CS 71058, 13781 Aubagne Cedex, France*
[4] *ENSC - Bordeaux INP, 33400 Talence, France*

Résumé. La régression inverse par tranches (Sliced Inverse Regression, SIR, en anglish) a ceci d'intéressant qu'elle permet la construction d'indices, comme combinaisons linéaires de la variable explicative multidimensionnelle, les plus associés (en un certain sens) à la variable réponse étudiée. Ce type de méthode est fort utile dans des contextes non linéaires. La problématique de la sélection de variables dans ce cadre est importante. Nous présentons dans cette communication deux approches fondées sur le seuillage doux ou le seuillage dur de la matrice d'intérêt de la méthode SIR. Nous indiquons comment sélectionner l'hyper-paramètre du seuillage considéré et nous présentons les performances numériques de ces nouvelles méthodologies obtenues dans le cadre de simulations.

Mots-clés. Modèle SIR, Réduction de dimension, Sélection de variables, Seuillage dur, Seuillage doux.

Abstract. An interesting feature of Sliced Inverse Rregression (SIR) is that it allows the construction of indices, as linear combinations of the multidimensional explanatory variable, most associated (in some sense) with the response variable under study. This type of method is very useful in non-linear regression contexts. The problem of variable selection in this context is important. In this paper we present two approaches based on soft thresholding or hard thresholding of the SIR interest matrix. We show how to select the hyper-parameter of the thresholding considered and we present the numerical performances of these new methodologies obtained in simulations.

Keywords. SIR model, Dimension reduction, Variable selection, Soft thresholding, Hard thresholding.

1 Introduction

Soient une variable aléatoire réelle Y et une variable aléatoire p-dimensionnelle $\mathbf{X} \in \mathbb{R}^p$ telle que $\boldsymbol{\mu} = \mathbb{E}(\mathbf{X})$ et $\boldsymbol{\Sigma} = \mathbb{V}(\mathbf{X})$. Soit $\boldsymbol{\beta}$ une direction de $\mathbb{R}^p$.

Dans cette communication, nous allons nous focaliser sur le modèle de régression semiparamétrique suivant :

$$Y = g(\boldsymbol{\beta}'\mathbf{X}, \varepsilon), \tag{1}$$

le terme d'erreur aléatoire ε est supposé être indépendant de $\mathbf{X}$. La fonction g est une fonction de lien arbitraire, ce qui fait du modèle précédent une méthode de réduction de dimension attirante pour l'étude de la liaison entre $\mathbf{X}$ et Y, via l'indice réel $\boldsymbol{\beta}'\mathbf{X}$. Il est courant de présenter le modèle sous la forme :

$$Y \perp \mathbf{X} \mid \boldsymbol{\beta}'\,\mathbf{X}, \tag{2}$$

ce qui signifie que Y est indépendant de $\mathbf{X}$ conditionnellement à $\boldsymbol{\beta}'\,\mathbf{X}$. Ainsi, il est possible de remplacer $\mathbf{X} \in \mathbb{R}^p$ par l'indice $\boldsymbol{\beta}'\mathbf{X} \in \mathbb{R}$ sans perte d'information dans la régression de Y sur $\mathbf{X}$. On a donc fait de la réduction de dimension de la partie explicative (en passant de la dimension p à la dimension 1), sans spécifier ni le lien entre l'indice $\boldsymbol{\beta}'\,\mathbf{X}$ et Y, ni la distribution du terme d'erreur ε.

Cependant, dans le cadre de ce modèle, seule la direction de $\boldsymbol{\beta}$ est identifiable. On définit alors l'espace EDR E (pour *effective dimension reduction* en anglais), comme le sous-espace linéaire de $\mathbb{R}^p$ engendré par $\boldsymbol{\beta}$. Tout vecteur appartenant à E est alors une direction EDR.

Dans la suite, nous verrons qu'il est possible d'estimer une direction EDR sans avoir à estimer la fonction de lien g, ce qui est fort appréciable. Dans la section 2, nous introduirons la version classique de la méthode SIR, pour *sliced inverse regression* en anglais, et régression inverse par tranches (ou par tranchage) en français, qui permet d'obtenir l'estimation d'une direction EDR. Puis à la section 3, nous proposerons des versions de la méthode SIR, basées sur des seuillages doux ou dur de la matrice d'intérêt (que l'on définira), permettant de faire de la sélection de variables en basculant à zéro des composantes du vecteur $\boldsymbol{\beta}$.

Ces différentes approches seront alors comparées sur des simulations numériques, montrant l'intérêt de faire de la sélection de variables lorsque $X \in \mathbb{R}^p$ est en grande dimension (tout en ayant ici une taille d'échantillon n supérieure à p afin d'avoir une estimation non dégénérée de la matrice de variances-covariances $\boldsymbol{\Sigma}$).

2 Présentation de la méthode SIR classique

Afin d'estimer l'espace EDR, plusieurs méthodes de type SIR ont été proposées dans la littérature. Théoriquement la méthode SIR nécessite une condition sur la distribution de la variable explicative $\mathbf{X}$ appelée condition de linéarité (CL) :

$$\forall b \in \mathbb{R}^p, \mathbb{E}\left[b'\mathbf{X} \mid \boldsymbol{\beta}'\mathbf{X}\right] \text{ est linéaire en } \mathbf{X}'\boldsymbol{\beta}. \tag{3}$$

Vu que $\boldsymbol{\beta}$ est inconnu, il n'est pas possible de vérifier a priori cette hypothèse (CL) en pratique. Cependant, cette hypothèse est valide lorsque la distribution de $\mathbf{X}$ est elliptique, par exemple, c'est le cas pour la loi normale multidimensionnelle.

Considérons maintenant une transformation monotone T. Sous le modèle (1) et la condition (CL), [1] ont montré que la courbe de régression inverse centrée vérifie :

$$\mathbb{E}[\mathbf{X} \mid T(Y)] - \boldsymbol{\mu} \in \text{Span}(\boldsymbol{\Sigma}\beta). \tag{4}$$

Ainsi le sous-espace engendré par cette courbe, $\{\mathbb{E}[x \mid T(Y)] - \mathbb{E}[x] \ : \ Y \in \mathcal{Y}\}$ où $\mathcal{Y}$ est le support de la variable à expliquer Y, appartient à l'espace EDR. À partir de (4), la courbe de régression inverse centrée peut clairement être utilisée pour retrouver l'espace EDR. En effet, une conséquence directe de ce résultat est que la matrice de variances-covariances de cette courbe,

$$\Gamma = \mathbb{V}(\mathbb{E}[\mathbf{X} \mid T(Y)]),$$

est dégénérée dans toute direction Σ-orthogonale à $\boldsymbol{\beta}$. En conséquence, le vecteur propre associé à la valeur propre non nulle de la matrice d'intérêt $\Sigma^{-1}\Gamma$ est une direction EDR, ce qui signifie qu'elle engendre l'espace EDR E.

Dans l'étape de tranchage de la méthode SIR, le support de Y est découpé en H tranches $\{s_1, \ldots, s_H\}$. Avec une fonction T correspondant à ce tranchage, la matrice Γ peut alors s'écrire simplement sous la forme

$$\Gamma = \sum_{h=1}^{H} p_h (m_h - \mu)(m_h - \mu)'$$

où $p_h = P(Y \in s_h)$ et $m_h = \mathbb{E}[\mathbf{X} \mid Y \in s_h]$. Cette matrice est ainsi facilement estimable.

Considérons un échantillon $\{(\mathbf{X}_i, Y_i), i = 1, \ldots, n\}$ généré à partir du modèle (1). En substituant les moments empiriques aux moments empiriques $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, p_h et m_h, un estimateur d'une base de l'espace of E est obtenu via le vecteur propre $\hat{b}_{\text{SIR}}$ associé à la plus grande valeur propre de l'estimateur $\widehat{\Sigma}_n^{-1}\widehat{\Gamma}_n$ de $\Sigma^{-1}\Gamma$ où

$$\widehat{\boldsymbol{\Sigma}}_n = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{X}_i - \bar{X}_n)(\mathbf{X}_i - \bar{X}_n)' \ \text{ and } \ \widehat{\Gamma}_n = \sum_{h=1}^{H} \hat{p}_{h,n}(\hat{m}_{h,n} - \bar{X}_n)(\hat{m}_{h,n} - \bar{X}_n)',$$

avec

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i, \quad \hat{p}_{h,n} = \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}[Y_i \in s_h] = \frac{\hat{n}_{h,n}}{n}, \quad \hat{m}_{h,n} = \frac{1}{\hat{n}_{h,n}}\sum_{i \in s_h}\mathbf{X}_i,$$

la notation $\mathbb{I}[.]$ correspondant à la fonction indicatrice. La méthode SIR a été étudiée par de nombreux auteurs dans la littérature.

3 Approches de sélection de variables par seuillage doux ou seuillage dur

Plusieurs méthodes de sélection de variables ont été proposées en régression SIR dans la littérature. Elles peuvent être classées en 2 types : citons par exemple,

- celles fondées sur un problème d'optimisation comme Shrinkage-SIR [2], Sparse Ridge-SIR [3], Dantzig-SIR [4], Refined sparse SIR [5],
- celles fondées sur une approche purement computationelle telles que celle reposant sur la mesure de l'importance des variables [6], ou bien encore celle reposant sur la génération et l'analyse de sous-modèles (CSS pour *closest submodel selection* en anglais) [7].

Dans le cadre de cette communication, nous allons considérer deux nouvelles approches computationnelles utilisant le seuillage doux ou le seuillage dur de la matrice d'intérêt. Notons qu'une approche par seuillage, appelée *Diagonal Thresholding* (DT-SIR, pour seuillage de la diagonale en anglais), a déjà été proposée par [8] pour le cas $p \gg n$.

Tout d'abord, définissons les opérateurs de seuillage doux (S comme *soft* en anglais) et de seuillage dur (H comme *hard* en anglais) qui s'appliquent à toute matrice M de dimensions $p \times p$. Pour tout réel $\lambda \geq 0$ et $(i,j) \in \{1,\ldots,p\} \times \{1,\ldots,p\}$,

$$S_\lambda(M)_{i,j} = signe(M_{i,j}) \times \begin{cases} |M_{i,j}| - \lambda & \text{si } |M_{i,j}| - \lambda > 0, \\ 0 & \text{sinon.} \end{cases} \tag{5}$$

$$H_\lambda(M)_{i,j} = \begin{cases} M_{i,j} & \text{si } |M_{i,j}| - \lambda > 0, \\ 0 & \text{sinon.} \end{cases} \tag{6}$$

où *signe* renvoie le signe de l'élément considéré (par convention, le signe d'un nombre réel est 1, 0 ou -1 si la valeur observée est positive strictement, nulle ou négative strictement).

Nous proposons d'évaluer numériquement les performances des deux méthodes suivantes :

- *ST-SIR* pour seuillage doux de la matrice d'intérêt de l'approche SIR, qui considère la matrice $S_\lambda(\widehat{\Sigma}_n^{-1}\widehat{\Gamma}_n)$,
- *HT-SIR* pour seuillage dur de la matrice d'intérêt de l'approche SIR, qui considère la matrice $H_\lambda(\widehat{\Sigma}_n^{-1}\widehat{\Gamma}_n)$.

Pour chacune des approches, un vecteur propre associé à la plus grande valeur propre de $S_\lambda(\widehat{\Sigma}_n^{-1}\widehat{\Gamma}_n)$ ou de $H_\lambda(\widehat{\Sigma}_n^{-1}\widehat{\Gamma}_n)$ est calculé, pour une valeur de λ donnée.

Le choix de l'hyper-paramètre λ de seuillage est fondamental, car il doit non seulement permettre de faire de la sélection de variables (en basculant à zéro des composantes du vecteur propre considéré) sans pour autant trop déformer la direction estimée avec les données d'origine. Une méthodologie fournissant une valeur de λ pertinente sera décrite.

Une fois les p^* variables pertinentes/utiles identifiées, un nouveau modèle SIR peut alors être estimé sur la base de ces p^* variables sélectionnées.

Des simulations illustreront le comportement numérique des approches proposées pour

- différentes valeurs de n et de p,
- différentes structures de données (choix de $\mathbf{\Sigma}$, choix de g, choix de la distribution de ϵ),
- et différentes valeurs du nombre p^* de variables pertinentes, c'est-à-dire associées à un coefficient β_k non nul du paramètre $\boldsymbol{\beta}$.

References

[1] Duan, N. & Li, K. Slicing regression: a link-free regression method. *The Annals Of Statistics*. **19** pp. 505-530 (1991)

[2] Ni, L., Cook, R. & Tsai, C. A note on shrinkage sliced inverse regression. *Biometrika*. **92**, 242-247 (2005)

[3] Li, L. & Yin, X. Sliced inverse regression with regularizations. *Biometrics*. **64**, 124-131 (2008)

[4] Yu, Z., Zhu, L., Peng, H. & Zhu, L. Dimension reduction and predictor selection in semiparametric models. *Biometrika*. **100**, 641-654 (2013,4), https://doi.org/10.1093/biomet/ast005

[5] Tan, K., Shi, L. & Yu, Z. Sparse SIR: Optimal rates and adaptive estimation. *The Annals Of Statistics*. **48**, 64 - 85 (2020), https://doi.org/10.1214/18-AOS1791

[6] Jlassi, I. & Saracco, J. Variable importance assessment in sliced inverse regression for variable selection. *Communications In Statistics - Simulation And Computation*. **48**, 169-199 (2019),

[7] Coudret, R., Liquet, B. & Saracco, J. Comparison of sliced inverse regression approaches for underdetermined cases. *Journal De La Société Française De Statistique*. **155**, 72-96 (2014)

[8] Lin, Q., Zhao, Z. & Liu, J. Sparse sliced inverse regression via lasso. *Journal Of The American Statistical Association*. **114**, 1726-1739 (2019)

Liens entre les décompositions de Hoeffding–Sobol et de Möbius

Cécile Mercadier [1] & Olivier Roustant [2] & Christian Genest [3]

[1] *Université de Lyon, Université Claude-Bernard Lyon 1, Institut Camille-Jordan, UMR CNRS 5208, 43, boulevard du 11 novembre 1918, 69622 Villeurbanne*
mercadier@math.univ-lyon1.fr
[2] *Institut de Mathématiques de Toulouse, Université de Toulouse, INSA, 31077 cedex 4 Toulouse*
roustant@insa-toulouse.fr
[3] *Department of Mathematics and Statistics, McGill University, 805, rue Sherbrooke ouest, Montréal (Québec) Canada H3A 0B9*
christian.genest@mcgill.ca

Résumé.
Cet exposé présente les résultats obtenus dans *Linking the Hoeffding–Sobol and Möbius formulas through a decomposition of Kuo, Sloan, Wasilkowski, and Woźniakowski,* publié récemment dans *Statistics & Probability Letters*. Il s'agit de comprendre les liens entre les décompositions fonctionnelles de Hoeffding–Sobol et de Möbius. Toutes deux sont utilisées en statistique et permettent d'écrire une fonction comme la somme de termes de complexité croissante. Nous montrons qu'elles sont étroitement liées à celle obtenue en 2010 par Kuo, Sloan, Wasilkowski et Woźniakowski. Afin de les retrouver dans un énoncé commun, nous avons dû généraliser le résultat de Kuo et coll. (*Math. Comp.*, 79 (2010), 953–966).

Mots-clés. Décomposition fonctionnelle, de Hoeffding–Sobol, de Möbius.

Abstract. This talk presents the results obtained in *Linking the Hoeffding–Sobol and Möbius formulas through a decomposition of Kuo, Sloan, Wasilkowski, and Woźniakowski,* recently published in *Statistics & Probability Letters*. The goal is to understand the links between the functional decompositions of Hoeffding–Sobol and of Möbius. Both are used in statistics and make it possible to write a function as the sum of terms of increasing complexity. We prove that they are closely related to the one obtained by Kuo, Sloan, Wasilkowski, and Woźniakowski in 2010. In order to derive them from a common statement, we had to generalize the result of Kuo et al. (*Math.Comp.*, 79 (2010), 953–966).

Keywords. Functional decomposition, Hoeffding–Sobol, Möbius.

1 Les décompositions de Hoeffding–Sobol et de Möbius

Diverses constructions permettent d'exprimer une fonction f à valeurs réelles définie sur un domaine $D \subseteq \mathbb{R}^d$ pour $d \geq 2$ comme

$$f = \sum_{u \subseteq \{1,\ldots,d\}} f_u \tag{1}$$

où l'élement f_u, associé au sous-ensemble $u \subseteq \{1,\ldots,d\}$, resterait à préciser bien sûr. Nous nous intéressons à deux de ces formules, la décomposition de Hoeffding–Sobol et celle dite de Möbius. Utilisées en statistique dans des domaines différents, elles n'avaient jusqu'ici jamais été citées conjointement dans un papier. Commençons par rappeler le contexte statistique de l'une et de l'autre, ainsi que leurs définitions respectives.

1.1 Sur fond d'analyse de sensibilité

Soit $f : [0,1]^d \to \mathbb{R}$ une fonction de $L^2([0,1]^d, \lambda)$ pour $\lambda = \lambda_1 \times \cdots \times \lambda_d$ un produit de mesures de probabilité sur $[0,1]$. Une manière de comprendre la structure de f est de la décomposer en une somme de fonctions de complexités croissantes. C'est précisément ce que permet l'analyse fonctionnelle de la variance, FANOVA. Elle repose sur la décomposition de Hoeffding–Sobol qui décrit f comme (1) avec

$$f_u(\mathbf{x}) = \sum_{v \subseteq u} (-1)^{|u \setminus v|} \int f(\mathbf{x}) d\lambda_{-v}(\mathbf{x}) \tag{2}$$

lorsque $\mathbf{x} \in [0,1]^d$, $d\lambda_u(\mathbf{x}) = \prod_{i \in u} d\lambda_i(x_i)$ et $-v = \{1,\ldots,d\} \setminus v$. Le terme f_u ne dépend que des composantes de $\mathbf{x}$ associées à u. Le terme constant $f_\emptyset$ vaut $\int f d\lambda$ alors que la variance globale s'obtient comme $\sigma^2 = \int (f - f_\emptyset)^2 d\lambda$. En posant $\sigma_u^2 = \int f_u^2 d\lambda_u$ et $\sigma_\emptyset^2 = 0$, et en utilisant les arguments d'orthogonalité de [Efron and Stein, 1981], on obtient

$$\sigma^2 = \sum_{u \subseteq \{1,\ldots,d\}} \sigma_u^2 \,,$$

qui est à l'origine de plusieurs indices d'intérêt en analyse de sensibilité, comme par exemple les indices de [Sobol', 1993]. Depuis ces travaux fondateurs, la communauté scientifique n'a jamais cessé d'investiguer et reste soucieuse de proposer des outils performants pour son utilisation industrielle ; on pourra consulter par exemple [Iooss and Prieur, 2019, Roustant et al., 2020] pour citer deux publications récentes. À noter que cette décomposition a été récemment appliquée dans [Mercadier and Roustant, 2019] à la *stable tail dependence function*, aussi appelée fonction caudale, qui mesure la dépendance asymptotique en théorie des valeurs extrêmes multivariées. En résulte la définition d'indices permettant de visualiser la dépendance caudale asymptotique sur un graphe nommé le *tail dependograph*.

1.2 A l'origine des tests d'indépendance sur les copules

Considérons maintenant la classe des copules d-variées, fonctions agissant sur $[0,1]^d$, à valeurs dans $[0,1]$ et décrivant les possibles structures de dépendance d'un vecteur aléatoire. La décomposition de Möbius a été introduite par [Deheuvels, 1981] pour tester l'indépendance mutuelle dans un vecteur de variables aléatoires continues.

Pour $v \subseteq \{1,\dots,d\}$ et un vecteur $\mathbf{x} = (x_1,\dots,x_d) \in [0,1]^d$, soit $\mathbf{x}_{v,\mathbf{1}}$ le vecteur $(t_1,\dots,t_d)$ tel que $t_i = x_i$ si $i \in v$ et $t_i = 1$ sinon. La décomposition de Möbius de la copule C est alors donnée, pour tout $\mathbf{x} \in [0,1]^d$, par

$$C(\mathbf{x}) = \Pi(\mathbf{x}) + \sum_{u \subseteq \{1,\dots,d\}, |u| \geq 2} \mathcal{M}_u(C)(\mathbf{x}) \times \prod_{k \notin u} x_k,$$

avec

$$\mathcal{M}_u(C)(\mathbf{x}) = \sum_{v \subseteq u} (-1)^{|u \setminus v|} C(\mathbf{x}_{v,\mathbf{1}}) \times \prod_{k \in u \setminus v} x_k.$$

On peut calquer la présentation sur celle de la décomposition précédente en écrivant

$$C(\mathbf{x}) = \sum_{u \subseteq \{1,\dots,d\}} C_u(\mathbf{x})$$

pour

$$C_u(\mathbf{x}) = \mathcal{M}_u(C)(\mathbf{x}) \times \prod_{k \notin u} x_k \ . \tag{3}$$

Ce champ d'investigation reste d'actualité, comme en témoigne par exemple le travail [Genest et al., 2019] et la lignée bibliographique le reliant au papier pionnier de Deheuvels.

2 Nos questions

L'objectif de notre travail [Mercadier et al., 2022] était de comprendre les points communs et les différences entre les deux décompositions citées. Malgré leurs similitudes, la décomposition de Möbius est peu connue et reste isolée dans la litérature pour son utilisation dans les tests d'indépendance. Nous avons donc pris comme point de départ le Théorème 2.1 de [Kuo et al., 2010], bien connu dans la communauté de l'analyse de sensibilité. Son exploration nous a permis de comprendre comment nous devions l'étendre afin d'obtenir la décomposition de Möbius comme un cas particulier. Et nous y sommes parvenus sur deux aspects. L'exposé tentera de présenter ces extensions et nos espoirs sur leurs futurs croisements.

Bibliographie

[Deheuvels, 1981] Deheuvels, P. (1981). An asymptotic decomposition for multivariate distribution-free tests of independence. *Journal of Multivariate Analysis*, 11(1):102–113.

[Efron and Stein, 1981] Efron, B. and Stein, C. (1981). The jackknife estimate of variance. *The Annals of Statistics*, 9(3):586–596.

[Genest et al., 2019] Genest, C., Nešlehová, J. G., Rémillard, B., and Murphy, O. A. (2019). Testing for independence in arbitrary distributions. *Biometrika*, 106(1):47–68.

[Iooss and Prieur, 2019] Iooss, B. and Prieur, C. (2019). Shapley effects for sensitivity analysis with correlated inputs: Comparisons with Sobol' indices, numerical estimation and applications. *Int. J. Uncertain. Quantif.*, 9(5):493–514.

[Kuo et al., 2010] Kuo, F. Y., Sloan, I. H., Wasilkowski, G. W., and Woźniakowski, H. (2010). On decompositions of multivariate functions. *Math. Comp.*, 79(270):953–966.

[Mercadier and Roustant, 2019] Mercadier, C. and Roustant, O. (2019). The tail dependograph. *Extremes*, 22(2):343–372.

[Mercadier et al., 2022] Mercadier, C., Roustant, O., and Genest, C. (2022). Linking the Hoeffding–Sobol and Möbius formulas through a decomposition of Kuo, Sloan, Wasilkowski, and Woźniakowski. *Statistics & Probability Letters*, 185:109419.

[Roustant et al., 2020] Roustant, O., Gamboa, F., and Iooss, B. (2020). Parseval inequalities and lower bounds for variance-based sensitivity indices. *Electron. J. Statist.*, 14(1):386–412.

[Sobol', 1993] Sobol', I. M. (1993). Sensitivity estimates for nonlinear mathematical models. *Math. Modeling Comput. Experiment.*, 1(4):407–414 (1995).

Mixture of longitudinal factor analysis for modelling heterogeneous longitudinal multivariate data

Amine Ounajim [1,2] & Yousri Slaoui [1] & Pierre-Yves Louis [3] & Denis Frasca [4] & Philippe Rigoard [2]

[1] *Université de Poitiers, Laboratoire de Mathématiques et applications, UMR 7348, 11 Boulevard Marie et Pierre Curie, 86962 Futuroscope Chasseneuil, France*
[2] *PRISMATICS Lab (Predictive Research in Spine/Neuromodulation Management and Thoracic Innova-tion/Cardiac Surgery), Poitiers University Hospital, 86000 Poitiers, France.*
[3] *AgroSup Dijon, PAM UMR 02.102, Université Bourgogne Franche-Comté, 21000 Dijon, France.*
[4] *Department of Anaesthesiology and Critical Care, Poitiers University Hospital, 86021 Poitiers, France.*
Amine.OUNAJIM@chu-poitiers.fr
yousri.slaoui@math.univ-poitiers.fr
pierre-yves.louis@agrosupdijon.fr
denis.frasca@chu-poitiers.fr
Philippe.RIGOARD@chu-poitiers.fr

Résumé. Afin d'étudier l'évolution de plusieurs indicateurs observés parmi des individus, il est important de se concentrer sur les tendances longitudinales parmi les variables latentes en utilisant une modélisation conjointe basée sur les structures de covariance entre ces indicateurs observés. Cependant, ce type de données peut représenter une hétérogénéité dans le temps et selon les groupes d'individus. Dans cet article, nous proposons une extension de l'analyse factorielle classique dans laquelle la non-invariance des groupes est prise en compte en utilisant un modèle de mélange. Nous commençons par définir le mélange du modèle d'analyse factorielle longitudinale et ses paramètres. Ensuite, nous proposons un algorithme EM pour estimer le modèle. Nous proposons également un critère d'information bayésien pour identifier le nombre optimal de composantes du mélange. L'initialisation, la convergence et le temps de calcul seront discutés. Nous montrons l'applicabilité du modèle en utilisant des données simulées et réelles.

Mots-clés. Analyse factorielle longitudinale, modèle à effets mixtes, mélange de modèles, classes latentes, non-invariance factorielle.

Abstract. In order to study the evolution of several observed outcomes among patients, it is important to focus on longitudinal trends among latent variables using joint modeling based on covariance structures between these observed outcomes. However this type of data might represent heterogeneity over time and among groups of individuals. In this paper, we propose an extension to the factor analysis framework where group non-invariance is taken into account using a mixture model. We start by defining the mixture of the longitudinal factor analysis model and its parameters. Then, we propose an EM algorithm to estimate the model. We also propose a Bayesian information criterion to identify the number of components in the mixture. Initialisation, convergence and running time are discussed. We show the applicability of the model using simulated and real data.

Keywords. Longitudinal factor analysis, Mixed models, mixture models, latent class analysis, factor non-invariance.

1 Introduction

In this work, we are interested in the analysis of multivariate longitudinal data where the relationship between the variables depends on unobserved groups.
Several methods for multivariate longitudinal data analysis are available in the literature, including latent factor(s) linear mixed models (Roy and Lin 2000; An, Yang, and Bentler 2013), curve of factors models (McArdle 1988), dynamic factor model and dynamic SEM (Browne and Nesselroade 2005). These models represent the observed outcomes using one or several latent factors.

Generally, in order to study the change in several noisy outcomes among patients sampled from a homogeneous population, it is important to focus on longitudinal trends among the latent variables using joint modeling based on the covariance structures of the observed outcomes. This allows researchers to draw more general and joint conclusions instead of conducting a separate analysis for each observed outcome. One of the assumptions of the factor models is the invariance of factor structure among subgroups of subjects. This assumption supposes that the studied population is homogeneous regarding the structure of the constructs being measured. In education and health sciences, it is possible that factors structure vary among latent subpopulations due to cultural or physiological differences. To address this issue, some authors proposed the mixture of factor analysers models (McLachlan, Peel, and Bean 2003). These models estimate different factor loadings for different subpopulations that are represented by a latent class variable. To our knowledge, no extension to longitudinal data of these models was proposed previously in the literature. In this paper we propose an extension to the longitudinal factor analysis framework where group non-invariance is taken into account using a mixture of factor analysers.

2 Mixture of longitudinal factor analysis models (MLFA)

Throughout this paper, we suppose that there are n subjects with J outcomes at t_{n_i} time points $(i = 1, \ldots, n)$. The total number of observations will be noted $N = \sum_{i=1}^{n} n_i$. For subject i, time t and outcome $(y_{ijt})_{j\in\{1,\ldots,J\}}$, the proposed mixture of longitudinal factor analysis model can be written in the following manner:

$$y_{ijt} = \sum_{c=1}^{C} \mathbb{1}_{\{v_i=c\}}(\Lambda_{jc}\eta_{i.tc} + \epsilon_{ijtc}), \tag{1}$$

The quantity v_i is a categorical latent variable representing the latent class of subject i. The variable v takes values in $\{1, \ldots, C\}$ associated with the probabilities $\{\pi_1, \ldots, \pi_C\}$ $(\sum_{c=1}^{C} \pi_c = 1)$. The indicator function $\mathbb{1}_{\{v_i=c\}}$ is equal to 1 if subject i belongs to class c and 0 otherwise. The vector Λ_{jc} contains the factor loadings associated with the outcome $y_{.jt}$ for cluster c. The vector $\eta_{i.tc} = (\eta_{i1tc}, \ldots, \eta_{iKtc})$ contains the K latent factor scores for subject i in cluster c at time t. We assume that the number of latent factors is similar between classes.
We suppose that ϵ_{ijtc} follows the distribution $\mathcal{N}(0, \sigma^2_{jc})$ where σ^2_{jc} is the variance of the error terms associated with class c and the outcome j.
The latent factors for each class are modelled using the following mixed effects model:

$$\eta_{iktc} = X_{iktc}\beta_{kc} + Z_{iktc}\xi_{ikc} + \omega_{itc} \tag{2}$$

where X_{iktc} and Z_{iktc} are respectively the fixed effects and random effects covariate matrices for class c. The coefficients β_{kc} are the fixed effects associated with the factor structure of the latent class c. The random effects associated with the latent class c are represented by the random variable ξ_{ikc} which are assumed to be normally distributed with mean 0 and variance covariance matrix Σ_{ξ_c}. The quantity ω_{itc} represents the error term of the structural model which follows a Gaussian distribution $\mathcal{N}(0, \Sigma_{\omega_c})$.
Since we represent the J outcomes in y using a much smaller number K of latent factors in η, the estimation of β is simpler than estimating these effects on each one of the original J outcomes. Another advantage is that different latent factors and different classes can be modeled using different covariate matrices X_{kc} and Z_{kc}. This is useful since in practice, it can be of interest to study the effect of different variables on different latent factors or study different time-trajectories (linear, quadratic, ...) for factors in different classes.

3 Parameters estimation of the MLFA model with EM algorithm

In this section, we propose an EM algorithm (Dempster, Laird, and Rubin 1977) for estimating the proposed model. An advantage of using the EM algorithm for this model is that it is easy to implement, because closed form solutions are available for both E and M steps.

The expected log-likelihood of the complete data $\{y, X, Z, \eta, \xi\}$ can be written as follows:

$$\mathbb{E}(\log L_{comp}(\theta)|y_i) = \sum_{i=1}^{n}\sum_{c=1}^{C}\mathbb{E}(v_{ic}\log(\pi_c) + v_{ic}\log(\phi(y_i, X_{ic}, Z_{ic}, \eta_{ic}, \xi_{ic}, \theta_c))) + constant \tag{3}$$

We have the following:

$$\begin{aligned}\log(\phi(y_i, X_{ic}, Z_{ic}, \eta_{ic}, \xi_{ic}, \theta_c)) &= \log(\phi(y_i|\eta_{ic}, \theta_c)) + \log(\phi(\eta_{ic}|\xi_{ic}, \theta_c)) + \log(\phi(\xi_{ic}, \theta_c)) \\ &= \sum_{j=1}^{J}\{-\frac{t_{n_i}}{2}\log\sigma_{jc}^2 - \frac{1}{2\sigma_{jc}^2}(y_{ij} - \eta_{ic}^T\Lambda_{jc})^T(y_{ij} - \eta_{ic}^T\Lambda_{jc})\} \\ &- \frac{1}{2}\sum_{t=1}^{t_{n_i}}(\log|\Sigma_\omega| + (\eta_{itc} - Z_{itc}\xi_{ic} - X_{itc}\beta_c)^T\Sigma_\omega^{-1}(\eta_{itc} - Z_{itc}\xi_{ic} - X_{itc}\beta_c)) \\ &- (\frac{1}{2}\log|\Sigma_{\xi_c}| + \frac{1}{2}\xi_{ic}^T\Sigma_{\xi_c}^{-1}\xi_{ic}).\end{aligned}$$

Maximizing the expected log-likelihood given in equation (3) gives the following estimation equations:

$$\Lambda_c = ((\sum_{i=1}^{n}\sum_{t=1}^{t_{n_i}}\mathbb{E}(v_{ic}|y_i)\,\mathbb{E}(\eta_{itc}\eta_{itc}^T|v_{ic}, y_i)^{-1}(\sum_{i=1}^{n}\sum_{t=1}^{t_{n_i}}\mathbb{E}(v_{ic}|y_i)\,\mathbb{E}(\eta_{itc}|v_{ic}, y_i)y_{it}^T))^T, \tag{4}$$

$$\Sigma_{\xi_c} = \frac{1}{\sum_{i=1}^{n}\mathbb{E}(v_{ic}|y_i)}\sum_{i=1}^{n}\mathbb{E}(v_{ic}|y_i)\,\mathbb{E}(\xi_{ic}\xi_{ic}^T|y_i, v_{ic}). \tag{5}$$

$$\beta_c = (\sum_{i=1}^{n}\sum_{t=1}^{t_{n_i}}\mathbb{E}(v_{ic}|y_i)X_{itc}^T\Sigma_{\omega_c}^{-1}X_{itc})^{-1}(\sum_{i=1}^{n}\sum_{t=1}^{t_{n_i}}\mathbb{E}(v_{ic}|y_i)X_{itc}^T\Sigma_{\omega_c}^{-1}(\mathbb{E}(\eta_{ict}|v_{ic}, y_i) - Z_{itc}\,\mathbb{E}(\xi_{ic}|v_{ic}, y_i))), \tag{6}$$

and the probability class membership can be obtained as:

$$\pi_c = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(v_{ic}|y_i). \tag{7}$$

The quantities σ_{jc} and Σ_{ω_c} are be obtained similarly.
In the E-step of the EM algorithm, we calculate the sufficient statistics (*i.e.*, the conditional expectations on the unobserved variables). In the M-step, we maximize the expected log-likelihood using equations (4) to (7). We repeat the E and M steps until convergence.

4 Number of mixture components

In this section, we propose a BIC criterion to identify the optimal number C of mixture components. the BIC of the model could be calculated as:

$$BIC_{MLFA} = -2\log L(\theta) + (C(\#parameters) + C - 1)\log(N) \tag{8}$$

where $\#parameters = (JK + J + \frac{K(K+1)}{2} + K(p+1) + \frac{(q+1)K((q+1)K+1)}{2})$ is the number of free parameters in the model. The function L is the observed data likelihood. the quantities p and q are the number of fixed and random effect covariates.
To identify the number of components, we fit the model with different number of components and select the model yielding the smallest BIC_{MLFA}.

5 Simulations study

We used simulated data to evaluate the computational and practical properties of the proposed model. We also evaluate the performance of the BIC criterion proposed in this paper at identifying the number of mixture components. We repeat $S = 100$ simulations. We evaluate the accuracy of the parameters estimates using Mean Absolute Error ($MAE = \frac{1}{S}\sum_{s=1}^{S}|p - \hat{p}_s|$, where p is the true parameter and $\hat{p}_s$ is the estimated parameter at simulation s). We take a two-component mixture of longitudinal factor analysis models. The simulations included 500 subjects with 5 observations each. Two latent factors were considered in the simulations. These two factors were measured using 10 items. A mixed effects model with random intercept and slopes was used to simulate the relationship between the latent factors and the explanatory variables. To fix the scale of latent factors, $\lambda_{1,1,c}$ and $\lambda_{6,2,c}$ are fixed to 1 for $c = 1, 2$. The following parameters were considered in the simulations:

$$C = 2, \qquad t \in \{1, 2, 3, 4, 5\},$$
$$\pi_1 = 0.6,\ \pi_2 = 1 - \pi_1,\ \Lambda_1 = \begin{pmatrix} 1&1&1&1&1&0&0&0&0&0 \\ 0&0&0&0&0&1&1&1&1&1 \end{pmatrix}^T, \qquad \Lambda_2 = \begin{pmatrix} 1&0&0&1&1&0&1&1&0&0 \\ 0&1&1&0&0&1&0&0&1&1 \end{pmatrix}^T,$$
$$\sigma^2_{j1} = 1 \text{ for all } j \in \{1, \ldots, 10\} \text{ and } \sigma^2_{j2} = 0.75^2 \text{ for all } j \in \{1, \ldots, 10\},$$
$$X \sim \mathcal{N}((0,0)^T,\ I_2), \qquad Z = X,$$
$$\beta_{111}(t) = -1, \qquad \beta_{112}(t) = 1,$$
$$\beta_{121}(t) = 1, \qquad \beta_{122}(t) = -1,$$
$$\beta_{211}(t) = 1, \qquad \beta_{212}(t) = -1,$$
$$\beta_{221}(t) = 0, \qquad \beta_{222}(t) = -2,$$
$$\xi_{ickp} = a_{ickp} \text{ for } i \text{ in } \{1, \ldots, 500\} \text{ and } k \text{ in } \{1, 2\},$$
$$\text{where } (a_{ic11}, a_{ic12}, a_{ic21}, a_{ic22})^T \sim N((0,0,0,0)^T, \Sigma_\xi) \text{ for } c \text{ in } \{1, 2\},$$
$$\omega_{ic} \sim \mathcal{N}((0,0)^T, \Sigma_{\omega_c}),$$

where $\Sigma_\xi = \begin{pmatrix} 1 & 0.5^2 & 0.5^2 & 0.5^2 \\ 0.5^2 & 1 & 0.5^2 & 0.5^2 \\ 0.5^2 & 0.5^2 & 1 & 0.5^2 \\ 0.5^2 & 0.5^2 & 0.5^2 & 1 \end{pmatrix}$ and $\Sigma_{\omega_1} = \begin{pmatrix} 0.5^2 & 0.2^2 \\ 0.2^2 & 0.5^2 \end{pmatrix}$ for $c = 1$ and $\Sigma_{\omega_2} = \begin{pmatrix} 0.75^2 & 0.5^2 \\ 0.5^2 & 0.75^2 \end{pmatrix}$ for $c = 2$. We considered both subject-specific intercept and slopes.

Results of the first simulation can be found in Table 1. The model estimates the true parameters with relatively low error. Regarding the evaluation of the BIC criterion, we found that the proposed BIC_{MLFA} the optimal number of mixture components. The number of mixture components was correctly identified for 86% of the simulations for the model with a single component and for 91% of the simulations for the model with a two components.

6 Real data example

The majority of evaluation tools used in the pain literature are developed and validated on chronic pain populations which do not share the same goals regarding pain treatments. Therefore, weighting unimportant items differently when evaluating these patients would give a better representation of the true evolution of their health following a therapy (Butowicz et al. 2019). In this application, we consider the situation where these observed longitudinal item scores estimate several underlying longitudinal latent outcomes with structures that vary between individuals based on their latent sociodemographic

Table 1: Results of the simulation study (N = 2500). For each parameter, we reported the mean absolute error and its standard deviation over 100 simulations.

Component 1 parameters		Component 2 parameters	
Parameter	MAE (SD of AE)	Parameter	MAE (SD of AE)
π_1	0.018 (0.014)	π_2	-
$\lambda_{2,1,1}$	0.017 (0.012)	$\lambda_{2,2,2}$	0.015 (0.012)
$\lambda_{3,1,1}$	0.015 (0.013)	$\lambda_{3,2,2}$	0.013 (0.010)
$\lambda_{4,1,1}$	0.018 (0.014)	$\lambda_{4,1,2}$	0.015 (0.012)
$\lambda_{5,1,1}$	0.017 (0.014)	$\lambda_{5,1,2}$	0.017 (0.013)
$\lambda_{7,2,1}$	0.016 (0.013)	$\lambda_{7,1,2}$	0.015 (0.012)
$\lambda_{8,2,1}$	0.017 (0.013)	$\lambda_{8,1,2}$	0.016 (0.014)
$\lambda_{9,2,1}$	0.016 (0.011)	$\lambda_{9,2,2}$	0.014 (0.010)
$\lambda_{10,2,1}$	0.017 (0.012)	$\lambda_{10,2,2}$	0.013 (0.010)
$\beta_{1,1,1}$	0.127 (0.039)	$\beta_{2,1,1}$	0.042 (0.032)
$\beta_{1,1,2}$	0.040 (0.026)	$\beta_{2,1,2}$	0.063 (0.040)
$\beta_{1,2,1}$	0.049 (0.028)	$\beta_{2,2,1}$	0.059 (0.039)
$\beta_{1,2,2}$	0.101 (0.038)	$\beta_{2,2,2}$	0.109 (0.050)

and behavioural characteristics.
We applied the model to a set of chronic pain evaluation questionnaires to evaluate if there is heterogeneity in the factorial structure of the items. We used the PREDIBACK study data, which comprises of clinical, sociodemographic, cognitive-behavioural variables and a set of evaluation questionnaires of 200 patients with chronic pain after spinal surgery, evaluated each 3 months over a 1-year follow-up (Ounajim et al. 2021). Therefore, each patient has at least one observation and at most 5 observations ($n_i \in \{1, \ldots, 5\}$). Observations with missing were removed from the analysis (8 patients had missing data at all visits). The number of patients was $n = 192$ and the total number of observations was $N = 630$. Specifically, 34 patients had data at one visit, 25 at $n_i = 2$ visits, 36 at $n_i = 3$ visits, 47 at $n_i = 4$ visits and 50 at $n_i = 5$ visits. We used in this example $J = 23$ items which consist of Likert scales (ordinal variables) evaluating pain intensity, functional disability, quality of life and psychological distress. The number of mixture components was identified using the BIC_{MLFA}.
The main difference observed between the two clusters is that for the first component ($n = 57$), the first latent factor η_{11} mainly explains the items scoring "mobility", "usual activities", "social life", "enjoying things you used to enjoy", "laughing and seeing the good side of things", "feeling cheerful", "feeling slowed down" and "looking forward with enjoyment to things". These variables represent the subjective perception of the enjoyment that social and daily activities confer to the patient. On the other hand, the first factor η_{12} of the second majority cluster ($n = 135$) explains the items scoring "mobility", "self-care", "usual activities", "lifting", "walking", "standing", "social life" and "travelling" which are more associated with a more objective perception of ones true functional capacity.
The mixed effect model included in the model allowed us to estimate the polynomial time trajectories of the latent factors which can be found in Figure 1.

7 Conclusion

In this paper, we propose a mixture of longitudinal factor analyzers to model multivariate longitudinal heterogeneous data. This model could be applied to many fields in which subjects are evaluated using questionnaires or several evaluation criteria over a period of time, in order to achieve cluster-based holistic evaluation.

References

An, Xinming, Qing Yang, and Peter M. Bentler (Oct. 2013). "A latent factor linear mixed model for high-dimensional longitudinal data analysis". eng. In: *Statistics in medicine* 32.24, pp. 4229–4239. ISSN: 0277-6715. URL: https://pubmed.ncbi.nlm.nih.gov/23640746.

Browne, Michael W. and John R. Nesselroade (2005). *Representing Psychological Processes With Dynamic Factor Models: Some Promising Uses and Extensions of Autoregressive Moving Average Time Series Models.* Multivariate applications book series. Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers, pp. 415–452.

Butowicz, Courtney M et al. (Nov. 2019). "Characterizing and Understanding the Low Back Pain Experience Among Persons with Lower Limb Loss". In: *Pain Medicine* 21.5, pp. 1068–1077. ISSN: 1526-2375. DOI: 10.1093/pm/pnz293. eprint: https://academic.oup.com/painmedicine/article-pdf/21/5/1068/33194030/pnz293.pdf. URL: https://doi.org/10.1093/pm/pnz293.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). "Maximum Likelihood from Incomplete Data via the EM Algorithm". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 39.1, pp. 1–38. ISSN: 00359246. URL: http://www.jstor.org/stable/2984875.

McArdle, J. J. (1988). *Dynamic but structural equation modeling of repeated measures data.* Perspectives on individual differences. New York, NY, US: Plenum Press, pp. 561–614. DOI: 10.1007/978-1-4613-0893-5_17.

McLachlan, GJ, D Peel, and RW Bean (2003). "Modelling high-dimensional data by mixtures of factor analyzers". In: *Computational Statistics & Data Analysis* 41.3. Recent Developments in Mixture Model, pp. 379–388. ISSN: 0167-9473. DOI: https://doi.org/10.1016/S0167-9473(02)00183-4. URL: https://www.sciencedirect.com/science/article/pii/S0167947302001834.

Ounajim, Amine et al. (Oct. 2021). "Finite Mixture Models Based on Pain Intensity, Functional Disability and Psychological Distress Composite Assessment Allow Identification of Two Distinct Classes of Persistent Spinal Pain Syndrome after Surgery Patients Related to Their Quality of Life". eng. In: *Journal of clinical medicine* 10.20, p. 4676. ISSN: 2077-0383. URL: https://pubmed.ncbi.nlm.nih.gov/34682799.

Roy, Jason and Xihong Lin (2000). "Latent Variable Models for Longitudinal Data with Multiple Continuous Outcomes". In: *Biometrics* 56.4, pp. 1047–1054. ISSN: 0006341X, 15410420. URL: http://www.jstor.org/stable/2677036.

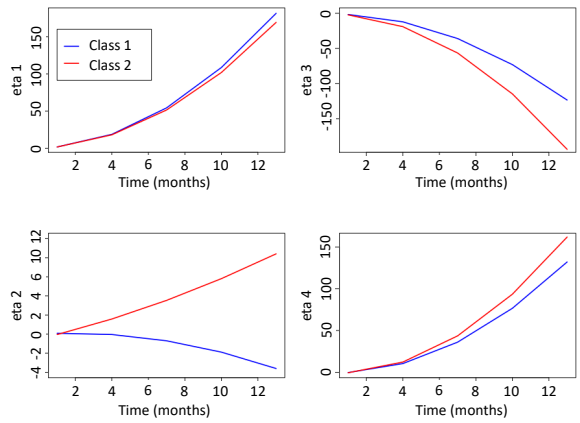


Figure 1: Time trajectories of the four extracted latent factors for class 1 (blue lines) and class 2 (red lines) patients.

Estimation adaptative dans un modèle de régression-convolution en base d'Hermite

Ousmane Sacko

MAP5 UMR CNRS 8145, Université Paris Cité
ousmane.sacko@u-paris.fr

Résumé. Dans ce travail, on considère le modèle de régression-convolution suivant : $y(x_k) = f \star g(x_k) + \varepsilon_k, x_k = kT/n, k = -n, \dots, n-1$, T fixé ; g est connue et f est la fonction inconnue que l'on cherche à estimer. Les erreurs $(\varepsilon_k)_{-n\leqslant k\leqslant n-1}$ sont i.i.d., centrées de variance finie et connue. Nous proposons un estimateur par projection en exploitant les propriétés de la base d'Hermite. Une borne du risque est prouvée : sous des conditions de régularité, nous obtenons une vitesse de convergence. Une procédure adaptative pour choisir la dimension pertinente est aussi proposée en s'inspirant de la méthode de Goldenshluger and Lepski (2011) et nous prouvons que l'estimateur résultant satisfait une inégalité oracle pour ε sous-Gaussien. Enfin, nous illustrons la procédure par des graphes.

Mots-clés. Base d'Hermite ; Estimateur par projection ; Modèle de convolution ; Régression non-paramétrique ; Sélection de modèle.

Abstract. In this work, we consider the following regression-convolution model : $y(x_k) = f \star g(x_k) + \varepsilon_k, x_k = kT/n, k = -n, \dots, n-1$, fixed T ; g is known and f is the unknown function to be estimated. The errors $(\varepsilon_k)_{-n\leqslant k\leqslant n-1}$ are i.i.d. centered with finite known variance. We propose a projection estimator by exploiting the properties of the Hermite basis. A risk bound is proved : under regularity conditions, we derive rates of convergence. An adaptive procedure to select the relevant dimension inspired by the Goldenshluger and Lepski (2011) method is also proposed and we prove that the resulting estimator satisfies an oracle inequality for sub-Gaussian ε's. Finally, we illustrate the procedure by some figures.

Keywords. Convolution model ; Hermite basis ; Model selection ; Nonparametric regression ; Projection estimator.

1. Introduction

Considérons le modèle de régression-convolution suivant :

$$y(x_k) = h(x_k) + \varepsilon_k, \quad k = -n, \dots, n-1, \tag{1}$$

où $h(x) = f \star g(x) = \int_{\mathbb{R}} f(x-y)g(y)dy$, la fonction g appelée *noyau* est supposée connue et f est la fonction inconnue que l'on souhaite reconstruire. Les erreurs $(\varepsilon_k)_{-n\leqslant k\leqslant n-1}$ sont i.i.d. avec $\mathbb{E}[\varepsilon_k] = 0$ et $\mathrm{Var}(\varepsilon_k) = \sigma_\varepsilon^2 < \infty$, connu. Les points $(x_k = kT/n)_{-n\leqslant k\leqslant n-1}$ sont déterministes et équirépartis sur $[-T, T]$, où $0 < T < \infty$ est fixé.

Ce modèle se trouve dans beaucoup de domaines d'applications : en analyse des données d'imagerie dynamique avec produit de contraste (voir *e.g.* Cuenod et al. (2006), Comte

et al. (2017)) et dans l'étude de la spectroscopie de fluorescence résolue en temps (voir *e.g.* McKinnon et al. (1977) et Abramovich et al. (2013)).

L'estimation de la fonction inconnue h à partir des données $(x_k, y(x_k))$ est connue comme *problème de régression non paramétrique en "fixed design"*. Ce problème a été largement étudié dans la littérature, voir Barron et al. (1999), Baraud (2000) par exemple. Un autre problème classique est celui où l'on cherche à estimer la densité f d'une variable X alors qu'on observe $Z = X + \varepsilon$. Si g désigne la densité de ε, le problème est aussi de reconstruire f à partir d'un estimateur de $f_Z = f \star g$. Ce problème est connu comme un *problème de déconvolution*. Il s'agit d'un *problème inverse* qui est aussi intensivement étudié dans la littérature, voir *e.g.* Fan (1991), Pensky and Vidakovic (1999), Comte and Lacour (2011), Mabon (2017), Sacko (2020). Le modèle (1) cumule deux questions *régression et déconvolution*, c'est pourquoi il est difficile. Les résultats théoriques sur ces deux problèmes traités *indépendamment* sont bien connus.

Si le support de f et g est inclus dans $\mathbb{R}^+$, le problème (1) a été étudié entre autres par : Dey et al. (1998), Abramovich et al. (2013), Vareschi (2015), Comte et al. (2017) et Benhaddou et al. (2019).

Dans ce travail, on désire estimer la fonction f sans a priori sur le support des fonctions f et g, à partir des données $(x_k, y(x_k))$ issues du modèle (1). Nous considérons la base d'Hermite qui est à support non compact et est donc bien adaptée à notre problème général. Nous considérons l'hypothèse suivante, classique en déconvolution :

(**H1**) Il existe des constantes c_1, c_1', γ, $c_1 \geqslant c_1' > 0$ et $\gamma > 0$ telles que

$$c_1'(1+t^2)^{\gamma} \leqslant |g^*(t)|^{-2} \leqslant c_1(1+t^2)^{\gamma}, \quad \forall t \in \mathbb{R}, \quad t^*(u) = \int e^{iux} t(x) dx. \tag{2}$$

Nous allons proposer une approche de projection, fondée sur une décomposition de f et h en base d'Hermite. Nous obtenons un estimateur de f en remplaçant h par son estimateur des moindres carrés dans l'expression des coefficients de f.

Les résultats énoncés ici proviennent de l'article Sacko (2022).

2. Définition de l'estimateur

2.1. **Base d'Hermite.** C'est une base orthonormée sur $\mathbb{L}^2(\mathbb{R})$, définie à partir du polynôme d'Hermite $(H_j)_{j\geqslant 0}$ (voir Bongioanni and Torrea (2009)) :

$$\varphi_j(x) = c_j H_j(x) e^{-x^2/2}, \; H_j(x) = (-1)^j e^{x^2} \frac{d^j}{dx^j}(e^{-x^2}), \; c_j = (2^j j! \sqrt{\pi})^{-1/2}, x \in \mathbb{R}.$$

2.2. **Procédure d'estimation.** Soit $(x_k, y(x_k))_{-n\leqslant k\leqslant n-1}$ issues du modèle (1) et $S_m = \text{Vect}\{\varphi_0, \ldots, \varphi_{m-1}\}$, l'espace linéaire engendré par $(\varphi_j)_{0\leqslant j\leqslant m-1}$. Pour $f \in \mathbb{L}^2(\mathbb{R})$, nous avons $f = \sum_{j=0}^{\infty} a_j(f)\varphi_j, a_j(f) = \langle f, \varphi_j\rangle = \int f(x)\varphi_j(x)dx$ et sa projection orthogonale sur S_m est : $f_m = \sum_{j=0}^{m-1} a_j(f)\varphi_j$. Pour estimer f, nous construisons m-estimateurs des coefficients $a_j(f)$. A cette fin, on remarque sous (**H1**) et par le théorème de Plancherel :

$$a_j(f) = \langle \frac{h^*}{g^*}, \varphi_j^* \rangle, \quad \varphi_j^* = \sqrt{2\pi}(i)^j \varphi_j. \tag{3}$$

Ainsi, on obtient un estimateur des $a_j(f)$ en remplaçant h^* par un estimateur. Pour reconstruire h, on utilise la méthode des moindres carrés. Posons :

$$\Phi_d = (\varphi_j(x_i))_{-n\leqslant i\leqslant n-1, 0\leqslant j\leqslant d-1}, \quad \Psi_d = \frac{T}{n}\Phi_d^t\Phi_d,$$

où A^t désigne la transposée de la matrice A. En minimisant $\|\vec{y} - \Phi_d\vec{b}\|^2_{\mathbb{R}^{2n}}$ en $\vec{b} \in \mathbb{R}^d$, on estime h par :

$$\widehat{h}_d = \sum_{j=0}^{d-1} \widehat{b}_j^{(d)}\varphi_j, \quad \text{où} \quad \vec{\widehat{b}}^{(d)} = (\widehat{b}_0^{(d)}, \dots, \widehat{b}_{d-1}^{(d)})^t = (\Phi_d^t\Phi_d)^{-1}\Phi_d^t\vec{y} = \frac{T}{n}\Psi_d^{-1}\Phi_d^t\vec{y}, \tag{4}$$

$\vec{y} = (y(x_{-n}), \dots, y(x_{n-1}))^t$. Notons qu'ici Ψ_d est inversible pour $d \leqslant n$. Ainsi, l'estimateur $\widehat{h}_d$ est bien défini. En injectant la transformée de Fourier de $\widehat{h}_d$ donné par (4) dans la formule (3), on définit l'estimateur suivant :

$$\widehat{f}_{m,d} = \sum_{j=0}^{m-1} \widehat{a}_{j,d}\varphi_j, \quad \widehat{a}_{j,d} = \frac{(-i)^j}{\sqrt{2\pi}} \int \frac{\widehat{h}_d^*(u)}{g^*(u)}\varphi_j(u)du. \tag{5}$$

L'estimateur $\widehat{f}_{m,d}$ est bien défini car la base d'Hermite, qui décroit en $e^{-\xi x^2}$ rend la fonction $(\widehat{h}_d^*/g^*)\varphi_j$ intégrable pour toutes fonctions g satisfaisant **(H1)**. Les dimensions m et d doivent être optimisées. Clairement, la performance de $\widehat{f}_{m,d}$ est liée à celle de $\widehat{h}_d$: une étude complète sur $\widehat{h}_d$ se trouve dans Sacko (2022) (voir Section 2).

3. Borne du risque

3.1. Majoration du risque de $\widehat{f}_{m,d}$.

On introduit la notation suivante :

$$\Delta(m) = \sup_{|u|\leqslant\sqrt{\rho m}} |g^*(u)|^{-2} + \sum_{j=0}^{m-1} \int_{|u|\geqslant\sqrt{\rho m}} |\varphi_j(u)|^2|g^*(u)|^{-2}du, \quad \rho > 0. \tag{6}$$

Proposition 3.1. *Soient $f \in \mathbb{L}^2(\mathbb{R})$ et $h \in \mathbb{L}^2(\mathbb{R})$. Sous* **(H1)**, *on a*

(7)

$$\mathbb{E}\left[\|\widehat{f}_{m,d} - f\|^2\right] \leqslant \|f - f_m\|^2 + 2\Delta(m)\left(\|h - h_d\|^2 + \lambda_{max}\left(\Psi_d^{-1}\right)\|h - h_d\|_n^2 + \sigma_\varepsilon^2\frac{T}{n}\mathrm{tr}\left(\Psi_d^{-1}\right)\right),$$

où $\|t\|_n^2 = \frac{T}{n}\sum_{i=-n}^{n-1} t^2(x_i)$ est la norme empirique de t, $\|t\|^2 = \int t^2(x)dx$ sa version intégrale, $\lambda_{max}(A)$ est le rayon spectral de la matrice A et $\mathrm{tr}(A)$ sa trace.

- Le premier terme de droite dans (7) est le terme classique de biais : il décroît avec m.
- Le terme $\Delta(m)$ correspond à l'aspect déconvolution du problème : il est étudié en utilisant **(H1)**. Il croît avec m.
- Enfin, le terme dans la grande parenthèse représente l'aspect régression du problème, c'est le risque d'estimation de h pour la norme intégrale.

3.2. **Vitesse de convergence.** Pour déduire les résultats de convergence, nous allons faire deux compromis biais-variance, d'abord pour la partie régression ensuite pour le risque global en injectant cette valeur dans (7). On fera ensuite un autre compromis vis-a-vis de m pour obtenir la vitesse de $\widehat{f}_{m,d}$. Il faut quantifier les ordres de grandeur de chaque terme donné en (7). Pour ce faire, on considère l'hypothèse suivante :

(**H2**) Il existe une constante $\lambda > 0$ telle que :

$$\lambda_{\max}(\Psi_d^{-1}) \leqslant \lambda < +\infty, \text{ uniformément en } d,$$

et nous rappelons aussi la définition de l'espace de *régularité* associé à la base d'Hermite :

Définition 3.1 (Boule de Sobolev-Hermite)**.** *Soient $s, L > 0$, on définit la boule de Sobolev-Hermite de régularité s et de rayon L par :*

$$W_H^s(L) = \{\theta \in \mathbb{L}^2(\mathbb{R}), \sum_{k\geqslant 0} k^s a_k^2(\theta) \leqslant L\}, \quad a_k(\theta) = \int \theta(x)\varphi_k(x)dx.$$

Théorème 3.1. *Supposons que les hypothèses* (**H1**), (**H2**) *vérifiées et $h \in W_H^{s+\gamma}(L)$. Alors pour $d_{opt} = m_{opt} = [n^{1/(s+\gamma+1)}]$ avec $s + \gamma > 11/6$, on en déduit que*

$$\sup_{f\in W_H^s(L)} \mathbb{E}\left[\|\widehat{f}_{m_{opt},d_{opt}} - f\|^2\right] = \mathcal{O}\left(n^{-\frac{s}{s+\gamma+1}}\right).$$

L'estimateur $\widehat{f}_{m_{opt},d_{opt}}$ converge à une vitesse polynomiale comme dans le cas de déconvolution d'une densité. Cette vitesse n'est pas standard (voir Fan (1991) pour la déconvolution d'une densité).

Toutefois, *l'oracle* $\widehat{f}_{m_{opt},d_{opt}}$ n'est pas calculable puisqu'il dépend de s, qui est inconnue, voir la section suivante pour un choix adaptatif *i.e.* conduite par les données.

4. Adaptation

Le but de cette section est de proposer une méthode de sélection automatique pour $\widehat{f}_{m,d}$. Elle consiste à substituer le biais et la variance de $\widehat{f}_{m,d}$ par des quantités calculables. Nous nous inspirons des idées développées par Goldenshluger and Lepski (2011). Dans la suite, on fixe $m = d$ (voir Théorème 3.1) et on considère l'estimateur suivant :

$$\widetilde{f}_{(m)} = \widehat{f}_{m,m}, \quad \text{où } \widehat{f}_{m,m} \text{ est donné en (5).} \tag{8}$$

On estime le biais par :

$$\widetilde{A}(m) = \max_{m'\in\mathcal{M}_n} \left\{ \left(\|\widetilde{f}_{(m')} - \widetilde{f}_{(m\wedge m')}\|^2 - \kappa_1 V(m') \right)_+ \right\}, \quad \kappa_1 > 0,$$

où $m \wedge m' = \min(m, m')$, $V(m) = 4\,(1 + 24\log(n))\,\sigma_\varepsilon^2 \Delta(m) \frac{\lambda m T}{n}$ avec $\Delta(m)$ défini en (6), et $\mathcal{M}_n$ est une collection finie de modèles. Ainsi, on sélectionne m comme suit :

$$\widetilde{m} := \arg\min_{m\in\mathcal{M}_n} \left\{ \widetilde{A}(m) + \kappa_2 V(m) \right\}, \tag{9}$$

où $\kappa_2 \geqslant \kappa_1 > 0$ doivent être *calibrées*. On ajoute l'hypothèse suivante sur le bruit :

(**H3**) ε_1 est une variable sous-Gaussienne de paramètre $b > 0$: *i.e.*

$$\mathbb{E}[e^{t\varepsilon_1}] \leqslant e^{\frac{b^2t^2}{2}}, \quad \forall t \in \mathbb{R}.$$

L'hypothèse (**H3**) est vérifiée par une variable Gaussienne centrée et pour toutes les variables bornées.

Théorème 4.1 (Inégalités oracles)**.** *Supposons que les hypothèses* (**H1**) *à* (**H3**) *vérifiées. Soit* $\widetilde{f}_{\widetilde{m}}$ *défini en* (8) *avec* $\widetilde{m}$ *sélectionné par* (9)*. Alors, il existe* $\kappa_1 \geqslant 12$*, tel que*

$$\mathbb{E}[\|\widetilde{f}_{(\widetilde{m})} - f\|^2] \leqslant C \inf_{m \in \mathcal{M}_n^{(2)}} \left(\|f - f_m\|^2 + R_b(m) + V(m)\right) + C' \frac{\log(n)}{n}, \tag{10}$$

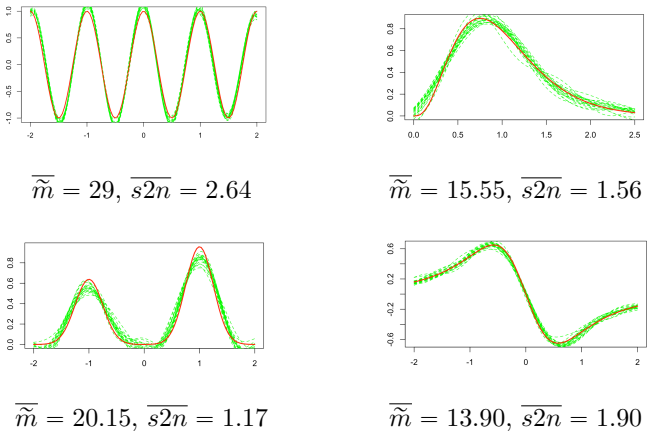
où $R_b(m) = \max_{m' \in \mathcal{M}_n^{(2)}, m \leqslant m'} \Delta(m') \|h - \mathbb{E}[\widehat{h}_{m'}]\|^2$*,* C *est une constate et* $C' = C'(\mathbb{E}[\varepsilon_1^4], \gamma, c_1, \xi, \lambda, C'_\infty)$ *avec* c'_1, γ *donnés en* (**H1**)*,* ξ, C'_∞ *liées à la base d'Hermite et* λ *défini en* (**H2**)*.*
De plus, si $f \in W_H^s(L)$ *et* $h \in W_H^{s+\gamma}(L')$ *avec* $s + \gamma \geqslant 17/6$*, on en déduit*

$$\mathbb{E}[\|\widetilde{f}_{(\widetilde{m})} - f\|^2] \leqslant C_1 \inf_{m \in \mathcal{M}_n^{(2)}} \left(m^{-s} + V(m)\right) + C'_1 \frac{\log(n)}{n}, \tag{11}$$

où $C_1 = C_1(C, L, L', s, \gamma)$ *et* $C'_1 = C'_1(C', s, \gamma)$*.*

Les Inégalités (10) et (11) sont non asymptotiques et montrent que $\widetilde{f}_{(\widetilde{m})}$ réalise automatiquement un compromis biais-variance à un facteur log près, et un terme résiduel $C' \log(n)/n$, qui est négligeable. En outre, on en déduit du Théorème 3.1 (où n est remplacé par $\log(n)/n$) que : $\mathbb{E}[\|\widetilde{f}_{(\widetilde{m})} - f\|^2] = \mathcal{O}((\frac{n}{\log(n)})^{-\frac{s}{s+\gamma+1}})$. Notons que la valeur $\kappa_1 \geqslant 12$ préconisée par la théorie est trop grande ; ainsi en pratique on a pris $\kappa_1 = 10^{-3}$ et $\kappa_2 = 2 \times 10^{-3}$.

5. Illustration



$\overline{\widetilde{m}} = 29$, $\overline{s2n} = 2.64$ — $\overline{\widetilde{m}} = 15.55$, $\overline{s2n} = 1.56$

$\overline{\widetilde{m}} = 20.15$, $\overline{s2n} = 1.17$ — $\overline{\widetilde{m}} = 13.90$, $\overline{s2n} = 1.90$

Figure 1. 20 estimateurs de $\widetilde{f}_{(\widetilde{m})}$ pour $n = 1000$. La vraie fonction est en trait continu (rouge) et les estimés en pointillés (vert).

Dans la Figure 1, nous illustrons la procédure adaptative. Nous représentons dans chaque cas la vraie fonction inconnue et 1 faisceau de 20 estimateurs pour différentes fonctions. La dimension sélectionnée par la méthode et le ratio Signal-bruit $s2n$ sont donnés en dessous de chaque graphe. Nous observons que la méthode donne des résultats très satisfaisants, visuellement. Des études numériques plus complètes et des comparaisons avec une autre stratégie se trouvent dans l'article Sacko (2022).

Références

Abramovich, F., Pensky, M., and Rozenholc, Y. (2013). Laplace deconvolution with noisy observations. *Electron. J. Stat.*, 7 :1094–1128.

Baraud, Y. (2000). Model selection for regression on a fixed design. *Probab. Theory Related Fields*, 117(4) :467–493.

Barron, A., Birgé, L., and Massart, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, 113(3) :301–413.

Benhaddou, R., Pensky, M., and Rajapakshage, R. (2019). Anisotropic functional Laplace deconvolution. *J. Statist. Plann. Inference*, 199 :271–285.

Bongioanni, B. and Torrea, J. L. (2009). What is a Sobolev space for the Laguerre function systems ? *Studia Math.*, 192(2) :147–172.

Comte, F., Cuenod, C.-A., Pensky, M., and Rozenholc, Y. (2017). Laplace deconvolution on the basis of time domain data and its application to dynamic contrast-enhanced imaging. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 79(1) :69–94.

Comte, F. and Lacour, C. (2011). Data-driven density estimation in the presence of additive noise with unknown distribution. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 73(4) :601–627.

Cuenod, C., Fournier, L., Balvay, D., and Guinebretiere, J.-M. (2006). Tumor angiogenesis : pathophysiology and implications for contrast-enhanced mri and ct assessment. *Abdominal imaging*, 31(2) :188–193.

Dey, A. K., Martin, C. F., and Ruymgaart, F. H. (1998). Input recovery from noisy output data, using regularized inversion of the laplace transform. *IEEE Transactions on Information Theory*, 44(3) :1125–1130.

Fan, J. (1991). Asymptotic normality for deconvolution kernel density estimators. *Sankhyā Ser. A*, 53(1) :97–110.

Goldenshluger, A. and Lepski, O. (2011). Bandwidth selection in kernel density estimation : oracle inequalities and adaptive minimax optimality. *Ann. Statist.*, 39(3) :1608–1632.

Mabon, G. (2017). Adaptive deconvolution on the non-negative real line. *Scand. J. Stat.*, 44(3) :707–740.

McKinnon, A., Szabo, A., and Miller, D. (1977). The deconvolution of photoluminescence data. *The Journal of Physical Chemistry*, 81(16) :1564–1570.

Pensky, M. and Vidakovic, B. (1999). Adaptive wavelet estimator for nonparametric density deconvolution. *Ann. Statist.*, 27(6) :2033–2053.

Sacko, O. (2020). Hermite density deconvolution. *ALEA Lat. Am. J. Probab. Math. Stat.*, 17(1) :419–443.

Sacko, O. (2022). Hermite estimation in noisy convolution model. *Preprint hal-03517620*.

Vareschi, T. (2015). Noisy Laplace deconvolution with error in the operator. *J. Statist. Plann. Inference*, 157/158 :16–35.

A local version of R-hat for MCMC convergence diagnostic

Théo Moins[1], Julyan Arbel[1], Anne Dutfoy[2] & Stéphane Girard[1]

[1] *Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France*
{theo.moins, julyan.arbel, stephane.girard}@inria.fr
[2] *EDF R&D dept. Périclès, 91120 Palaiseau, France*
anne.dutfoy@edf.fr

Résumé. Diagnostiquer la convergence des chaînes de Markov est un enjeu crucial pour les méthodes de Monte Carlo par chaîne de Markov. Parmi les méthodes les plus populaires, le diagnostique de Gelman–Rubin, communément appelé $\hat{R}$, est un indicateur qui permet de vérifier la convergence en se basant sur une comparaison des variances inter- et intra-chaîne. Nous proposons ici une version localisée $\hat{R}(x)$ qui se concentre sur un quantile x de la distribution, et analysons certaines propriétés de la valeur théorique associée $R(x)$. Ceci conduit à proposer un nouvel indicateur $\hat{R}_\infty$, qui permet à la fois de localiser la convergence de la chaîne en différents quantiles de la distribution, et en même temps de traiter certains problèmes de convergence non détectés par d'autres versions de $\hat{R}$.

Mots-clés. Chaînes de Markov, Statistiques bayésienne, Diagnostique de convergence

Abstract. Diagnosing convergence of Markov chain is crucial for Markov Chain Monte Carlo methods and remains an essentially unsolved problem. Among the most popular methods, the Gelman–Rubin potential scale reduction factor, commonly named $\hat{R}$, is an indicator that monitors the convergence of output chains to a stationary distribution, based on a comparison of the between- and within-variances of the chains. Here, we propose a localized version $\hat{R}(x)$ that focuses on quantiles of the distribution and we analyse some properties of the associated population value $R(x)$. This leads to proposing a new indicator $\hat{R}_\infty$, which is shown to allow both for localizing the Markov chain Monte Carlo convergence in different quantiles of the distribution, and at the same time for handling some convergence issues not detected by other $\hat{R}$ versions.

Keywords. Markov chains, Bayesian statistics, Convergence diagnostic

1 Introduction

Markov chain Monte Carlo (MCMC) algorithms have strongly contributed to the popularity of Bayesian models to sample from posterior distributions, especially in high-dimensional or high computational settings. The fundamental idea behind those algorithms is the convergence of the sampling distribution to the target (typically the posterior) when the number of samples goes to infinity. A major challenge is therefore to

know if the behaviour for a finite number of draws is satisfactory or not. This allows for a handle on the number of iterations to be drawn, which is all the more crucial in complex models with costly sampling schemes. See Roy (2020) for a recent literature review on convergence diagnostics.

A property that is frequently mentioned to verify chain convergence is mixing (see Vats and Flegal (2021) for a discussion). It refers in practice to the exploration of the support of F: slow mixing chains correspond to chains that only explore a subset of the parameter space, which can lead to strong bias in the distribution.

A common way to avoid mixing issues is to run several chains in parallel with different starting points, which also allows comparing the chains together. We place ourselves in that case: consider m chains of size n, with $\theta^{(i,j)}$ denoting the ith draw from chain j. We focus here on the Gelman–Rubin diagnostic (Gelman and Rubin, 1992), named potential reduction scale factor and commonly denoted by $\hat{R}$. It is by far one of the most popular methods to assess MCMC convergence, used in particular in Stan, PyMC3, or NIMBLE. The original heuristic for $\hat{R}$ construction is the comparison between two estimators that converge to the target variance $\mathrm{Var}[\theta]$, based on $\hat{W}$ and $\hat{B}$, respectively the estimated within- and between- variances. This diagnostic has the advantage of being scalar even in the case of a huge number of chains and comes with a rule of thumb that makes it very easy to use: generally $\hat{R} \geq 1$, and if it is greater than a given threshold (for example 1.01), then a convergence issue is raised.

The use of the original version of Gelman and Rubin (1992) has some limitations, which can be found for example in Vats and Knudson (2021); Vehtari et al. (2021): lack of interpretability, lack of robustness for certain types of non-convergence, arbitrary choice of a threshold, etc. We hope to take a step forward in addressing these limitations with a localized version of $\hat{R}$ introduced in Moins et al. (2021) and developed here and in Moins et al. (2022): we analyze $\hat{R}(x)$, a local version of $\hat{R}$ associated with a given quantile x, and the corresponding population value $R(x)$. This study leads us to suggest a new indicator $\hat{R}_\infty$, which in addition to being more interpretable, shows better results in terms of MCMC convergence diagnostic.

2 A local version of $\hat{R}$

Population version. For all $x \in \mathbb{R}$, introduce the Bernoulli random variable $I_x = \mathbb{I}\{\theta \leq x\}$, where $\mathbb{I}\{\cdot\}$ denotes the indicator function. The idea of our local convergence estimate is decidedly simple: we use I_x in place of θ in the original Gelman–Rubin construction. If $Z \in \{1, \ldots, m\}$ denotes the corresponding index of the chain, the population within-chain and between-chain variances at point x are then defined respectively as $W(x) = \mathbb{E}\left[\mathrm{Var}[I_x \mid Z]\right]$ and $B(x) = \mathrm{Var}\left[\mathbb{E}[I_x \mid Z]\right]$. Note that both quantities exist whatever the tail heaviness of θ distribution thanks to introduction of the indicator function, thus relaxing moment conditions of the original $\hat{R}$. We define the associated population $R(x)$

as

$$R(x) = \sqrt{\frac{W(x) + B(x)}{W(x)}}.$$

It turns out that under the assumption of stationarity for each chain, $R(x)$ can be expressed in closed form with respect to the chains' distribution.

Proposition 2.1 *Suppose that, for any $j \in \{1, \dots, m\}$, $\mathbb{P}(Z = j) = 1/m$ and θ given $Z = j$ has cumulative distribution function (cdf) F_j. Then, one has for any $x \in \mathbb{R}$:*

$$R(x) = \sqrt{1 + \frac{\sum_{j=1}^{m} \sum_{k=j+1}^{m} \left(F_j(x) - F_k(x)\right)^2}{m \sum_{j=1}^{m} F_j(x)(1 - F_j(x))}}. \tag{1}$$

Thus, using I_x instead of θ defines a local convergence estimate at any point x which quantifies a distance between the F_j's. This allows for diagnostic convergence relatively to a quantile one wants to estimate (for a posterior credible interval for example).

In order to summarize this continuous index into a scalar one, we may also consider its supremum over $\mathbb{R}$:

$$R_\infty = \sup_{x \in \mathbb{R}} R(x). \tag{2}$$

Note that R_∞ is finite simply as soon as the F_j's are continuous with overlapping supports. Considering R_∞ amounts to considering the local version $R(x)$ corresponding to the quantile x with the poorest convergence when no information is given on the posterior interval used for inference.

Sample version. Population version $R(x)$ can be estimated by replacing $F_j(x)$ in (1) by its empirical counterpart $\hat{F}_j(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{\theta^{(i,j)} \leq x\}$. This is equivalent to computing the original version of $\hat{R}$ on indicator variables $I_x^{(i,j)} = \mathbb{I}\{\theta^{(i,j)} \leq x\}$ instead of $\theta^{(i,j)}$. This connects with the Raftery–Lewis diagnostic (Raftery and Lewis, 1992) and more recently with Vehtari et al. (2021) who suggest this transformation for effective sample size (ESS) to construct graphical diagnostics or "tail-versions" of this diagnostic. Moreover, a rank-normalization step is added in Vehtari et al. (2021)'s versions to prevent from infinite moments, although using $I_x^{(i,j)}$ ensures the index existence whatever the $\theta^{(i,j)}$ distribution is. Skipping this step for $\hat{R}$ yields an explicit expression of what is estimated in the stationary case with (1). This makes the diagnostic more interpretable and allows us to obtain key theoretical results for the associated theoretical R and R_∞.

3 Illustrations

In this section, we consider toy distributions for the chains, where the computation of R_∞ can be done explicitly. In particular, we first focus on two cases raised by Vehtari

et al. (2021) of deficient behaviour of the traditional $\hat{R}$. Then, we exhibit a situation of discrepancy of rank-$\hat{R}$, the updated version of Vehtari et al. (2021). All these theoretical behaviours are illustrated on a small simulation study.

Example 1: chains with same mean and different variances. To tackle the first situation of poor behaviour of the traditional $\hat{R}$, we consider m chains following centered uniform distributions with different variances. More specifically, assume that the $m-1$ first chains have the cdf $F_1 = \cdots = F_{m-1}$ of the uniform distribution $\mathcal{U}(-\sigma, \sigma)$ while the last chain has the cdf F_m of the uniform distribution $\mathcal{U}(-\sigma_m, \sigma_m)$ with $0 < \sigma \leq \sigma_m$. In such a case, the between-variance is zero and it is thus expected that $\hat{R} \approx 1$. In contrast, R_∞ can be written

$$R_\infty = \sqrt{1 + \frac{m-1}{m}\left(1 - \frac{2}{1+\sigma_m/\sigma}\right)}.$$

It appears that R_∞ is an increasing function of σ_m/σ starting from $R_\infty = 1$ when $\sigma_m/\sigma = 1$, and upper bounded by $\sqrt{2 - 1/m}$ when $\sigma_m/\sigma \to \infty$. Results are illustrated in the first column of Figure 1. In the third row, the histograms of replications confirm that $\hat{R}_\infty$ is able to spot the same convergence issue as the one Vehtari et al. (2021) suggests.

Example 2: chains with heavy-tails and different locations. As a second example of poor behaviour of $\hat{R}$, we consider chains following Pareto(α, η) distributions, with cdf

$$F(x \mid \alpha, \eta) = 1 - (x/\eta)^{-\alpha}, \quad \forall x \in [\eta, +\infty),$$

shape parameter $\alpha > 0$ and lower bound $\eta > 0$. Let us recall that such a distribution is heavy-tailed (Embrechts et al., 2013, Table 3.4.2) and has a finite first moment when $\alpha > 1$. We focus on the case where one chain is shifted from the other ones: $F_1(x) = \cdots = F_{m-1}(x) = F(x \mid \alpha, \eta)$ and $F_m(x) = F(x \mid \alpha, \eta_m)$ with $0 < \eta \leq \eta_m$ and $0 < \alpha \leq 1$. Here, the within- and between-variances do not exist and it is expected in practice that $\hat{R} \approx 1$. In contrast, R_∞ can be written as

$$R_\infty = \sqrt{1 + \frac{1}{m}\left(\left(\frac{\eta_m}{\eta}\right)^{\alpha} - 1\right)}.$$

Clearly, R_∞ is an increasing function of η_m/η starting from $R_\infty = 1$ when $\eta_m = \eta$ and such that $R_\infty \to \infty$ as $\eta_m/\eta \to \infty$. Results are shown in the second column of Figure 1. This experiment corresponds to the second example of convergence issue raised by Vehtari et al. (2021). The same observations as for Example 1 can be made here: $\hat{R}_\infty$ is more prone to indicating a convergence issue than rank-$\hat{R}$.

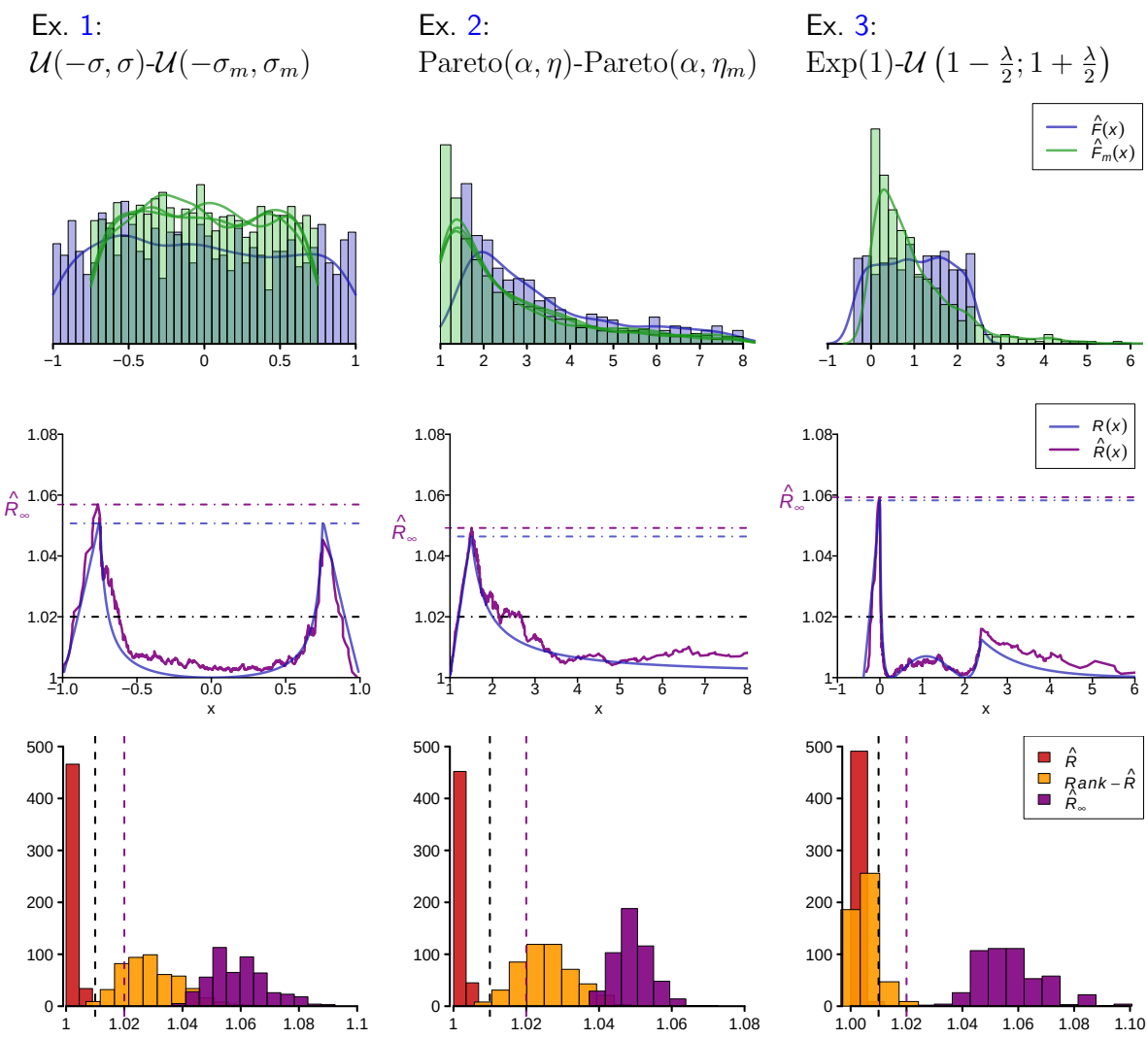


Figure 1: Illustrations with $m = 4$ chains, $n = 200$ independent iterations each. Top row: Empirical distributions $F_1 = \cdots = F_{m-1}$ in green distinct from F_m in blue. For the uniform example (left), $\sigma = 3/4$ and $\sigma_m = 1$, for the Pareto (middle) $\eta = 1$ and $\eta_m = 1.5$, and for the uniform with Laplace (right) $\sigma = 1/4$. Second row: The corresponding population version $R(x)$ and empirical version $\hat{R}(x)$ as function of x for one replication. Bottom row: Histograms of 500 replications of $\hat{R}$, rank-$\hat{R}$ and $\hat{R}_\infty$. The dashed lines correspond to the threshold of 1.01 for $\hat{R}$ and rank-$\hat{R}$ and 1.02 for $\hat{R}_\infty$.

Example 3: chains with same mean and mean over the median. Finally, we look at an example where both $\hat{R}$ and rank-$\hat{R}$ fail to detect non-convergence. Consider $m-1$ exponential chains Exp(1) and one uniform $\mathcal{U}(1-2\log 2, 1+2\log 2)$. This results in chains with same mean and mean over the median, which by construction should fool rank-$\hat{R}$. Results are illustrated in the third row of Figure 1: the histograms of replications confirm that $\hat{R}_\infty$ is able to spot the convergence issue that both $\hat{R}$ and rank-$\hat{R}$ do not.

In the communication, we shall also present applications to other models in a more practical case for Bayesian inference.

References

Embrechts, P., Klüppelberg, C., and Mikosch, T. (2013). *Modelling extremal events*, volume 33. Springer Science & Business Media.

Gelman, A. and Rubin, D. B. (1992). "Inference from iterative simulation using multiple sequences." *Statistical Science*, 7(4), 457–472.

Moins, T., Arbel, J., Dutfoy, A., and Girard, S. (2021). "Contributed discussion: Rank-Normalization, Folding, and Localization: An Improved $\widehat{R}$ for Assessing Convergence of MCMC." *Bayesian Analysis*, 16(2), 711–712.

— (2022). "On the use of a local $\hat{R}$ to improve MCMC convergence diagnostic." *Preprint*. URL https://hal.inria.fr/hal-03600407

Raftery, A. E. and Lewis, S. (1992). "How Many Iterations in the Gibbs Sampler?" *Bayesian Statistics*, 4, 763–773.

Roy, V. (2020). "Convergence diagnostics for Markov chain Monte Carlo." *Annual Review of Statistics and Its Application*, 7, 387–412.

Vats, D. and Flegal, J. M. (2021). "Invited discussion: Rank-normalization, folding, and localization: An improved $\widehat{R}$ for assessing convergence of MCMC." *Bayesian Analysis*, 16(2), 695–701.

Vats, D. and Knudson, C. (2021). "Revisiting the Gelman–Rubin Diagnostic." *Statistical Science*, 36(4), 518 – 529.

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., and Bürkner, P.-C. (2021). "Rank-normalization, folding, and localization: An improved $\widehat{R}$ for assessing convergence of MCMC (with discussion)." *Bayesian Analysis*, 16(2), 667–718.

Statistique mathématique – Valeurs Manquantes

Vitesse de convergence pour une prediction linéaire en présence de données manquantes

Alexis Ayme [1], Claire Boyer [1], Aymeric Dieuleveut [2] & Erwan Scornet [2]

[1] *Sorbonne Université, CNRS, Laboratoire de Probabilités, Statistique et Modélisation (LPSM), F-75005 Paris, France,* {alexis.ayme,claire.boyer}@sorbonne-universite.fr
[2] *CMAP, UMR7641, Ecole Polytechnique, IP Paris, 91128 Palaiseau, France,* {aymeric.dieuleveut,erwan.scornet}@polytechnique.edu

Résumé. Les valeurs manquantes apparaissent dans la plupart des bases de données. Leur nature même empêche généralement d'exécuter les algorithmes classiques d'apprentissage supervisé. Dans cet article, nous étudions les modèles linéaires largement étudiés, mais qui en présence de valeurs manquantes, s'avère être une tâche difficile. En effet, cela nécessite finalement de résoudre un nombre de tâches d'apprentissage, exponentiel par rapport au nombre de variables, ce qui rend les prédictions impossibles sur de vraies bases de données. Nous établissons une première borne non-asymptotique sur l'excès de risque d'un estimateur de type moindres carrés, qui croît exponentiellement avec la dimension. Nous proposons ensuite un nouvel algorithme et une borne qui s'adapte à la distribution des valeurs manquantes et qui, de plus, s'avère être optimale au sens minimax. Des expériences numériques mettent en évidence les avantages de notre méthode par rapport aux algorithmes fréquemment utilisés pour la prédiction avec valeurs manquantes.

Mots-clés. Régression linéaire, données manquantes, apprentissage supervisé.

Abstract. Missing values arise in most real-world data sets Their very nature usually prevents us from running standard learning algorithms. In this paper, we focus on the extensively-studied linear models, but in presence of missing values, which turns out to be quite a challenging task. Indeed, a naive approach amounts to solve a number of learning tasks exponential in the number of input features, which makes predictions impossible for current real-world datasets. First, we establish a bound on the excess risk of a least-squares type estimator which increases exponentially in the dimension. Then, we leverage the missing data distribution to propose a new algorithm, and derive associated adaptive risk bounds that turn out to be minimax optimal. Numerical experiments highlight the benefits of our method compared to state-of-the-art algorithms used for predictions with missing values.

Keywords. Linear regression, missing values, supervised learning.

1 Introduction

Context Missing values are more and more present as the size of datasets increases. These missing values can occur for a variety of reasons, such as sensor failures, refusals

to answer poll questions, or aggregations of data coming from different sources (with different methods of data collection).

If theoretical developments have been mainly done for imputation tasks, there is actually little work on prediction with missing values. Pelckmans et al. [2005] adapt the SVM classifier to the case of missing values. Josse et al. [2019] study the consistency of imputation strategies prior to non-parametric learning methods. Prediction under linear models has been studied in [Le Morvan et al., 2020a,b], by exploiting the peculiar pattern-by-pattern structure of the Bayes predictor (i.e. decomposable into predictors specific to each missing pattern), and estimating it when the input variables are assumed to be Gaussian. Le Morvan et al. [2020b] obtain risk bounds, that suffer from the curse of dimensionality, and are actually not compatible with their Gaussian assumption.

Contributions In this paper, we study pattern-by-pattern predictors for regression with missing input variables. First, we provide a synthetic overview of all assumptions that allow to obtain a pattern-wise linear Bayes predictor, and we propose a detailed study on how these assumptions are related (Section 2). Second, we provide a distribution-free excess risk bound for a least-square estimator handling unbounded features (Section 3), but suffering from the curse of dimensionality. We therefore introduce a novel thresholded estimator for which we establish an excess risk bound adaptive to the missing pattern distribution (Section 4). The latter actually applies to all types of missing data (MCAR, MAR, MNAR) and is shown to be minimax optimal.

2 Typology of missing value and its consequence on the Bayes predictor

Setting In a context of regression, we observe $n \in \mathbb{N}$ input/output observations $(X_i, Y_i)_{i\in[n]}$, i.i.d. copies of a generic pair $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$, assuming that the underlying model linking Y to X is linear.

Assumption 1 (Linear Model)**.** $Y = \beta_0 + \beta^\top X + \epsilon$, with a Gaussian noise $\epsilon \sim \mathcal{N}(0, \sigma^2)$ independent of X.

The (unknown) model parameters are therefore $(\beta_0, \beta) \in \mathbb{R}^{d+1}$. Although standard linear regression is a well-understood problem in statistics, we consider here that only a fraction of the components of X is available: to the data $X \in \mathbb{R}^d$ one associates the missing values pattern $M \in \{0, 1\}^d$, such that $M_j = 1$ if and only if X_j is missing. Let $\mathcal{M} = \{0, 1\}^d$ be the set of missing values patterns. For $m \in \mathcal{M}$, we denote by obs(m) (resp. mis(m)) the set of indexes of the observed variables (resp. the missing variables) and $X_{\text{obs(m)}}$ (resp. $X_{\text{mis(m)}}$) the vector of observed components (resp. unobserved components) of X. Thus, under a linear model with missing covariates, our goal is to predict Y given $\big(X_{\text{obs(m)}}, M\big)$, denoted Z in the sequel.

Bayes predictor The Bayes predictor for the quadratic loss can be decomposed according to the possible missing data patterns, as

$$\begin{aligned} f^\star(Z) &= \mathbb{E}\left[Y|Z\right] = \mathbb{E}\left[Y|X_{\text{obs(m)}}, M\right] \\ &= \sum_{m\in\mathcal{M}} f_m^\star(X_{\text{obs(m)}})\mathbb{1}_{M=m}, \end{aligned}$$

where $f_m^\star(X_{\text{obs(m)}}) := \mathbb{E}\left[Y|X_{\text{obs(m)}}, M=m\right]$ can be seen as the Bayes predictor conditionally on the event "$M=m$".

Data scenarios There exist two main approaches for modelling the joint distribution of X and M: selection models [Heckman, 2012] (including the classical MCAR, MAR, MNAR mechanisms) and pattern-mixture ones [Little, 1993].

Assumption 2 (Gaussian Pattern Mixture Model-GPMM)**.** For all $m \in \mathcal{M}$, $X|(M = m) \sim \mathcal{N}(\mu^{(m)}, \Sigma^{(m)})$.

Linearity of the Bayes predictor In this subsection, we give an overview of the properties that ensure the linearity of $f_m^\star$.

Definition 2.1. Consider the vector space of linear predictors in the observed variables i.e. $f \in \mathcal{F}_b$ if $f(., m)$ is linear for all $m \in \mathcal{M}$. The dimension of $\mathcal{F}_b$ is $p := 2^{d-1}(d+2)$.

Proposition 2.2. *Assume the Gaussian Pattern Mixture Model (Assumption 2). Then $f^\star \in \mathcal{F}_b$ i.e. for all $m \in \mathcal{M}$ there exist $\delta_0^{(m)} \in \mathbb{R}$ and $\delta^{(m)} \in \mathbb{R}^{|m|}$ such that*

$$f_m^\star(X_{\text{obs(m)}}) = \delta_0^{(m)} + \left(\delta^{(m)}\right)^\top X_{\text{obs(m)}}.$$

Note that this proposition also holds for Gaussian covariates with M(C)AR mechanisms. The proposition highlights that there is a wide variety of possible assumptions such that $f^\star \in \mathcal{F}_b$. In particular, 2 may not only include MAR but MNAR scenarios as well, the latter known to be challenging in an inference setting. However, this comes at the cost that the dimension p of $\mathcal{F}_b$ grows exponentially with the ambient dimension d ($p = 2^{d-1}(d+2)$).

3 A distribution-free bound on excess risk

In the framework of missing values, let the excess risk be

$$\mathcal{E}\left(\widehat{f}\right) := \mathbb{E}\left[\left(\widehat{f}(Z) - f^\star(Z)\right)^2 \middle| \mathcal{D}_n\right], \tag{1}$$

and its integrated version $\mathbb{E}[\mathcal{E}(\widehat{f})]$. This quantity measures the quality of performance made by a prediction function $\widehat{f}$ compared to the optimal predictor $f^\star$.

Consider the set $\mathcal{K}_D := \left\{(x_{\text{obs(m)}}, m), \quad \left\|x_{\text{obs(m)}}\right\|_\infty \leq D\right\}$, for some $D > 0$, which consists in taking the covariates with all observed components in an ℓ^∞-ball of radius D.

For instance, with sub-Gaussian covariates, an observation Z falls into the bounded set $\mathcal{K}_D$ with high probability. One can then adapt the results in Györfi et al. [2006], Audibert and Catoni [2011] when X is on a bounded set to the sub-Gaussian case. To do so, consider the modified least-squares (D-LS) estimator taking into account only the observations falling into $\mathcal{K}_D$:

$$\hat{f}^{(D\text{-LS})} \in \operatorname*{argmin}_{f \in \mathcal{F}_b} \sum_{Z_i \in \mathcal{K}_D} (f(Z_i) - Y_i)^2 \tag{2}$$

if $\mathcal{K}_D \neq \emptyset$, and $\hat{f}^{(D\text{-LS})} = 0$ otherwise. Computing $\hat{f}^{(D\text{-LS})}$ amounts to perform one ordinary least-square procedure per missing pattern (as $\mathcal{F}_b$ is composed of functions that are linear on each missing pattern). Finally, for technical purposes, to ensure that the prediction is bounded, we consider the *clipped* estimator at level L, $T_L \hat{f} := (-L) \vee \hat{f} \wedge L$.

In the framework of Proposition 2.2, the Bayes predictor is ensured to be linear. This wipes the approximation error and lead to the following Theorem.

Theorem 3.1. *Under Assumption 2, assume the covariates to be γ-sub-Gaussian, with $\gamma = \max_{\substack{m \in \mathcal{M} \\ j \in [d]}} \mathbb{E}\left[X_j^2 | M = m\right]$. Choosing $D = \sqrt{\gamma}(1 + \sqrt{\gamma \log(n)})$, and $L = (D+1)(B+1)$ with $B = \max_{m \in \mathcal{M}} \max[|\delta_0^{(m)}|, \|\delta^{(m)}\|_1]$, we have*

$$\mathbb{E}\left[\mathcal{E}\left(T_L \hat{f}^{(D\text{-LS})}\right)\right] \lesssim (\log(n) + 1)\left(\sigma_{na}^2 \vee L^2\right) 2^d \frac{d}{n}.$$

This theorem is the first theoretical result that provides a control on the excess risk of a least-square-type predictor. The obtained upper bound is the multiplication of three terms. The first factor $(\log(n) + 1)$ is due to [Györfi et al., 2006, Theorem 11.3] on which our result is built upon. The second factor $\sigma_{\text{na}}^2 \vee L^2$ should be seen as a tight bound for $\mathbb{E}[Y^2]$, which corresponds to the risk of the trivial predictor (predicting 0 for any value of Z). Note that the coefficient L logarithmically depends on n: the truncation of the predictor should be less stringent with an increasing number of observations. The rate of convergence is eventually dictated by the factor $2^d \frac{d}{n}$, which remains problematic as it grows exponentially with the dimension. It reflects the fact that a different regression model is required for each missing value pattern. Overall, the bound ensures that when $n > d2^d$, the least-square predictor is better than the zero one. This curse of dimensionality is the price to pay as the result is valid for a wide variety of missing pattern distribution.

4 Main result: an excess risk bound adaptive to the missing pattern distribution

The error bound obtained in Theorem 3.1 holds for very general assumptions on the missing pattern distribution. However, this bound is pessimistic when some missing

patterns are not observed or, more generally, when the missing pattern distribution is non-uniform, i.e. of low entropy. In this section, we leverage the distribution of the missing patterns in order to derive better theoretical bounds compared to Theorem 3.1. To this end, we propose a refined version of the predictor introduced in Equation (2).

Regression only on high frequency missing patterns For any missing pattern $m \in \mathcal{M}$, we denote $E_m = \{i \in [n], M_i = m\}$ and $\mathcal{D}_n^{(m)} = ((X_{i,\mathrm{obs(m)}}, Y_i))_{i \in E_m}$ respectively the observation indices and the sub-sample with missing pattern m. For any $m \in \mathcal{M}$, we build an estimator $\widetilde{f}_m$ of $f_m^\star$ as

$$\widetilde{f}_m \in \arg\min_{f \in \mathcal{F}_m} \sum_{i \in \overline{E}_m} (f(X_i) - Y_i)^2, \tag{3}$$

if $\overline{E}_m := \left\{ i \in E_m, \left\| X_{\mathrm{obs(m)}} \right\|_\infty \leq D \right\}$ is non-empty, and $\widetilde{f}_m = 0$ otherwise. The global predictor is then obtained by combining the previous pattern-by-pattern predictors for all patterns $m \in \mathcal{M}$ that appear with a frequency $\widehat{p}_m := \frac{|E_m|}{n}$ larger than a threshold $\tau \in [0, 1]$,

$$\widehat{f}^{(\tau)}(Z) = \sum_{m \in \mathcal{M}} \widetilde{f}_m \left(X_{\mathrm{obs(m)}} \right) \mathbb{1}_{\widehat{p}_m > \tau} \mathbb{1}_{M=m}. \tag{4}$$

Contrary to the naive estimator $\hat{f}^{(D\text{-LS})}$ defined in (2), computing $\widehat{f}^{(\tau)}$ may not require to perform up to 2^d linear regressions. This new predictor (4) enjoys the following risk bounds.

Theorem 4.1. *Under the same assumptions as in Theorem 3.1, for any $\tau \geq 1/n$, the generalization bound for the predictor $\widehat{f}^{(\tau)}$ defined in* (4)*, reads as*

$$\mathbb{E}\left[\mathcal{E}\left(T_L \widehat{f}^{(\tau)} \right) \right] \lesssim (\log(n) + 1)\left(\sigma_{na}^2 \vee L^2\right) \left(1 \vee \frac{d}{n\tau} \right) \mathfrak{C}_p(\tau). \tag{5}$$

with the missing patterns distribution complexity $\mathfrak{C}_p(\tau)$ *defined by*

$$\mathfrak{C}_p(\tau) := \sum_{m \in \mathcal{M}} p_m \wedge \tau. \tag{6}$$

The upper bound in Inequality (5) *is minimal for the choice $\tau = d/n$ which leads to*

$$\mathbb{E}\left[\mathcal{E}\left(T_L \widehat{f}^{(d/n)} \right) \right] \lesssim (\log(n) + 1)\left(\sigma_{na}^2 \vee L^2\right) \mathfrak{C}_p\left(\frac{d}{n} \right). \tag{7}$$

Theorem 4.1 is the first result controlling the excess risk of a pattern-by-pattern least-square-type predictor with a bound depending on the missing pattern distribution through the complexity $\mathfrak{C}_p$, and actually holds for any type of missing patterns. Theorem 4.1 improves over Theorem 3.1, as the pattern distribution complexity $\mathfrak{C}_p$ is a lower bound of $2^d d/n$. The adaptivity of $\mathfrak{C}_p$ to the missing pattern distribution is illustrated in the following example.

Consider a subset $\mathcal{B} \subset \mathcal{M}$ of small cardinality $|\mathcal{B}|$, so that only missing patterns in $\mathcal{B}$ are very frequent and that the other missing patterns occur with a residual probability $\delta = \mathbb{P}(M \in \mathcal{B}^c)$. We have,

$$\mathfrak{C}_p\left(\frac{d}{n}\right) \leq |\mathcal{B}|\frac{d}{n} + \delta. \tag{8}$$

This bound clearly improves upon Theorem 3.1, as the complexity is now controlled by $|\mathcal{B}|\frac{d}{n}$ instead of $2^d\frac{d}{n}$. This bound reflects the good learning ability of the regressor $\hat{f}^{(d/n)}$ when there are few frequent missing patterns.

References

J.-Y. Audibert and O. Catoni. Robust linear least squares regression. *The Annals of Statistics*, 39(5):2766–2794, 2011.

L. Györfi, M. Kohler, A. Krzyzak, and H. Walk. *A distribution-free theory of nonparametric regression*. Springer Science & Business Media, 2006.

J. Heckman. The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *NBER Book Chapters*, 5, 02 2012.

J. Josse, N. Prost, E. Scornet, and G. Varoquaux. On the consistency of supervised learning with missing values. *arXiv preprint arXiv:1902.06931*, 2019.

M. Le Morvan, J. Josse, T. Moreau, E. Scornet, and G. Varoquaux. NeuMiss networks: differentiable programming for supervised learning with missing values. In *NeurIPS 2020 - 34th Conference on Neural Information Processing Systems*, 2020a.

M. Le Morvan, N. Prost, J. Josse, E. Scornet, and G. Varoquaux. Linear predictor on linearly-generated data with missing values: non consistency and solutions. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020b.

R. J. Little. Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, 88(421):125–134, 1993.

K. Pelckmans, J. De Brabanter, J. A. Suykens, and B. De Moor. Handling missing values in support vector machine classifiers. *Neural Networks*, 18(5-6):684–692, 2005.

Imputation multiple dans le modèle linéaire fonctionnel avec une covariable partiellement observée et des valeurs manquantes dans la réponse

Christophe Crambes [1,†] & Chayma Daayeb [1,2,††] & Ali Gannoun [1,‡] & Yousri Henchiri [2,3,*]

[1] *Institut de Mathématiques et de Modélisation de Montpellier, CNRS, Université de Montpellier, France.*
[2] *École nationale d'ingénieurs de Tunis, Laboratoire de Modélisation Mathématique et Numérique dans les Sciences de l'Ingénieur LAMSIN, Tunisie.*
[3] *Institut Supérieur des Arts Multimédia de la Manouba (ISAMM), Tunisie.*
[†] *christophe.crambes@umontpellier.fr,* [††] *chayma.daayeb@umontpellier.fr,*
[‡] *ali.gannoun@umontpellier.fr,* [*] *yousri.henchiri@umontpellier.fr*

Résumé. Les données manquantes sont courantes dans les études de météorologie, de biomécanique, d'économie, de médecine et de sciences sociales et posent souvent un sérieux défi dans l'analyse des données. Les méthodes d'imputation multiple (IM) sont des outils populaires et naturels pour traiter les données manquantes, qui offrent une méthode réalisable sur le plan informatique pour étudier un large spectre de problèmes en présence de données manquantes, en remplaçant chaque valeur manquante par un ensemble de valeurs raisonnables qui représentent une incertitude sur les valeurs à imputer.

Mots-clés. Modèle linéaire fonctionnel, Données manquantes, Courbe partiellement observée, Composantes principales fonctionnelles, Imputation Multiple.

Abstract. Missing data are common in meteorology, biomechanics, economics, medical and social science studies and often raise a serious challenge in data analysis. Multiple imputation (MI) methods are popular and natural tools to deal with missing data, that offer a computationally feasible method to study a wide range of problems under missingness mechanisms by replacing each missing value with a set of reasonable values that represent the uncertainty about the values to be imputed.

Keywords. Functional linear model, Missing data, Partially observed curve, Functional Principal Components, Multiple imputation,

1 Introduction

Ces dernières années, les travaux sur les données fonctionnelles se sont concentrés en grande partie sur les situations où les données sont entièrement observées. Ce n'est pas toujours le cas, et des données manquantes peuvent apparaître, par exemple lorsque un appareil de mesure tombe en panne. Les problématiques de données manquantes restent très répandues dans

les problèmes statistiques et nécessitent un traitement particulier. Ignorer les données manquantes peut entraîner une perte de précision lors de la modélisation de ces données. L'objectif de ce travail est d'apporter une réponse à cette problématique par l'imputation multiple. Cette dernière technique est conçue pour tirer parti de l'informatique moderne pour traiter les données manquantes. Chaque valeur manquante est remplacée par deux ou plusieurs valeurs imputées afin de représenter l'incertitude quant à la valeur à imputer. Nous observons l'échantillon $\mathcal{D}_n \triangleq \big\{(X_1, Y_1), \ldots, (X_n, Y_n)\big\}$, où les (X_i, Y_i) sont indépendants et identiquement distribués de même loi que (X, Y). La fonction aléatoire X prend ses valeurs dans l'espace $\mathcal{L}_2(\mathcal{I})$ des fonctions de carré intégrable définies sur un intervalle $\mathcal{I} \subset \mathbb{R}$ et la réponse réelle Y est générée par le modèle de régression suivant

$$Y = \alpha + \int_{\mathcal{I}} \theta(t)X(t)\mathrm{d}t + \varepsilon. \tag{1.1}$$

Ici, α est une constante, représentant l'intercept du modèle et θ est une fonction appartenant à $\mathcal{L}_2(\mathcal{I})$, représentant la fonction de pente. Nous supposons que les erreurs ε_i sont indépendantes et identiquement distribuées avec une variance finie et une moyenne nulle et sont également indépendantes des variables explicatives X_i.
La méthodologie de régression linéaire fonctionnelle sur composantes principales est fondée sur les décompositions spectrales de la covariance de X et de son estimateur et est construite comme suit : nous définissons l'opérateur de covariance croisée empirique $\widehat{\Delta}_n$ donné par $\widehat{\Delta}_n u = \frac{1}{n}\sum_{i=1}^n \langle X_i, u\rangle Y_i$ pour tout $u \in \mathcal{L}_2(\mathcal{I})$, l'opérateur de covariance empirique $\widehat{\Gamma}_n$ donné par $\widehat{\Gamma}_n u = \frac{1}{n}\sum_{i=1}^n \langle X_i, u\rangle X_i$ pour tout $u \in \mathcal{L}_2(\mathcal{I})$. Soient $(\widehat{\phi}_j)_{j=1}^{k_n}$ les fonctions propres associées à $\widehat{\Gamma}_n$ correspondant à k_n valeurs propres maximales $\widehat{\lambda}_1 > \ldots > \widehat{\lambda}_{k_n} > 0$ (avec k_n entier qui dépend de n). Nous définissons l'opérateur de projection orthogonale $\widehat{\Pi}_{k_n}$ sur le sous espace engendré par $(\widehat{\phi}_1, \ldots, \widehat{\phi}_{k_n})$ par $\widehat{\Pi}_{k_n} u = \sum_{j=1}^{k_n} \langle \widehat{\phi}_j, u\rangle \widehat{\phi}_j$ pour tout $u \in \mathcal{L}_2(\mathcal{I})$. Le but du problème d'estimation, en (1.1), est d'estimer α et θ ainsi de récupérer

$$\eta(X) \triangleq \alpha + \int_{\mathcal{I}} \theta(t)X(t)\mathrm{d}t, \tag{1.2}$$

basé sur un échantillon d'apprentissage $\mathcal{D}_n$. Soit ℓ_n une fonction qui mesure la qualité de l'ajustement de η aux données. Alors, l'estimateur $\widehat{\eta}_n$ de η est donné par

$$\widehat{\eta}_n \triangleq \mathrm{argmin}_{\eta_0} \Big(\ell_n\big(\eta_0 \mid \mathcal{D}_n\big)\Big) \quad \text{où la minimisation est prise sur} \tag{1.3}$$

$$\Big\{\eta_0 \mid \eta_0(X) = \alpha_0 + \textstyle\int_{\mathcal{I}} \theta_0(t)X(t)\mathrm{d}t : \alpha_0 \in \mathbb{R}, \theta_0 \in \mathrm{Span}\big(\widehat{\phi}_1, \ldots, \widehat{\phi}_{k_n}\big)\Big\}.$$

Généralement, ℓ_n est choisie de telle sorte qu'elle soit convexe en η_0 et $\mathbb{E}(\ell_n(\eta_0))$ est uniquement minimisé par η. Le choix le plus courant de la fonction d'ajustement des données est l'erreur quadratique $\ell_n(\eta_0 \mid \mathcal{D}_n) \triangleq \frac{1}{n}\sum_{i=1}^n \Big(Y_i - \eta_0(X_i)\Big)^2$. De manière équivalente, la minimisation peut être prise sur (α_0, θ_0) au lieu de η_0 pour obtenir des estimations à la fois de l'intercept

et de la fonction pente, désignées respectivement par $\widehat{\alpha} = \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ et $\widehat{\theta} = \sum_{j=1}^{k_n} \widehat{\mathsf{s}}_j \widehat{\phi}_j$ avec $\widehat{\mathsf{s}}_j = \frac{1}{n\widehat{\lambda}_j}\sum_{i=1}^{n}\langle X_i, \widehat{\phi}_j\rangle Y_i$.

Nous décrivons à présent les deux types de mécanismes de données manquantes qui feront l'objet de notre travail. Pour le mécanisme de données manquantes sur la réponse réelle, nous considérons une variable aléatoire dichotomique $\delta^{[Y]}$ conduisant à l'échantillon $(\delta_i^{[Y]})_{i=1,\ldots,n}$ tel que $\delta_i^{[Y]} = 1$ si la valeur Y_i est disponible et $\delta_i^{[Y]} = 0$ si la valeur Y_i est manquante, pour tout $i = 1, \ldots, n$. Ici, nous considérons que les données de la réponse sont "missing at random" (MAR) : le fait que la valeur Y soit manquante ne dépend pas de la réponse du modèle, mais peut éventuellement dépendre de la covariable,

$$\mathbb{P}(\delta^{[Y]} = 1 \mid X, Y) = \mathbb{P}(\delta^{[Y]} = 1 \mid X).$$

De façon équivalente, l'hypothèse MAR implique que la distribution de Y est la même pour les unités avec $\delta_i^{[Y]} = 1$ que pour celles qui ont $\delta_i^{[Y]} = 0$, après prise en compte de X. En conséquence de cette hypothèse MAR, la variable $\delta^{[Y]}$ (le fait qu'une observation est manquante) est indépendante de l'erreur du modèle ε. Dans la suite, le nombre de valeurs manquantes parmi les $Y_1, \ldots, Y_n$ est désigné par

$$m_n^{[Y]} = \sum_{i=1}^{n} \mathbb{1}_{\{\delta_i^{[Y]}=0\}}.$$

Pour le mécanisme de données manquantes sur la covariable fonctionnelle, nous adoptons le paradigme des fonctions partiellement observées comme dans [3] et [4]. Plus précisément, pour chaque courbe X_i, $i = 1, \ldots, n$, nous considérons la partie observée $O_i \subseteq \mathcal{I}$ de X_i et la partie manquante $M_i = \mathcal{I} \setminus O_i$. La partie observée O_i fait référence à un intervalle (ou plusieurs intervalles) où la courbe X_i est observée en certains points de mesure de O_i. Basée sur des observations ponctuelles, la courbe entière peut être reconstruite sur O_i avec les méthodes habituelles (par exemple splines de lissage, splines de régression, régression locale ou régression polynomiale locale). Cependant, aucune information n'est disponible sur la partie manquante M_i. Dans la suite du document, nous écrirons "O" et "M" pour désigner une production donnée de O_i et M_i. De plus, nous indiquons les parties observées et manquantes de X_i par X_i^O et X_i^M.

L'objectif est de prédire une nouvelle valeur de la réponse Y étant donné une nouvelle observation test sur la variable explicative X, une fois que les courbes partiellement observées X sont reconstruites et les données manquantes Y sont imputées à l'aide d'une méthode d'imputation aléatoire.

2 Imputation Multiple

2.1 Imputation déterministe

Récemment, le cas de données manquantes dans la régression linéaire fonctionnelle lorsque la covariable est partiellement observée et que la réponse est affectée par des données man-

quantes a été étudié dans [2]. L'estimateur $\widehat{\eta}_n$ de η dans (1.3) peut être calculé sur la base des courbes complètes reconstruites en adoptant la méthodologie de reconstruction des courbes présenté dans [3]. En ce qui concerne les données manquantes sur la réponse, une imputation déterministe par régression est utilisée, comme dans [1]. Nous considérons d'abord le problème de reconstruction des courbes explicatives reliant la partie manquante des courbes à la partie observée, en écrivant

$$X_i^M(s) = \mathcal{L}(X_i^O(t)) + \mathcal{Z}_i(s),$$

pour tout $t \in O$ et $s \in M$, où $\mathcal{L} : \mathcal{L}_2(O) \to \mathcal{L}_2(M)$ est un opérateur linéaire de reconstruction et $\mathcal{Z}_i \in \mathcal{L}_2(M)$ est l'erreur de reconstruction. Suivant [3], l'opérateur de reconstruction linéaire optimal, minimisant le risque suivant $\mathbb{E}\Big(\Big(X_i^M(u) - \mathcal{L}(X_i^O)(u)\Big)^2\Big)$, pour tout $u \in M$, donné par $\widetilde{\mathcal{L}}(X_i^O)$, est estimé par $\widehat{\mathcal{L}}_{\tilde{k}_n}(X_i^O)$, où $\tilde{k}_n$ est un nombre entier positif, appelé *paramètre de troncature*. Considérant maintenant l'échantillon suivant $\widetilde{\mathcal{D}}_n \triangleq \big\{(X_1^\star, \delta_1^{[Y]}, Y_1), \ldots, (X_n^\star, \delta_n^{[Y]}, Y_n)\big\}$, avec des courbes explicatives éventuellement reconstruites

$$X_i^\star(t) = \begin{cases} X_i^O(t), & \text{if} \quad t \in O, \\ \widehat{\mathcal{L}}_{\tilde{k}_n}(X_i^O)(t), & \text{if} \quad t \in M, \end{cases}$$

l'échantillon $\widetilde{\mathcal{D}}_n$ est divisé en deux sous-ensembles : $\widetilde{\mathcal{D}}_{resp}$ est l'ensemble des variables d'intérêt observées et $\widetilde{\mathcal{D}}_{noresp}$ est l'ensemble des variables d'intérêt non observées, avec $\widetilde{\mathcal{D}}_n = \widetilde{\mathcal{D}}_{resp} \cup \widetilde{\mathcal{D}}_{noresp}$. Nous définissons l'opérateur de covariance avec les courbes reconstruites par $\widehat{\Gamma}_{n,rec}^{obs} = \frac{1}{n - m_n^{[Y]}} \sum_{i=1}^n \langle X_i^\star, . \rangle \delta_i^{[Y]} X_i^\star$. Soit $\widehat{\Pi}_{k_n,rec}^{obs}$ l'opérateur de projection sur le sous-espace $\text{Span}(\widehat{\phi}_{1,rec}^{obs}, \ldots, \widehat{\phi}_{k_n,rec}^{obs})$ où $\widehat{\phi}_{1,rec}^{obs}, \ldots, \widehat{\phi}_{k_n,rec}^{obs}$ sont les k_n premières fonctions propres de l'opérateur de covariance $\widehat{\Gamma}_{n,rec}^{obs}$. Dans ce cas, l'estimateur $\widetilde{\eta}_n$ de η est donné par

$$\widetilde{\eta}_n \triangleq \text{argmin}_{\widetilde{\eta}_0} \Big(\widetilde{\ell}_n\big(\eta_0 \mid \widetilde{\mathcal{D}}_n\big)\Big), \quad \text{où la minimisation se fait sur}$$

$$\Big\{\eta_0 \;\mid\; \eta_0(X) = \alpha_0 + \textstyle\int_{\mathcal{I}} \theta_0(t) X(t) \mathrm{d}t \;:\; \alpha_0 \in \mathbb{R}, \theta_0 \in \text{Span}\big(\widehat{\phi}_{1,rec}^{obs}, \ldots, \widehat{\phi}_{k_n,rec}^{obs}\big)\Big\},$$

avec $\widetilde{\ell}_n(\eta_0 \mid \widetilde{\mathcal{D}}_n) \triangleq \frac{1}{n - m_n^{[Y]}} \sum_{i=1}^n \delta_i^{[Y]} \Big(Y_i - \eta_0(X_i^\star)\Big)^2$. De manière équivalente, les estimateurs de l'intercept et de la fonction pente, désignées respectivement par $\widetilde{\alpha}$ et $\widetilde{\theta}$, sont comme suit

$$\widetilde{\alpha} = \overline{Y}_{obs} = \frac{1}{n - m_n^{[Y]}} \sum_{i=1}^n \delta_i^{[Y]} Y_i, \;\; \widetilde{\theta} = \sum_{j=1}^{k_n} \widetilde{\mathsf{s}}_j \widehat{\phi}_{j,rec}^{obs}, \;\; \text{avec } \widetilde{\mathsf{s}}_j = \frac{1}{(n - m_n^{[Y]}) \widehat{\lambda}_{j,rec}^{obs}} \sum_{i=1}^n \langle X_i^\star, \widehat{\phi}_{j,rec}^{obs} \rangle \delta_i^{[Y]} Y_i.$$

Si nous considérons une valeur manquante Y_ℓ sur la réponse, telle que $\delta_\ell^{[Y]} = 0$, nous définissons la valeur imputée $Y_{\ell,imp}$ par

$$Y_{\ell,imp} = \widetilde{\eta}_n(X_\ell^\star) \triangleq \widetilde{\alpha} + \sum_{j=1}^{k_n} \widetilde{\mathsf{s}}_j \langle X_\ell^\star, \widehat{\phi}_{j,rec}^{obs} \rangle.$$

La précision de l'imputation peut être mesurée par le risque,

$$\mathcal{E}(\widetilde{\eta}_n) \triangleq \mathbb{E}^\star\big(\widetilde{\eta}_n(X_\ell^\star) - \eta(X_\ell^\star)\big)^2 = \mathbb{E}^\star\Big(\widetilde{\alpha} + \langle\widetilde{\theta}^\star, X_\ell^\star\rangle - \alpha - \langle\theta, X_\ell^\star\rangle\Big)^2 \tag{2.1}$$

où $(X_\ell^\star, , \delta_\ell^{[Y]}, Y_\ell)$ est une copie de $(X^\star, \delta^{[Y]}, Y)$ et $\mathbb{E}^\star$ représente l'espérance par rapport à $X_\ell^\star$ et Y_ℓ. Des résultats concernant l'étude asymptotique de l'erreur d'imputation sont présentés dans [2].

2.2 Imputation aléatoire

Nous utilisons l'imputation par régression avec résidu standardisé observé, pour $s = 1, \ldots, q$, donnée par

$$\widetilde{Y}_\ell^{(s)} \triangleq \widetilde{\eta}_n(X_\ell^\star) + \varepsilon_\ell^{\star(s)} = \widetilde{\alpha} + \langle\widetilde{\theta}, X_\ell^\star\rangle + \varepsilon_\ell^{\star(s)}, \quad \ell \in \widetilde{\mathcal{D}}_{noresp}, \tag{2.2}$$

où $\varepsilon_\ell^\star$ est un bruit aléatoire associé à l'élément $\ell \in \widetilde{\mathcal{D}}_{noresp}$, tiré au hasard, généralement avec remise, parmi tous les résidus standardisés observés. Plus précisément, $\varepsilon_\ell^\star$ est tiré avec remise dans l'ensemble

$$\Big\{e_i \mid e_i = \widetilde{e}_i - \bar{e}, \;\; i \in \widetilde{\mathcal{D}}_{resp}\Big\}$$

avec $\widetilde{e}_i = \widetilde{\sigma}^{-1}(\widetilde{\alpha} + \langle\widetilde{\theta}, X_i^\star\rangle - Y_i))$, $\widetilde{\sigma} = \frac{1}{n - m_n^{[Y]}} \sum_{i=1}^n \delta_i^{[Y]}(\widetilde{\alpha} + \langle\widetilde{\theta}, X_i^\star\rangle - Y_i)^2$
et $\bar{e} = \frac{1}{n - m_n^{[Y]}} \sum_{i=1}^n \delta_i^{[Y]} \widetilde{e}_i$. Nous posons alors

$$Y_i^{\star(s)} = \delta_i^{[Y]} Y_i + (1 - \delta_i^{[Y]})\widetilde{Y}_i^{(s)}, \quad \text{pour tout } i = 1, \ldots, n.$$

Les tirages aléatoires dans (2.2) sont de nature non paramétrique car les imputations sont produites à partir de la distribution empirique des résidus standardisés observés et aucune distribution paramétrique (telle que la distribution normale) n'est imposée. Par conséquent, on peut envisager tout type de distribution de résidus (asymétrique, multimodale, ...).

3 Estimation de η et prédiction

Nous disposons de q échantillons $\mathcal{D}_n^{\star(s)} \triangleq \big\{(X_1^\star, Y_1^{\star(s)}), \ldots, (X_n^\star, Y_n^{\star(s)})\big\}$, avec $s = 1, \ldots q$, une fois que la base de données a été reconstruite avec q imputations aléatoires. Nous estimons, pour $s = 1, \ldots, q$, l'intercept et la fonction pente du modèle (1.1) sur la base de l'échantillon $\mathcal{D}_n^{\star(s)}$, par

$$\widehat{\alpha}^{(s)} = \frac{1}{n}\sum_{i=1}^n Y_i^{\star(s)} \quad \text{et} \quad \widehat{\theta}^{(s)} = \sum_{j=1}^{k_n} \widehat{\mathsf{s}}_j^{(s)} \widehat{\phi}_{j,rec}^\star \quad \text{avec} \quad \widehat{\mathsf{s}}_j^{(s)} = \frac{1}{n\widehat{\lambda}_{j,rec}^\star} \sum_{i=1}^n \langle X_i^\star, \widehat{\phi}_{j,rec}^\star\rangle Y_i^{\star(s)}.$$

Pour une nouvelle courbe X_{new}, la réponse est prédite par

$$\widehat{Y}_{new} = \frac{1}{q}\sum_{s=1}^{q} \widehat{Y}^{\star}_{new}{}^{(s)} \quad \text{avec} \quad \widehat{Y}^{\star}_{new}{}^{(s)} = \hat{\alpha}^{(s)} + \langle \widehat{\theta}^{(s)}, X^{\star}_{new} \rangle.$$

Une étude asymptotique de l'erreur de prévision de Y_{new} par $\widehat{Y}_{new}$ a été réalisée.

Références

[1] Crambes, C., Henchiri, Y. (2019) *Regression imputation in the functional linear model with missing values in the response*. Journal of Statistical Planning and Inference 201, 103-119.

[2] Crambes, C., Daayeb, C., Gannoun, A., Henchiri, Y. (2021). *Functional linear model with partially observed covariate and missing values in the response*. Preprint at https ://hal.archives-ouvertes.fr/hal-03083293v2.

[3] Kneip, A., Liebl, D. (2020). *On the optimal reconstruction of partially observed functional data*. The Annals of Statistics 48, 1692-1717.

[4] Kraus, D. (2015). *Components and completion of partially observed functional data*. Journal of the Royal Statistical Society : Series B 77, 777-801.

[5] Ramsay, J.O., Silverman, B.W. (2005). *Functional Data Analysis* (Second Edition). Springer, New York.

Functional Data Analysis : Theoretical and pratical effects of surrogate outcomes on the regression linear model

Firas IBRAHIM, Mustapha RACHDI, Jacques DEMONGEOT

E-mail: firas.ibrahim@univ-ubs.fr

Résumé: Dans ce travail, on traite le probème de l'estimation non paramétrique de la régression d'une variable de réponse scalaire incomplète, mais ayant des valeurs de substitution (surrogate data), à partir d'une variable aléatoire fonctionnelle. Nous construisons un estimateur de l'opérateur de régression en utilisant, en plus des valeurs disponibles (vraies) de la variable réponse, les valeurs des données de substitution. Nous avons établi ensuite des propriétés asymptotiques de l'estimateur construit, en termes de convergence en probabilité, presque complète et en moyenne quadratique. Notons que les résultats obtenus ici, dans le cadre fonctionnel, généralisent une partie des résultats obtenus dans le cadre de dimension finie. Puis, nous avons appliqué les résultats obtenus sur des données simulées et nous avons appliqué notre méthodologie sur des données médicales (signaux respiratoires) afin de prédire le PaO2. De plus, nous avons ainsi démontré la supériorité de notre estimateur sur les estimateurs classiques, lorsque nous manquons de données complètes, en procédant à une comparaison des estimations, quand (i) les observations sont totalement complètes, (ii) à partir de données avec valeurs d'imputation et enfin (iii) à partir des données de substitution.

Abstract : In this work, we deal with the problem of the non-parametric estimation of the regression operator of an incomplete scalar response variable but having substitution values (surrogate data), from a functional random variable. We started by constructing an estimator of the regression operator by replacing the missing responses with the surrogate data. We then established asymptotic properties of the constructed estimator, in terms of convergence in probability, almost completely and in mean squares. Then, we applied the obtained results on simulated data and then we applied our methodology to some real data using surrogate medical data (respiratory signals) to predict PaO2. In addition, we compared the estimators when (i) the observations are completely observed, (ii) by using data with imputation values, and finally (iii) from surrogate data.

Keywords : functional data analysis, surrogate data, regression estimation.

Recently, the statistical analysis of functional random variables (FDA) has attracted considerable interest. In fact, with advances in modern technologies, more and more data is continuously recorded over a period of time or, intermittently, at several distinct times. The data collected are examples of functional data. Overall, the FDA deals with the analysis and theory of complex data like functions or curves, surfaces, images, ... For an overview on this topic, we can cite, for example, Ramsay and Silverman (2002), Ferraty and Vieu (2006), Horvath and Kokoszka (2012), Cuevas (2014), Zhang (2014), Bongiorno et al. (2014), Hsing and Eubank (2015), Goia and Vieu (2015) and Wang et al. (2016) and the references therein.

We are interested in the following model:

$$Y = m(X) + \varepsilon, \tag{0.1}$$

where m is the regression operator, X is a functional covariate which belongs to a semi-metric space (E, d), Y is the response variable, and ε is a random error.

The problem we are addressing in this work, the unavailability of some data in the response variable, can be motivated both from a practical and a theoretical point of view. In fact, it may be difficult or expensive to accurately measure some response observations Y.Instead, some substitution observations, usually available through relatively simple measurement methods, can be obtained and used. This problem has received less attention in FDA, mainly because standard methodology could not be used to handle this case and the treatment requires some technical tools which are not available in the high dimensional statistics framework. These methods, however, may be sensitive to the assumed models. Our goal is then to improve the modeling by filling/recovering some of the information missed in the response variable with this surrogate variable. In this case, one solution is to use the help of validation data to capture the underlying relation between the true variables and surrogate ones.

To estimate the regression operator for surrogate data and functional covariates, we adopt an approach based on validation data ideas. In fact, the idea is to introduce the information contained in both the validation data and the brut data (the primary data), which includes surrogate data and the corresponding observations of the covariate X.

Application to respiration data

We attempt to predict the scalar response variable Y (PaO_2) from a functional variable X (thoracic volume). We compare the estimators in three different cases: just data with known true observations, data with imputation values, and finally known true data which are completed with surrogate data.

The database consists of 100 curves of thoracic volumes (Fig.1 shows an example of the thoracic volume), abdominal volumes and (PaO_2) (partial pressure of oxygen in arterial, a marks of the efficiency of the respiratory system), obtained from 100 normal individuals. For each individual i, we observe a curve of thoracic volume and a curve of abdominal volume. The problem which intervenes here is that we have some individuals whose (PaO_2) is not available (it is not easy to measure the (PaO_2) for all individuals).

In order to compare the three estimators, we start first with the estimator $\widehat{m}_V$ from only the validation set data obtained by deleting all the individuals who have unknown (PaO_2) (40 individuals). Then, we consider the estimator $\widehat{m}_{imp}$ from imputation data and finally the estimator $\widehat{m}_S$ from the added surrogate data.

We evaluate the performance of the estimator, in terms of prediction, by computing the relative mean squared errors (RMSE) on the test sample.

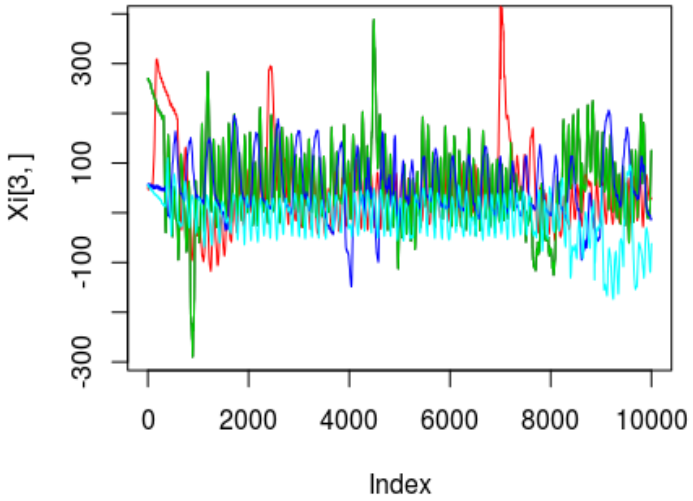


Figure 1: Curves of thoracic volume of the 30 individuals

How to choose surrogate data ? There exists a relation between (PaO_2) and the maximum of the respiratory signal (peaks of the abdominal volume). So, we can compute the mean of these peaks for the variable abdominal volume and use it as surrogate data for the individuals for whom the (PaO_2) is not observed. Using the predictor values of the (PaO_2), we can also predict the biological age of each individual (cf. Cardus et al., 1997).

Results given in Table 1 show that our estimator $\widehat{m}_S$ outperforms $\widehat{m}_{imp}$ and $\widehat{m}_V$.

Estimator	n	RMSE
$\widehat{m}_V$	10	0.935
$\widehat{m}_{imp}$	10	0.727
$\widehat{m}_S$	10	0.690

Table 1: `RMSE` of each estimator.

This application confirms the superiority and the usefulness, in practical situations, of the estimator with surrogate data over the estimator with imputation or with only validation data.

References

[1] Al-Awadhi, F., Benchiha, A., Kaid, Z., Laksaci, A. and Rachdi, M. (2019). FDA: *k*NN local linear regression with missing data at random. *Technical report*.

[2] Al-Awadhi, F., Kaid, Z., Laksaci, A., Ouassou, I. and Rachdi, M. (2019). Local linear functional quantile regression: L1 approach. *Stat. Methods Appl., (2018)*. https://doi.org/10.1007/s10260-018-00447-5

[3] Almanjahie, A.M., Elmezouar, Z.C., Laksaci, A. and Rachdi, M. (2018). *k*NN local linear estimation of the conditional cumulative function: Dependent functional data case (in French). *C. R. Acad. Sci. Paris, Mathématiques*, Sér. I, **356**, Issue 10, 1036–1039.

[4] Almanjahie, A.M., Elmezouar, Z.C., Laksaci, A. and Rachdi, M. (2019). FDA: strong consistency of the *k*NN local linear estimation of the functional conditional density and mode. *J. Nonparametr. Statist.*, **31**, Issue 1, 175–195.

[5] Al-Awadhi, F., Benchiha, A., Kaid, Z., Laksaci, A. and Rachdi, M. (2019). FDA: *k*NN local linear regression with missing data at random. *Technical report*.

[6] Al-Awadhi, F., Kaid, Z., Laksaci, A., Ouassou, I. and Rachdi, M. (2019). Local linear functional quantile regression: L1 approach. *Stat. Methods Appl., (2018)*. https://doi.org/10.1007/s10260-018-00447-5

[7] Almanjahie, A.M., Elmezouar, Z.C., Laksaci, A. and Rachdi, M. (2018). *k*NN local linear estimation of the conditional cumulative function: Dependent functional data case (in French). *C. R. Acad. Sci. Paris, Mathématiques*, Sér. I, **356**, Issue 10, 1036–1039.

[8] Benhenni, K., Ferraty, F., Rachdi, M. and Vieu, P. (2007). Local smoothing regression with functional data. *Comput. Statist.*, **22**, Issue 3, 353–369.

[9] Benhenni, K., Hedli-Griche, S. and Rachdi, M. (2017). Regression models with correlated errors based on functional random design. *TEST*, **26**, Issue 1, 1–21.

[10] Benhenni, K., Su, Y. and Rachdi, M. (2013). The effect of the regularity of the error process on the performance of kernel regression estimators. *Metrika*, **76**, Issue 6, 765–781.

[11] Bongiorno, E., Salinelli, E., Goia, A. and Vieu, P. (Eds.) (2014). *An overview of IWFOS'2014. In Contributions in infinite-dimensional statistics and related topics*, 1-6, Esculapio, Bologna.

[12] Chen, Y.H. (2002). Cox regression in cohort studies with validation sampling. *J. Roy. Statist. Soc. Ser. B (Statistical Methodology)*, **64**, Issue 1, 51–62.

[13] Ferraty, F., Sued, M. and Vieu, P. (2013). Mean estimation with data missing at random for functional covariables. *Statistics*, **47**, 688–706.

[14] El Methni, M. and Rachdi, M. (2011). Local weighted average estimation of the regression operator for functional data. *Commun. Stat. Theory Methods*, **40**, Issue 17, 3141–3153.

[15] Ferraty, F. and Vieu, P. (2006). *Nonparametric functional data analysis. Theory and Practice*. Springer Series in Statistics. New York.

[16] Wang, Q.H. (2006). Nonparametric regression function estimation with surrogate data and validation sampling. *J. Multivariate Anal.*, **97**, 1142–1161.

[17] Wang, J. L., Chiou, J.M. and Muller, H.G. (2016). Review of Functional Data Analysis. *Annu. Rev. Stat. Appl.*, **3**, 257–295.

Apprentissage semi-supervisé avec réponses manquantes informatives

Aude Sportisse ⋆ & Pierre-Alexandre Mattei ⋆ & Charles Bouveyron ⋆

⋆ *Université Côte d'Azur, Inria, Équipe Maasai, Laboratoire J.A. Dieudonné, CNRS*
{*aude.sportisse,pierre-alexandre.mattei,charles.bouveyron*}*@inria.fr*

Résumé. Pour prédire un phénomène, des données parfaites seraient constituées de plusieurs variables explicatives et d'une variable cible ; bien sûr, toutes seraient observées. En pratique, bien que la quantité de données disponibles soit souvent énorme, l'étiquetage des données est coûteux et prend du temps. L'apprentissage semi-supervisé vise à exploiter des données étiquetées et non étiquetées pour entraîner des modèles prédictifs. Dans ce travail, le non-étiquetage des données est considéré comme un problème de données manquantes. Nous nous concentrons sur le cas MNAR (Missing Not At Random), lorsque le manque des étiquettes dépend de leurs valeurs elles-mêmes. Par exemple, cela se produit lorsque les gens sont plus enclins à étiqueter les images de certaines classes qui sont faciles à reconnaître. Dans ce cadre, nous proposons d'estimer le mécanisme des données manquantes et nous proposons des estimateurs pondérés utilisant cette estimation qui tiennent compte de toutes les données.

Mots-clés. apprentissage semi-supervisé, données manquantes informatives

Abstract. To predict a phenomenon, perfect data would consist of many explanatory variables and one target variable; of course, all would be observed. In real life, although the amount of data available is often huge, labeling the data is costly and time-consuming. Semi-supervised learning (SSL) aims at leveraging both labeled and unlabeled data to learn predictive models. In this work, the unlabeled data are seen as missing values. We focus on the MNAR (Missing Not At Random) case, when the unavailability of the labels depends on their values themselves. For example, it occurs when people are more inclined to label images of some classes which are easy to recognize. We propose to estimate the missing-data mechanism and we propose weighted estimators using this estimation which account for all the data.

Keywords. semi-supervised learning, MNAR data

1 Contexte

Bien souvent, les données étiquetées sont beaucoup plus difficile à obtenir que les données non étiquetées. Par exemple, les images issues des radiographies sont nombreuses, mais leur étiquetage peut dans certains cas être fastidieux pour les médecins. Le but de

l'apprentissage semi-supervisé [Chapelle et al., 2009, Van Engelen and Hoos, 2020] est d'utiliser toutes les données possibles (étiquetées et non étiquetées) et d'améliorer les résultats obtenus seulement avec les données étiquetées (cas complet). Une donnée non étiquetée peut être vue comme ayant une variable cible manquante [Grandvalet and Bengio, 2004, Ahfock and McLachlan, 2019, Hu et al., 2022, Schmutz and Mattei, 2022].

Cadre mathématique Nous considérons n échantillons i.i.d. $D = \{(x_i, y_i)\}_{i=1}^n$, avec $x_i \in \mathbb{R}^d$ les variables explicatives et $y_i \in \mathcal{C} = \{0, \dots, K\}$ la variable cible, qui suivent la distribution jointe $p(x, y) = p(x)p(y|x)$. Dans le contexte semi-supervisé, certaines valeurs de la variable cible sont manquantes. On peut donc décomposer le jeu de données en deux parties: un jeu de données étiquetées $D_\ell = \{(x_i, y_i\}_{i=1}^{n_\ell}$ et un jeu de données non étiquetées $D_u = \{(x_i)\}_{i=n_\ell+1}^n$. On définit également l'indicateur de données manquantes, $r \in \{0, 1\}^n$, tel que $r_i = 1$ si y_i est observée et $r_i = 0$ si y_i est manquante. D'après la taxonomie de Rubin [Rubin, 1976], la variable cible est dite (i) MCAR (Missing Completely At Random) si la cause du manque est indépendante des données, i.e. $r \perp\!\!\!\perp x, y$, (ii) MAR (Missing At Random) si la cause du manque dépend des variables explicatives (observées), i.e. $r \perp\!\!\!\perp y|x$ et (iii) MNAR (Missing Not At Random) dans les autres cas. Par exemple, la variable cible sera MAR si les gens sont moins enclins à étiqueter les images qui ont un zoom important, mais sera MNAR si les gens sont plus enclins à étiqueter les images de certaines classes qui sont faciles à reconnaître. Cette dernière situation correspond au cas où les classes ont des *popularités* différentes, qui est assez fréquent en pratique [Oliver et al., 2018]. Dans la suite, nous considérons le cas MNAR auto-masqué, lorsque le manque dépend des valeurs des étiquettes elle-mêmes, i.e. $r \perp\!\!\!\perp x|y$.

Cas MCAR avec seulement les données étiquetées Nous souhaitons estimer le paramètre θ du modèle de prédiction $p(y|x; \theta)$ en minimisant le risque théorique:

$$\theta^\star = \text{argmin}_{\theta \in \Theta} \quad \mathcal{R}(\theta) := \mathbb{E}_{(x,y)\sim p(x,y)}[L(\theta; x, y)],$$

où L est une fonction de perte. Le risque théorique n'est pas une quantité observée, cela supposerait de connaître la vraie distribution des données $p(x, y)$. Lorsque toutes les données sont observées, un estimateur non biaisé du risque théorique est le risque empirique:

$$\hat{\theta} = \text{argmin}_{\theta \in \Theta} \quad \hat{\mathcal{R}}(\theta) := \frac{1}{n} \sum_{i=1}^{n} L(\theta; x_i, y_i).$$

Cette quantité n'est pas observée dans le cadre de l'apprentissage semi-supervisé, car certaines valeurs de la variable cible sont manquantes. Lorsque les valeurs manquantes sont de type MCAR, un estimateur non biaisé du risque théorique est un risque empirique qui n'utilise que les données étiquetées:

$$\hat{\mathcal{R}}^{\mathrm{CC}}(\theta) := \frac{1}{n}\sum_{i=1}^{n} r_i L(\theta; x_i, y_i).$$

Cas MCAR avec toutes les données La plupart des algorithmes qui proposent d'utiliser toutes les données ajoute un terme de régularisation calculé à partir des données non étiquetées également:

$$\hat{\mathcal{R}}^{\mathrm{SSL}}(\theta) := \frac{1}{n}\sum_{i=1}^{n} r_i L(\theta; x_i, y_i) + \frac{\lambda}{n}\sum_{i=1}^{n} H(\theta; x_i),$$

où H est un terme qui peut généralement être vu comme un substitut de L. Par exemple, Grandvalet and Bengio [2004] proposent une méthode appelée minimisation d'entropie, où H est l'entropie de Shannon. Il existe beaucoup d'autres méthodes, comme la pseudo-étiquette, qui consiste à remplacer les valeurs manquantes de la variable cible par la classe k qui maximise la probabilité prédite, i.e. $k \in \mathrm{argmax}_{\ell\in\{0,\dots,K\}} p(y = \ell|x; \theta)$ [Scudder, 1965].

2 Contribution

Dans le cas MNAR, les deux estimateurs du risque théorique vus précédemment, celui dans le cas complet $\hat{\mathcal{R}}^{\mathrm{CC}}$ et celui utilisant toutes les données $\hat{\mathcal{R}}^{\mathrm{SSL}}$, sont biaisés.

Pour rendre le risque empirique dans le cas complet $\hat{\mathcal{R}}^{\mathrm{CC}}$ sans biais, une stratégie consiste à le pondérer par l'inverse de la probabilité d'être observé.

$$\tilde{\mathcal{R}}^{\mathrm{IPW,MNAR}}(\theta) := \frac{1}{n}\sum_{i=1}^{n} \frac{r_i L(\theta; x_i, y_i)}{\pi(y_i)},$$

avec $\pi(y) = \mathbb{P}(r = 1|y)$. Le mécanisme de données manquantes π doit ainsi être estimé.

Nos contributions sont les suivantes:

1. Nous proposons d'estimer le mécanisme π par $\hat{\pi}$ en maximisant la vraisemblance observée. Nous considérons ainsi l'estimateur non biaisé suivant:

$$\hat{\mathcal{R}}^{\mathrm{IPW,MNAR}}(\theta) := \frac{1}{n}\sum_{i=1}^{n} \frac{r_i L(\theta; x_i, y_i)}{\hat{\pi}(y_i)}.$$

2. Nous proposons un estimateur non biaisé du risque théorique qui utilise également les données non étiquetées.

$$\hat{\mathcal{R}}^{\mathrm{SSL,MNAR}}(\theta) := \frac{1}{n}\sum_{i=1}^{n} r_i L(\theta; x_i, y_i) + \frac{\lambda}{n}\sum_{i=1}^{n} \frac{r_i - \hat{\pi}(y_i)}{\hat{\pi}(y_i)} H(\theta; x_i).$$

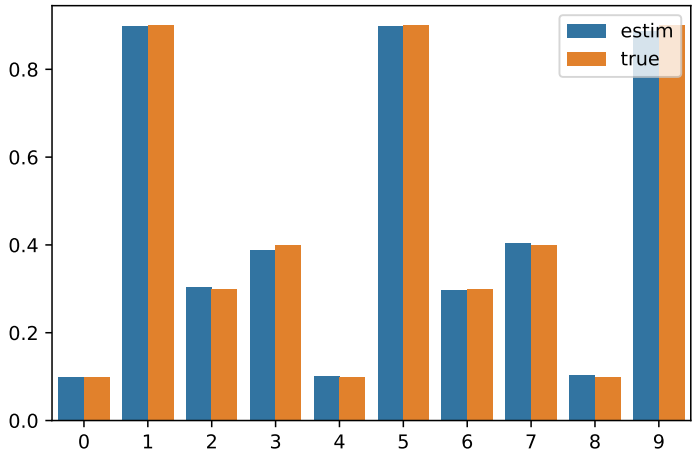


Figure 1: Estimation du mécanisme $\pi(y) = \mathbb{P}(r = 1|y), y \in \{0, \ldots, 9\}$ pour le jeu de données MNIST lorsque 55% des données sont non-étiquetées.

3. Ces deux approches sont étudiées numériquement sur des jeux de données classiques (MNIST, CIFAR) et comparées aux méthodes classiques.

Une première expérience numérique (voir Figure 1) sur le jeu de données MNIST permet de montrer que le mécanisme peut être estimé correctement en utilisant les estimateurs classiques issus de la maximisation de la vraisemblance observée. Les gradients sont calculés à l'aide du logiciel de différenciation automatique de Pytorch [Paszke et al., 2017].

References

Daniel Ahfock and Geoffrey J McLachlan. On missing label patterns in semi-supervised learning. *arXiv preprint arXiv:1904.02883*, 2019.

Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien. Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *IEEE Transactions on Neural Networks*, 20(3):542–542, 2009.

Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. *Advances in neural information processing systems*, 17, 2004.

Xinting Hu, Yulei Niu, Chunyan Miao, Xian-Sheng Hua, and Hanwang Zhang. On non-random missing labels in semi-supervised learning. In *International Conference on Learning Representations*, 2022.

Avital Oliver, Augustus Odena, Colin A Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. *Advances in neural information processing systems*, 31, 2018.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

Hugo Schmutz and Pierre-Alexandre Mattei. Don't fear the unlabelled: safe deep semi-supervised learning via simple debiasing. 2022.

Henry Scudder. Probability of error of some adaptive pattern-recognition machines. *IEEE Transactions on Information Theory*, 11(3):363–371, 1965.

Jesper E Van Engelen and Holger H Hoos. A survey on semi-supervised learning. *Machine Learning*, 109(2):373–440, 2020.

Estimation ciblée pour les modèles structurels marginaux

Herbert Susmann [1] & Antoine Chambaz [2]

[1] *University of Massachusetts Amherst and MAP5 (UMR CNRS 8145), Université de Paris, hsusmann@umass.edu*

[2] *MAP5 (UMR CNRS 8145), Université de Paris, antoine.chambaz@u-paris.fr*

Résumé. Deux des tâches principales de l'inférence causale sont de définir et d'estimer l'effet d'un traitement sur une issue d'intérêt. Formalement, ces effets sont definis comme un sommaire fonctionnel de la loi générant les données, et sont dénommés *paramètres d'intérêt*. L'estimation du paramètre d'intérêt peut être difficile, surtout lorsqu'il est de grande dimension. Les Modèles Structurels Marginaux (MSM) permettent de résumer tels paramètres en utilisant un modèle de travail spécifié par l'utilisateur. Nous déterminons la borne d'efficacité semi-parametrique pour l'estimation des paramètres des MSM dans un cadre général. Ensuite, nous présentons un estimateur fondé sur le principe de la minimisation ciblée de pertes qui atteint cette borne. Nos resultats peuvent être facilement adaptés à des structures de donné et paramètres d'intérêt spécifiques.

Mots-clés. analyse causale, apprentissage ciblé, statistique semi-paramétrique

Abstract.

Two of the principle tasks of causal inference are to define and estimate the effect of a treatment on an outcome of interest. Formally, such treatment effects are defined as a functional summary of the data generating distribution, and are referred to as *target parameters*. Estimation of the target parameter can be difficult, especially when it is high-dimensional. Marginal Structural Models (MSMs) provide a way to summarize such target parameters in terms of a lower dimensional user-supplied working model. We derive the semi-parametric efficiency bound for estimating MSM parameters in a general setting. We then present an estimator that achieve this bound based on Targeted Minimum Loss-Based Estimation. Our results are derived in a general context, and can be easily adapted to specific data structures and target parameters.

Keywords. causal inference, semi-parametric statistics, targeted learning

1 Introduction

We begin by describing a general framework for Marginal Structural Models (MSMs). Suppose we have n i.i.d. draws $O_1, \ldots, O_n$ from a distribution P_0 of a generic variable O. Let $\mathcal{O}$ be the support of P_0. We assume only that P_0 falls in a non-parametric model $\mathcal{M}$ of laws on $\mathcal{O}$. Further, suppose O can be decomposed into two variables, $O = (X, Z)$ (this is always possible, as we can always set Z to the trivial set). Let $\mathcal{X}$ and $\mathcal{Z}$ denote

the spaces to which X and Z belong. For any $P \in \mathcal{M}$, let $\Psi_P : \mathcal{X} \to \mathbb{R}$ be a functional summary of P, with Ψ_P a member of the function class $\mathcal{F}$. For later convenience, we introduce $\mathcal{T} := \cup_{P \in \mathcal{M}} \Psi_P(\mathcal{X})$ (that is, the union of the images of all Ψ_Ps). We focus on instances where $\mathcal{T}$ is an open set. We also assume that Ψ_P depends on P through a finite tuple of parameters $\bar{Q}_P = (\bar{Q}_P^{(0)}, \ldots, \bar{Q}_P^{(J)})$.

Example: Conditional Average Treatment Effect As an example, we will use the Conditional Average Treatment Effect (CATE) parameter. Let $Z = (A, Y)$, where A is a binary treatment indicator and Y is an outcome. Let $g_P(a, x) := P(A = a | X = X)$, and suppose that $g_P(1, X) > 0$ holds P-almost surely. Define $\Psi_P^\flat(x) = \bar{Q}_P^{(1)}(x) - \bar{Q}_P^{(0)}(x)$. This is interpretable as the expected difference in the outcome for treatment vs. non-treatment within the strata of covariates $X = x$.

For convenience, we will write $\psi_0 := \Psi_{P_0}$ to denote the functional under the true data generating distribution P_0. Estimating the functional ψ_0 across its entire domain may be challenging, especially if $\mathcal{X}$ is high dimensional. The idea with MSMs is to seek an approximation of the functional ψ_0 defined by a lower-dimensional working model. In other words, instead of estimating ψ_0, we estimate a parameter for the working model that leads to the best possible approximation of ψ_0 within the working model, as measured by the risk induced by a well-chosen, problem-specific loss function.

Formally, a loss function is a function $L : \mathcal{T} \times \mathbb{R} \to \mathbb{R}$ such that, in particular, for all $f \in \mathcal{F}$,

$$\mathbb{E}_P \left[L(\Psi_P(X), \Psi_P(X)) \right] \leq \mathbb{E}_P \left[L(\Psi_P(X), f(X)) \right].$$

The working model is a collection $\{m_\beta : \beta \in \mathcal{B}\}$ of functions $m_\beta : \mathcal{X} \to \mathbb{R}$ with $\mathcal{B}$ a parameter of dimension p. For convenience, we write $L_m(t, \beta)(X) := L(t, m_\beta(X))$ for all $t \in \mathcal{T}$ and $\beta \in \mathcal{B}$. The target parameter $B(P) \in \mathcal{B}$ is defined as any solution to the following optimization problem:

$$B(P) \in \arg\min_{\beta \in \mathcal{B}} \mathbb{E}_P \left[L_m(\Psi_P(X), \beta)(X) \right].$$

Note that $B(P)$ depends on P only through the functional summary Ψ_P and the marginal distribution of X, which we denote Q. From now on, we assume that there is a unique minimizer of (1) in the case where $P = P_0$. We call this minimizer $\beta_0 := B(P_0)$.

Prior Work Marginal Structural Models were first introduced by Robins (1998). Our approach for estimating the parameters of an MSM is based on Targeted Maximum Likelihood Estimation (TMLE, van der Laan and Rose, 2011; 2018). Non-parametric estimators have been developed for a variety of specific data structures, target parameters, and MSM loss functions and working models. These include estimators for vector-valued treatment

assignments, time-varying treatments, and instrumental variable designs (Petersen et al., 2014; Rosenblum & van Der Laan, 2010; Toth & van der Laan, 2016; Zheng, Luo, & van der Laan, 2018; Zheng, Petersen, & van der Laan, 2016). In practice, users may wish to test several alternative working model specifications and loss functions. This is difficult to do using existing approaches, as the estimators (and related software) are constructed for specific classes of working models and loss functions. Our results build off these works by proposing a TMLE based estimation strategy (and corresponding software implementation) in a generalized data structure and for arbitrary MSM loss functions and working models.

2 Semi-Parametric Efficiency Theory

Let $\beta_0 = B(P_0)$, the value of the parameter of interest under the true data generating distribution P_0. In this section, we derive the semi-parametric efficiency bound for estimating β_0 in a non-parametric model $\mathcal{M}$ (that is, $\mathcal{M}$ is the set of all laws P on $\mathcal{X} \times \mathcal{Z}$ such that $B(P)$ is well-defined.) Semi-parametric efficiency theory teaches us that the efficiency bound is given by the variance of the *efficient influence function*, denoted $D^*(P)$. The efficient influence function D^* exists as long as the parameter $B(P)$ is *pathwise differentiable*. Showing that $B(P)$ is pathwise differentiable, and subsequently deriving the form of $D^*(P)$, is a crucial step in understanding the behavior of non-parametric estimators of $B(P)$. In addition, in the next section we will use the form of D^* to construct an estimator that achieves the efficiency bound. In the next theorem, we present detailed conditions under which $B(P)$ is pathwise differentiable, and characterize the form of $D^*(P)$.

Theorem 1 *Efficient influence function for Marginal Structural Models. Key assumptions (several technical assumptions elided for brevity):*

1. *For every $P \in \mathcal{M}$, for every $t \in \mathcal{T}$, the following two conditions are met P-almost surely:*

   (a) *$\beta' \mapsto L_m(t, \beta')(X)$ is differentiable at every $\beta \in \mathcal{B}$ with derivative $\dot{L}_m(t, \beta)(X) \in \mathbb{R}^p$,*

   (b) *$\beta' \mapsto \dot{L}_m(t, \beta')(X)$ is differentiable at every $\beta \in \mathcal{B}$ with derivative $\ddot{L}_m(t, \beta)(X) \in \mathbb{R}^{p \times p}$.*

   Moreover, for every $P \in \mathcal{M}$, for every $\beta \in \mathcal{B}$, it holds P-almost surely that

   (c) *$t' \mapsto \dot{L}_m(t', \beta)(X)$ is differentiable at every $t \in \mathcal{T}$ with derivative $\nabla \dot{L}_m(t, \beta)(X) \in \mathbb{R}^p$.*

2. *For all $P \in \mathcal{M}$, for all $s \in L_0^2(P)$ such that $s \neq 0$, $\|s\|_\infty < \infty$, if $\{P_\epsilon : |\epsilon| < \|s\|_\infty^{-1}\}$ is characterized by $\frac{dP_\epsilon}{dP} = 1 + \epsilon s$ for all $\epsilon \in \mathbb{R}$ such that $|\epsilon| < \|s\|_\infty^{-1}$ and $P_\epsilon \in \mathcal{M}$, the*

mapping $\epsilon \mapsto \Psi_{P_\epsilon}(X)$ *is almost surely differentiable at 0 and*

$$\frac{d}{d\epsilon}\Psi_{P_\epsilon}(X)\bigg|_{\epsilon=0} = \mathbb{E}_P\left[\Delta^*(P)(O)s(O) \mid X\right]$$

for a function $\Delta^*(P) \in L_0^2(P)$.

The target functional $P \mapsto B(P)$ *is pathwise differentiable at every* $P \in \mathcal{M}$*, with an efficient influence function* $D^*(P)$ *given by*

$$D^*(P)(O) = M^{-1}\left[D_1^*(P)(O) + D_2^*(P)(X)\right],$$

where $D_1^*(P), D_2^*(P) \in L_0^2(P)$ *are given by*

$$\begin{aligned} D_1^*(P)(O) &= \nabla \dot{L}(\Psi_P(X), B(P))(X) \times \Delta^*(P)(O), \\ D_2^*(P)(X) &= \dot{L}(\Psi_P(X), B(P))(X), \end{aligned}$$

and the normalizing matrix M *is given by*

$$M = -\mathbb{E}_P\left[\ddot{L}_m(\Psi_P(X), B(P))(X)\right].$$

This result can be used to find the form of the EIF for specific data structures and functional summaries Ψ_P, as illustrated below.

Example: Conditional Average Treatment Effect (cont'd) Consider a linear working model $m_\beta(X) = \beta^\top X$ and a squared-error loss function $L(x, y) = (x - y)^2$. Let $M = -\mathbb{E}[XX^\top]$. Then the EIF for a marginal structural model defined for the CATE parameter $\Psi_P^\flat$ is given by

$$D_\flat^*(P)(O) = M^{-1}\left[\left\{\frac{I(A=1)}{g_P(1,X)} - \frac{I(A=0)}{g_P(0,X)}\right\}\left(Y - \bar{Q}_P^{(A)}\right) + \Psi_P^\flat(X) - B(P)^\top X\right].$$

3 Targeted Maximum Likelihood Estimation

In this section we give an outline of a non-parametric estimation strategy that achieves the efficiency bound presented in the previous section that is based on Targeted Minimum Loss-Based Estimation (TMLE, van der Laan and Rose 2011, 2018). The core idea of TMLE is to iteratively update an initial estimate of the parts of P_0 relevant to $B(P)$ in order to solve the efficient influence function D^*. We present the outline below as a pseudo-algorithm:

1. Obtain an initial estimate P_n^0 of the relevant parts of P_0 with respect to $B(P_0)$.

2. For any $P \in \mathcal{M}$, set a parametric sub-model $\{P_\epsilon : \epsilon \in \mathbb{R}^p\}$ such that $P_0 = P$ and the score of P_ϵ at $\epsilon = 0$ spans the EIF D^*.

3. Repeat the following iterative procedure:

   (a) For $k \geq 1$, recursively let $P^k_{n,\epsilon} = P^{k-1}_{n,\epsilon}$ where $\epsilon_{n,k}$ is estimated using maximum likelihood:

   $$\epsilon^k_n = \arg\max_{\epsilon \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \left[P^k_{n,\epsilon}(O_i) \right],$$

   and stop when $\epsilon^k_n \approx 0$ (rigorously, when $\epsilon^k_n = o_P(n^{-1/2})$).

4. Set $P^*_n = P^k_n$, where k is the final iteration of the above step.

We then estimate the value of the target parameter by plugging in the updated estimate P^*_n, setting $\beta^*_n := B(P^*_n)$. Importantly, it can be shown that $\frac{1}{n}\sum_{i=1}^{n} D^*(P^*_n)(O_i) = 0$. The following theorem, whose proof relies on this property of P^*_n, gives conditions under which the resulting estimator is asymptotically normal and efficient.

Theorem 2 *Asymptotic normality and efficiency of TMLE. Assumptions:*

- *Rate of convergence of second-order remainder:* $\beta^*_n - \beta_0 + P_0 D^*(P^*_n) = o_P(n^{-1/2})$.
- *Donsker conditions:* $D^*(P^*_n)$ *is in a* P_0*-Donsker class with probability tending to one, and* $P_0[D^*(P^*_n) - D^*(P_0)]^2 \xrightarrow{P_0} 0$*, where the operations and convergence in probability are to be understood element-wise.*

Then $\hat{\beta}^*_n$ *is asymptotically normal and efficient:*

$$\sqrt{n}\left(\beta^*_n - \beta_0\right) \rightsquigarrow N\left(0, P_0[D^*(P_0)D^*(P_0)^\top]\right).$$

The asymptotic normality of the estimator implies that asymptotically valid confidence intervals can be formed using the empirical variance of the estimated efficient influence function: $\left[\beta^*_n \pm 1.96\frac{\sigma^*_n}{\sqrt{n}}\right]$, where $\sigma^*_n = P_n[D^*(P^*_n)D^*(P^*_n)^\top]$.

4 Implementation with Automatic Differentiation

Practical implementation of the TMLE estimator requires calculating an estimate of the efficient influence function, both for forming the fluctuation submodel at the heart of the TMLE procedure and for forming confidence intervals for the resulting estimates. The form of the efficient influence function D^* depends on the specific choice of the MSM loss function and working model. Specifically, D^* depends on three derivatives of the loss function and working model (Theorem 1). Automatic differentiation can be used

to calculate these derivatives in software, allowing the user to specify any loss function or working model (as long as they satisfy the required conditions, such as differentiability.) We will present an implementation of the proposed TMLE estimator written in the Julia programming language, using the `ForwardDiff.jl` package for forward-mode autodifferentiation Bezanson, Edelman, Karpinski, and Shah (2017); Revels, Lubin, and Papamarkou (2016).

5 Simulation Study

Next, we present a simulation study that compares the finite sample performance of the frequentist and Bayesian targeted estimators for estimating Conditional Treatment Effects. The data generating distribution is defined by:

$$\begin{aligned}(X_1, X_2, X_3, X_4) &\sim \mathrm{N}_4(0, I_4),\\ A|X_1, X_2, X_3, X_4 &\sim \text{Bernoulli}(\text{logit}^{-1}(0.5X_1 - 0.5X_2 + 0.2X_3 - 0.1X4)),\\ Y|A, X_1, X_2, X_3, X_4 &\sim \mathrm{N}(X_2 + X_3 + 3A + 1.5AX_4, \sigma^2),\end{aligned}$$

where the outcome variance is set to $\sigma^2 = 0.1$. We define a marginal structural model with a linear working model and squared-error loss function given by

$$B(P) = \underset{\beta \in \mathbb{R}^2}{\arg\min}\, \mathbb{E}_P \left[\left(\Psi_P^\flat(X) - (1, X_4)\beta \right)^2 \right],$$

where $\Psi_P^\flat$ is defined as in the example and $\beta = (\beta_1, \beta_2)^\top \in \mathbb{R}^2$. Next, 500 datasets were drawn from the data generating distribution for each sample size $n \in \{100, 200, 300, \ldots, 1000\}$. The nuisance parameters $\bar{Q}_P^{(0)}$ and $\bar{Q}_P^{(1)}$ were estimated using correctly specified linear models. The TMLE estimator was applied to estimate β_n^* in each dataset and 95% confidence intervals were formed based on the estimated variance of the efficient influence function. The results of the simulation study indicate that the proposed TMLE estimator performs well in this finite sample setting. The empirical coverage of the 95% confidence intervals show that they are well calibrated (Figure 1).

Bibliography

Bezanson, J., Edelman, A., Karpinski, S., & Shah, V. B. (2017). Julia: A fresh approach to numerical computing. *SIAM Review*, *59*(1), 65–98. Retrieved from `https://epubs.siam.org/doi/10.1137/141000671` doi: 10.1137/141000671

Petersen, M., Schwab, J., Gruber, S., Blaser, N., Schomaker, M., & van der Laan, M. (2014). Targeted maximum likelihood estimation for dynamic and static longitudinal marginal structural working models. *Journal of Causal Inference*, *2*(2), 147–185. Retrieved from `https://doi.org/10.1515/jci-2013-0007` doi: doi:10.1515/jci-2013-0007

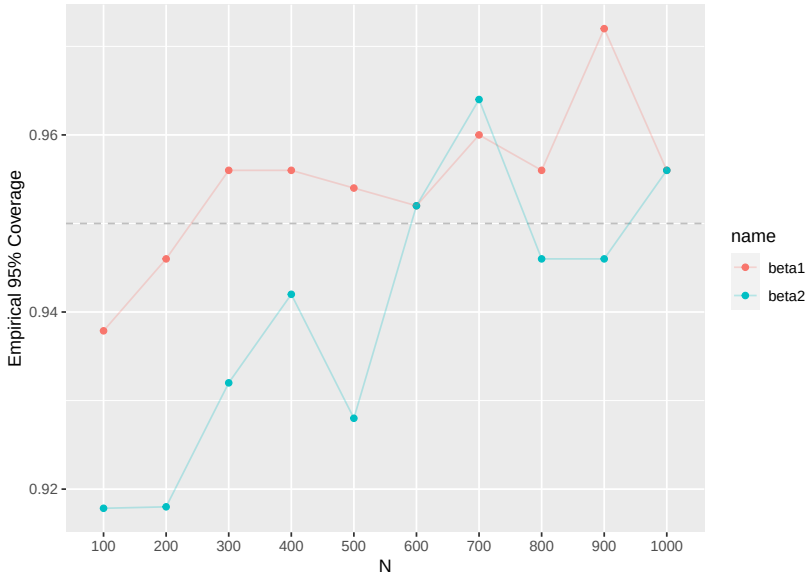


Figure 1: Empirical coverage of 95% confidence intervals for β_1 and β_2 in the simulation study.

Revels, J., Lubin, M., & Papamarkou, T. (2016). Forward-mode automatic differentiation in Julia. *arXiv:1607.07892 [cs.MS]*. Retrieved from `https://arxiv.org/abs/1607.07892`

Robins, J. M. (1998). Marginal structural models. *1997 proceedings of the American Statistical Association, section on Bayesian statistical science*, 1–10.

Rosenblum, M., & van Der Laan, M. J. (2010). Targeted maximum likelihood estimation of the parameter of a marginal structural model. *The international journal of biostatistics*, *6*(2).

Toth, B., & van der Laan, M. J. (2016). Tmle for marginal structural models based on an instrument.

van der Laan, M. J., & Rose, S. (2011). *Targeted learning: causal inference for observational and experimental data.* Springer Science & Business Media.

van der Laan, M. J., & Rose, S. (2018). *Targeted learning in data science: causal inference for complex longitudinal studies.* Springer.

Zheng, W., Luo, Z., & van der Laan, M. J. (2018). Marginal structural models with counterfactual effect modifiers. *The international journal of biostatistics*, *14*(1).

Zheng, W., Petersen, M., & van der Laan, M. J. (2016). Doubly robust and efficient estimation of marginal structural models for the hazard function. *The international journal of biostatistics*, *12*(1), 233–252.

Rencontre Jeunes statisticiens

Rencontres Jeunes Statisticiens et Conférenciers invités... et le quizz !

C'est aussi le grand retour des rencontres en présentiel ! Venez les rencontrer autour d'un apéro et du fameux quizz du groupe Jeunes, le jeudi 16 juin à partir de 18h, dans un lieu encore à préciser.

Pour nous permettre d'organiser au mieux ces rencontres, merci de bien vouloir vous inscrire en suivant ce lien : https://framaforms.org/inscription-aux-rencontres-jeunes-statisticiennes-et-conferencieres-invitees-1654079720

Café de la Statistique

Café Toï Toï le Zinc, 17-19 rue Marcel Dutartre, 69100 Villeurbanne,
ou visioconférence. Sur inscription à statcafe@sfds.asso.fr.

19h00 – 21h30 (entrée recommandée dès 18h30)

La transition énergétique : quels choix en France et en Europe ?

Cédric Philibert

Analyste énergie et climats – Chercheur associé à l'IFRI- Ancien responsable des énergies renouvelables à l'Agence Internationale de l'Energie

Les prévisions issues des données et des modèles climatiques présentées par le Groupe d'experts intergouvernemental sur l'évolution du climat (GIEC) font largement consensus dans la communauté scientifique et, en principe, entre les Etats qui se fixent un objectif global de neutralité carbone à l'horizon 2050. Cependant les débats en France et en Europe, dans le contexte d'une baisse rapide des coûts des renouvelables (éolien, solaire, biomasse) et d'un refus de la dépendance à l'égard du gaz et du pétrole russes, manifestent des oppositions et des incertitudes sur les voies à suivre et les moyens nécessaires en ce qui concerne notamment :

- les parts souhaitables et soutenables des énergies renouvelables, du nucléaire et des énergies carbonées...
- les besoins à satisfaire et les économies possibles dans les usages des énergies.

Au vu des données récentes sur les coûts et la faisabilité des options, telles qu'on les trouve présentées en particulier dans les derniers rapports de l'Agence internationale de l'énergie (AIE) et du Réseau de transport de l'électricité (RTE), notre Café de la Statistique du 16 juin, introduit par Cédric Philibert, ancien responsable de l'AIE et expert en matière d'énergies renouvelables, sera l'occasion d'éclairer ces questions essentielles. Un accent sera mis sur quelques problèmes statistiques du monde de l'énergie, à travers par exemple les notions d'énergies primaire et finale, et les répartitions d'émissions directes et indirectes des secteurs de consommation finale.

Pour participer à cette soirée débat, **inscrivez-vous** en adressant un courriel à statcafe@sfds.asso.fr en précisant votre choix présentiel ou visioconférence.
Pour permettre à tous ceux qui le souhaitent de suivre la conférence à distance, un lien de connexion vous sera adressé pour accéder à la conférence.
Pour permettre à tous ceux qui veulent avoir le temps de commander boissons, tartines ou planches avant le début de l'exposé, nous vous recommandons d'arriver vers 18h30.

Liste des auteurs

Abdi, Hervé, 921
Abdillahi Isman, Mohamed, 288
Acharki, Naoufal, 940
Adjabi, Smail, 732
Adjakossa, Eric, 1081
Agniel, Denis, 349
Ah-Pine, Julien, 288, 507
Ahmed, Mohamed-Salem, 45
Al Marjani, Aymen, 975
Alamichel, Louise, 82
Alaoui Belghiti, Siham, 1051
Alaoui, Enora, 413
Allard, Denis, 39
Allassonnière, Stéphanie, 58, 269
Allouche, Michael, 153
Alquier, Pierre, 312
Alsouki, Louna, 319
Amara-Ouali, Yvenn, 571
Amato, Francesco, 607
Amovin-Assagba, Messan Martial, 121
Andre, Manon, 441
Antero, Juliana, 707
Arbel, Julyan, 76, 82, 1111, 1145
Ardouin, Teddy, 1089
Arlot, Sylvain, 325
Armaut, Sara, 1075
Arnaud, Elise, 1069
Asadi, Arefe, 555
Asberg, Anders, 470
Aubin, Jean-Baptiste, 811
Ayme, Alexis, 1152
Azencott, Chloé-Agathe, 552
Azzaoui, Nourddine, 107

Bénesse, Clément, 941
Bachoc, Francois, 1118
Baldé, Oumar, 956
Ballarini, Paolo, 613
Bar-Hen, Avner, 550
Baragatti, Meili, 94
Barbe, Pierre, 413
Barbieri, Antoine, 482
Barbillon, Pierre, 257
Barbot, Benoit, 613
Barbu, Vlad Stefan, 559
Barriac, Vincent, 513
Barrier, Antoine, 969
Basak, Subhasish, 738
Batardière, Bastien, 526
Beaudoin, Annabelle, 921
Bect, Julien, 738
Beecham, Roger, 22, 453
Belhadji, Ayoub, 532
Bellet, Aurélien, 255
Ben Mrad, Oumaima, 999
Benoist, Clément, 470
Bernard, Renan, 652
Bernier, Jacques, 114
Berthoud, Françoise , 417
Bertoncello, Antoine, 940
Bertrand, Quentin, 202
Bethune, Louis, 646
Biernacki, Christophe, 101
Bleza, Esso-Ridah, 900
Bobbia, benjamin, 159
Bode, Nikolai, 88
Bonaldi, Christophe, 373
Bouaziz, Olivier, 873
Boudou, Alain, 1023
Bouhadjera, Feriel, 94
Boulin, Alexis, 165
Bouloré, Antoine, 956
Bousquet, Nicolas, 1005
Boutet, Jerome, 108
Boutin, Rémi, 263
Bouveyron, Charles, 263, 619, 625, 913, 933, 1169
Bouzalmat, Ibrahim, 1057
Boyer, Claire, 1152
Brault, Vincent, 108
Breeur, Marie, 640

Brunel, Victor-Emmanuel, 774
Bry, Xavier, 190
Bucci, Andrea, 668
Bunce, Catey, 18
Bunz, Yoann, 428
Buritica, Gloria, 843
Bystrova, Daria, 82

Côme, Etienne, 619
Cannamela, Claire, 566
Carrière, Mathieu, 247
Castanier, Bruno, 557
Celisse, Alain, 325
Celse, Benoit, 395
Cesa-Bianchi, Nicolò, 963
Chabridon, Vincent, 944
Chagneux, Mathis, 419
Chagny, Gaëlle, 500
Chainais, Pierre, 772
Chambaz, Antoine, 494, 1011, 1174
Chamroukhi, Faicel, 577, 600
Charlier, Benjamin, 1063
Chautru, Emilie, 849
Chauvet, Guillaume, 830, 891
Chavent, Marie, 132
Chazal, Frédéric , 1029
Chen, Chung Shue, 583
Chennetier, Guillaume, 383
Cheysson, Felix, 215
Chiquet, Julien, 526
Chrétien, Stéphane, 513
Chraïbi, Hassane, 383
Clémençon, Stephan, 544
Claeys, Emmanuelle, 613
Claudel, Sandra, 686
Clausel, Marianne, 1038
Clementz, Samy, 325
Colizza, Vittoria, 239
Colnet, Bénédicte, 950
Combes, Florian, 744
Combrisson, Damien, 428
Corneli, Marco, 625, 913
Cornelius, Victoria, 27–32, 476
Coulmy, Nicolas, 707
Courcoul, Léonie, 482
Courty, Nicolas, 652
Coutin, Laure, 1038
Crambes, Christophe, 1158
Crepey, Pascal, 239
Crivello, Jean-Claude, 1107
Cry, Pierre, 613
Cugliari, Jairo, 686
Cwiling, Ariane, 873

D'Amico, Guglielmo, 559
Döhler, Sebastian, 221
Déjean, Sébastien, 718
Da Veiga, Sébastien, 944
Daayeb, Chayma, 1158
Dabo-Niang, Sophie, 45
Damblin, Guillaume, 956
Daouia, Abdelaati, 798
David, Ingrid, 367
De Bortoli, Valentin, 1101
De Lara, Lucas, 646
De Larochelambert, Quentin, 694, 707
De Saporta, Benoîte, 1057
De Vilmarest, Joseph, 295
Delattre, Maud, 70
Delon, Julie, 1101
Deltreil, Guillaume, 488
Demangeot, Marine, 849
Demongeot, Jacques, 1164
Denis, Christophe, 228
Derquenne, Christian, 538
Descatha, Alexis, 488
Desolneux, Agnès, 1101
Di Bernardino, Elena, 177
Dieuleveut, Aymeric, 594, 1152
Difernand, Audrey, 701, 707
Digne, Julie, 254
Dombry, Clément, 64, 159, 855, 860
Drazek, Laurent, 1089
Du Roy De Chaumaray, Marie, 184
Ducombe, Stephanie, 694
Duprey, Corentin, 107
Durrleman, Stanley, 269
Dussap, Florian, 331
Dutfoy, Anne, 383, 1145
Duval, Laurent, 319
Duveau, Catherine, 140

El Harfaoui, Echarif, 300
Ella Mintsa, Eddy, 506
Enticott, Euan, 679
Erwan, Scornet, 950
Escobar-Bach, Mikael, 488
Esposito, Nicola, 557

Féron, Olivier, 594

Favre, Julien, 101
Feau, Cyril, 389
Fermanian, Jean-Baptiste, 337
Fernandez, Camila, 583
Fischer, Aurélie, 127
Flament, Guillaume, 818
Forbes, Florence, 76, 600
Fort, Gersende , 1030
Fougères, Anne-Laure, 855
Fouladirad, Mitra, 555
Fraiman, Ricardo, 744
Frasca, Denis, 1133
Fraysse, Guillaume, 513
Frellsen, Jes, 751
Frevent, Camille, 45
Friguet, Chloé, 652
Fromont, Magalie, 879

Gaillard, Pierre, 583
Galant, Julie, 107
Gamboa, Fabrice, 1005
Gannaz, Irène, 121, 811
Gannoun, Ali, 1158
Garcin, Camille, 824
Garcin, Matthieu, 171
Gares, Valerie, 652, 830, 879, 891
Garivier, Aurélien, 969, 993
Garnier, Josselin, 383, 389, 566, 940
Gastaud, Clément, 993
Gatulle, Nicolas, 494
Gauchy, Clément, 389
Genin, Michaël, 45
Geoffrey, Ecoto, 1011
Gerville-Réache, Léo, 713
Ghattas, Badih, 686, 744, 1051
Gibaud, Julien, 190
Giorgio, Massimiliano, 557
Giraldi, Loic, 956
Girard, Stéphane, 153, 1145
Giulini, Ilaria, 127
Gkelsinis, Thomas, 559
Gobet, Emmanuel, 153
Goga, Camelia, 64
Golovkine, Steven, 1017
Gonzalez-Sanz, Alberto, 646
Goude, Yannig, 571, 594, 686
Graczyk, Piotr, 488
Grave, Clémence, 373
Gribonval, Rémi, 993
Grill, Léo, 980
Grislain, Nicolas, 993
Grollemund, Paul-Marie, 94, 288
Grusea, Simona, 355
Guillemot, Vincent, 921
Guillin, Arnaud, 107

Hajage, David, 830
Hallin, Marc, 16
Hamrouche, Bachir, 571
Happe, André, 891
Harchaoui, Warith, 933
Harel, Michel, 300
Has, Sothea, 588
Hebiri, Mohamed, 228
Hejblum, Boris, 349
Helali, Salima, 750
Henchiri, Yousri, 1158
Hilgert, Nadine, 94
Hivert, Benjamin, 349
Humbert, Olivier, 927

Iapteff, Loïc, 395
Ibrahim, Firas, 1164
Il Idrissi, Marouane, 1005
Iooss, Bertrand, 1005
Ippoliti, Luigi, 668

Jacob, Jérôme, 565
Jacqmin-Gadda, Hélène, 482
Jacques, Julien, 121, 395, 513, 607
Jannot, Anne-Sophie, 58
Jeamart, Marion, 652
Jollois, François-Xavier, 550
Jolly, Caroline, 108
Joly, Alexis, 434, 824, 1063
Joly, Pierre, 373
Josse, Julie, 494, 594, 950
Jourdan, Astrid, 407
Jouvin, Nicolas, 619

Kamel, Mouna, 458
Katsahian, Sandrine, 879
Kaufmann, Emilie , 766
Kerleguer, Baptiste, 566
Khalfi, Abderaouf, 306
Klopfenstein, Quentin, 202
Klutchnikoff, Nicolas, 1017
Kon Kam King, Guillaume, 70, 82
Kpotufe, Samory , 771
Kratz, Marie, 787
Kuhn, Johann, 373

Kwon, Joon, 526

Lévy-Leduc, Céline, 361
La Rocca, Michele, 674
Labriffe, Marc, 470
Labyt, Etienne, 108
Lacroix, Perrine, 196
Laforgue, Pierre, 963
Lalanne, Clément, 993
Laloe, Thomas, 177
Lambert, Raphaël, 108
Lameiras Franco Da Costa, Victor, 395
Lasalle, Etienne, 275
Lasgorceux, Florian, 428
Latouche, Pierre, 263, 619, 913
Laurent, Béatrice, 1044
Lavenu, Audrey, 879
Laverny, Oskar, 343
Le Masson, Stéphane, 980
Le Minh, Tâm, 281
Le Strat, Yann, 373
Le Toquin, Bryan, 694
Leblanc, Frederique, 401
Lebreton, Noé, 507
Ledanois, Thibaut, 724
Lefort, Tanguy, 1063
Lejay, Antoine, 1038
Leoni, Samuela, 811
Leveau, Valentin, 434
Liang, Dingge, 913
Limnios, Myrto, 544
Lopez Sanchez, Javier, 718
Lorenzo, Hadrien, 1124
Loubes, Jean-Michel, 1005, 1044
Louis, Pierre-Yves, 707, 1133
Lu, Yunjiao, 108
Lucas, Felix, 413

Mélard, Guy, 1032
Mahé, Pierre, 1089
Makhlouf, Slimane, 550
Manou-Abi, Solym, 1057
Mantoux, Clément, 269
Marandon, Ariane, 632
Marbac, Matthieu, 184
Marbac-Lourdelle, Matthieu, 101
Marceau, Etienne, 792
Marchello, Giulia, 625
Mareau, Alexis, 879
Marie, Nicolas, 312
Marion, Pierre, 1095
Marquet, Pierre, 470
Marrel, Amandine, 944, 956
Marteau, Clément, 319
Marteau, Pierre-François, 900
Mas, Victoria, 413
Masmoudi, Afif, 999
Massart, Pascal, 571
Massias, Mathurin, 202
Massonaud, Clement, 239
Mattei, Pierre-Alexandre, 751, 927, 933, 1169
Maume-Deschamps, Véronique , 899
Meah, Iqraa, 221
Mercadier, Cécile, 1129
Merlier, Lucie, 565
Meyer, Nicolas, 804
Meynaoui, Anouar, 500
Michel, Bertrand, 247
Mignot, Raphael, 1038
Mikosch, Thomas, 843
Modeste, Thibault, 855
Mohdeb, Zaher, 222
Moins, Théo, 1145
Mokkadem, Abdelkader, 222
Monbet, Valerie, 900
Mougeot, Mathilde, 127
Mourer, Alex, 132
Murris, Juliette, 879

Nadjahi, Kimia, 658
Nagy, Stanislav, 780
Naulin, Jean-Philippe, 680
Naveau, Marion, 70
Naveau, Philippe, 860
Ngatchou-Wandji, Joseph, 300
Ngounou Bakam, Yves Ismaël, 227
Nguyen, Hien Duy, 76, 600
Nguyen, Thi Thanh Yen, 1011
Nguyen, TrungTin, 76, 600
Nguyen, Van Tuan, 898
Niglio, Marcella, 1032
Nortershauser, David, 980

Obst, David, 686
Oger, Emmanuel, 891
Ohl, Louis, 933
Olié, Valérie, 373
Olteanu, Madalina, 132
Opitz, Thomas, 428
Oppenheim, Georges, 686, 1038

Ounajim, Amine, 1133

Padoan, Simone A., 798
Pakzad, Cambyse, 177
Papaïx, Julien, 428
Papadakis, Nicolas, 553
Paparisteidi, Nefeli, 447
Paquelet, Stéphane, 891
Parent, Eric, 114
Patilea, Valentin, 1017
Paudice, Andrea, 963
Pellet, Aurélien, 447
Peng, Liang, 794
Perduca, Vittorio, 873
Perna, Cira, 674
Perrin, Tran Vi-vi Élodie, 680
Peyraud, Eliz, 885
Pham, Nhat Thien, 577
Philibert, Cédric, 1182
Philippe, Cathy, 921
Pic, Romain, 860
Pierrot, Amandine, 687
Pietrosemoli, Natalia, 921
Pinson, Pierre, 687
Poggi, Jean-Michel, 571
Pommeret, Denys, 227
Pontil, Massimiliano, 963
Potier, Axel, 101
Precioso, Frederic, 933
Prieur, Clémentine, 551
Prokopenko, Evgeny, 787
Pruilh, Solange, 58
Pujol, Louis, 208

Quinton, Jean-Charles, 108

Rabier, Charles-Elie, 355
Rachdi, Mustapha, 1164
Ragheb, Waleed, 434
Rahal, Mira, 1069
Ramiro, Lugo, 940
Rebafka, Tabea, 140, 632, 1107
Reise, Wojciech, 836
Reiss, Markus , 666
Rejeb, Sara, 140
Remlinger, Carl, 986
Ricard, Anne, 367
Rigoard, Philippe, 1133
Rigollet, Philippe, 640
Roche, Angelina, 500
Rodolpho Tomassi, Diego, 555
Rohmer, Jérémy, 680
Rohmer, Tom, 367
Rolland, Antoine, 464, 811
Roquain, Etienne, 221, 632
Rosier, Amélie, 312
Rouanet, Anaïs, 235
Roustant, Olivier, 680
Roux, Jonathan, 239
Rouzé, Timothé, 879

Sabourin, Anne, 849
Sacko, Ousmane, 1139
Sadoun, Mohamed Djemaa, 306
Sagaut, Pierre, 565
Saint-Pierre, Philippe, 718
Salmon, Joseph, 824, 1063
Salmona, Antoine, 1101
Samedy, Sonico, 447
Samson, Adeline, 1069, 1089
Sansonnet, Laure, 70, 526
Saracco, Jerôme, 1124
Sarazin, Gabriel, 944
Sarr, Djibril , 906
Sauliere, Guillaume, 694
Schipman, Julien, 694
Schmutz, Hugo, 927
Schneider, Ulrike, 759
Scornet, Erwan, 1152
Sedeaud, Adrien, 707
Selmani, Merwann, 1044
Senetaire, Hugo, 751
Servajean, Maximilien, 434, 824
Servien, Rémi, 519
Silva, Alonso, 583
Slaoui, Yousri, 750, 980, 999, 1133
Smits, Nathalie, 94
Sokolovska, Nataliya, 632, 1107
Sportisse, Aude, 1169
Spychala, Cécile, 64
Stepaniants, George, 640
Stoltz, Gilles, 969
Stupfler, Gilles, 798
Sultanov, Arsen, 1107
Susmann, Herbert, 1174

Taillardat, Maxime, 860
Tamo Tchomgui, Jean Steve, 513
Tardivel, Patrick, 488, 759
Ternès, Nils, 361

Thiébaut, Rodolphe, 349
Thomas-Agnan, Christine, 146
Todo, William, 1044
Torres, Isabel, 1069
Toussaint, Jean-François, 694, 707
Trottier, Catherine, 190
Tzourio, Christophe, 482

Usevich, Konstantin, 1038

Vaiter, Samuel, 253
Valentini, Pasquale, 668
Vandewalle, Vincent, 101
Varoquaux, Gaël, 950
Varron, Davit, 159
Vayatis, Nicolas, 544
Vazquez, Emmanuel, 738
Verbanck, Marie, 245
Vialaneix, Nathalie, 519
Viallon, Vivian, 640
Vicari, Donatella, 676
Viguier-Pla, Sylvie, 458, 1023
Vo, Thanh Huan, 891

Wahl, François, 319
Werge, Nicklas, 866
Wintenberger, Olivier, 295, 804, 843
Woillard, Jean-Baptiste, 470
Wolinski, Pierre, 1111

Yao, Anne-Françoise, 107, 288
Younso, Ahmad, 52

Zaffran, Margaux, 594
Zaouche, Mounia, 88
Zaoui, Ahmed, 228
Zhang, Li-Chun , 765
Zhao, Pan, 494
Zhu, Wencan, 361

Sponsors

aMies
LES MATHS
VECTEUR D'INNOVATION

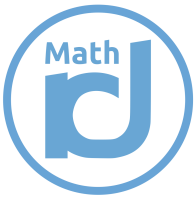
BIOMÉRIEUX

cea

ÉCOLE
CENTRALELYON

ENS DE LYON

eric


Math

Google

Inria

IXXI

ISFA
Institut de Science Financière et d'Assurances
GRANDE ÉCOLE D'ACTUARIAT & GESTION DES RISQUES

LIMOS

Lip

Lyon 1

UNIVERSITÉ
LUMIÈRE
LYON 2

LABEX
MILYON
UNIVERSITÉ DE LYON

MINES
Saint-Étienne

La Région
Auvergne-Rhône-Alpes

UMPA
ENS DE LYON