



MÉMOIRE

En vue de l'obtention d'une

HABILITATION À DIRIGER DES RECHERCHES

Délivrée par l'Université Toulouse I Capitole

École doctorale : **Mathématiques, Informatique et Télécommunications de Toulouse**

Présentée et soutenue publiquement par

GENDRE Xavier

le 14 mars 2022

Discipline : **Mathématiques et Applications**

Spécialité : **Statistique**

Unité de recherche : **ISAE SUPAERO Recherche**

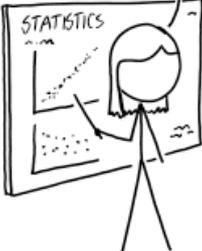
JURY

Philippe BERTHET
Aurélien GARIVIER
Nick HENGARTNER
Hervé MONOD
Mathieu RIBATET
Anne RUIZ-GAZEN
Florian SIMATOS
Christine THOMAS-AGNAN

Université Toulouse III Paul Sabatier
École Normale Supérieure de Lyon
Los Alamos National Laboratory
INRAE
École Centrale de Nantes
Université Toulouse I Capitole
ISAE SUPAERO
Université Toulouse I Capitole

Examineur
Rapporteur
Rapporteur
Rapporteur
Examineur
Examinatrice
Examineur
Marraine

IF YOU DON'T CONTROL FOR CONFOUNDING VARIABLES, THEY'LL MASK THE REAL EFFECT AND MISLEAD YOU.



A stick figure with a bob haircut stands next to a rectangular board. The board is titled "STATISTICS" at the top. It contains a scatter plot with a dashed regression line. The stick figure is pointing at the board with its right hand.

BUT IF YOU CONTROL FOR TOO *MANY* VARIABLES, YOUR CHOICES WILL SHAPE THE DATA, AND YOU'LL MISLEAD YOURSELF.



A stick figure with a bob haircut stands with its right hand raised and pointing upwards, gesturing as if speaking.

SOMEWHERE IN THE MIDDLE IS THE SWEET SPOT WHERE YOU DO BOTH, MAKING YOU DOUBLY WRONG. STATS ARE A FARCE AND TRUTH IS UNKNOWABLE. SEE YOU NEXT WEEK!



A stick figure with a bob haircut stands with its right hand raised and pointing upwards, gesturing as if speaking.

Remerciements

Mes premiers remerciements vont à Aurélien Garivier, Nick Hengartner et Hervé Monod pour avoir accepté de rapporter ce mémoire d'habilitation à diriger des recherches. Je remercie également Philippe Berthet, Mathieu Ribatet, Anne Ruiz-Gazen et Florian Simatos d'avoir accepté de prendre part au jury. J'adresse un merci tout particulier à Christine Thomas-Agnan qui m'a accompagné et encouragé pour organiser cette habilitation à diriger des recherches avec toute sa bienveillance. Je suis très heureux de compter dans mon jury des personnes d'une si grande qualité humaine et scientifique.

Je souhaite aussi remercier ici certains professeurs qui ont grandement marqué ma formation et nourri mon intérêt pour les mathématiques appliquées. Mon éveil doit beaucoup à Gérard Letac, à ses enseignements et à son soutien pour intégrer le magistère de mathématiques et d'informatique de l'École Normale Supérieure de Paris. Ma découverte de la théorie des probabilités avec Jean-François Le Gall a été à n'en pas douter le déclencheur de mon intérêt pour les domaines de l'aléatoire. Les enseignements de Pascal Massart et son enthousiasme pour la statistique ont confirmé mon envie de poursuivre dans cette voie. Enfin, ma formation doit beaucoup à Yannick Baraud qui a encadré mes travaux de thèse à l'origine de tout le reste.

En venant à l'Institut de Mathématiques de Toulouse, j'ai eu la chance de faire des rencontres d'une grande richesse dont certaines vont bien au-delà des mathématiques. Avec mes camarades de doctorat, Marcello Bernardara et Joan Millès, nous nous y sommes retrouvés avec plaisir. J'ai eu la chance de partager mon bureau avec des collègues d'une immense sympathie, merci à Katy Paroux, Pierre Petit et Sophie Jan pour tous les bons moments passés ensemble. Je remercie aussi toutes les personnes qui travaillent ou ont travaillé en soutien à la recherche de l'institut, en particulier Marie-Laure Ausset et Delphine Dallariva pour leur joie de vivre et leur aide, Françoise Michel et Nadège Parrier pour leur gentillesse et leur accompagnement sur les projets scientifiques et Christophe Ségui pour le café et toutes nos discussions. Mes années passées à l'Institut de Mathématiques de Toulouse ont été marquées par des opportunités pédagogiques et des échanges scientifiques pour lesquels

je remercie tout spécialement Nicolas Couellan, Fabrice Gamboa et Béatrice Laurent. Enfin et non des moindres, mes plus agréables sentiments vont à Philippe Berthet, Manon Costa, Sébastien Gadat, Aldéric Joulin, Thierry Klein, Agnès Lagnoux et Clément Pellegrini pour être qui ils sont et pour toutes nos soirées de midi.

Lorsque j'ai rejoint l'Institut Supérieur de l'Aéronautique et de l'Espace, je me souviens avoir été marqué par l'immense bienveillance qui y règne. Ces quelques lignes sont l'occasion d'adresser un grand merci à celles et ceux qui y travaillent avec une attention spéciale pour mes collègues du Département d'Ingénierie des Systèmes Complexes. Merci en particulier à Alain Haït, Rob Vingerhoeds et Odile Riteau pour le bon accueil qui m'a été fait et pour nos échanges réguliers. Je suis également heureux de pouvoir remercier mes collègues de l'équipe de Mathématiques Appliquées ainsi que celles et ceux qui en sont proches à l'image de Meryem Benammar et Olivier Besson.

Le métier d'enseignant-chercheur permet de formidables rencontres dont certaines dépassent largement le cadre scientifique. Depuis nos études jusqu'à maintenant, j'ai la chance de pouvoir compter Nicolas Verzelen parmi mes amis. J'ai rencontré Éric Matzner-Løber dans le cadre de mes activités liées au logiciel R et je suis très heureux de pouvoir le remercier ici pour les échanges et les très bons moments qui découlent de cette rencontre. Par-delà la statistique, j'ai développé une profonde appétence pour des mathématiques appliquées à des problèmes issus de différents domaines scientifiques au fil des dix dernières années. Cette ouverture est le fruit de collaborations enthousiasmantes avec Alexandra ter Halle et Philippe Garnier ainsi qu'avec Noémie Gaudio qui occupe une place centrale dans cette sphère du réel et que j'ai la joie d'avoir à mes côtés depuis la cours du lycée.

Participer à la vie de la communauté et organiser des événements scientifiques fait partie de notre métier. Les plus belles réussites auxquelles j'ai eu le privilège de contribuer en ce domaine sont les Rencontres R organisées à Toulouse en 2016 et surtout la conférence UseR! 2019. Je suis convaincu que la qualité de ces événements doit beaucoup à la fantastique équipe d'organisation que nous avons consituté. Pour tout cela, je remercie chaleureusement Aurore Archimbaud, Christophe Bontemps, Sébastien Déjean, Robert Faivre, Thibault Laurent, Élise Maigné, Pierre Neuvial, Anne Ruiz-Gazen, Rémi Servien, Matthias Zytnicki et tout particulièrement notre dictatrice bienveillante Nathalie Vialaneix.

Par-delà le travail, il y a le soutien des amis sans qui être ce que nous sommes n'aurait pas la même saveur. Ce n'est pas un exercice auquel nous nous prêtons généralement entre nous mais je ne laisse pas passer cette occasion de dire un grand merci à Nico, Anne, Matthieu, Jérémie, Séb, les deux Manu, Damien, Jean-Roc et tous les autres. Je remercie aussi Jelena et Mi-

ckaël qui sont les meilleurs voisins que l'on puisse avoir. Il y a aussi celles et ceux qui sont plus loin et que j'ai toujours autant de joie à retrouver. En particulier, merci à Charly et sa Bougeotte ainsi qu'à Delphine pour nos soirées parisiennes. J'ai également une pensée toute spéciale pour mon cher ami ultra-rhénan Marc.

Nous n'avons pas souvent l'occasion de remercier nos parents pour tout ce qu'ils ont fait pour nous et pour le soutien qu'ils nous témoignent. Je profite donc de ces quelques mots pour adresser mes plus tendres remerciements à mes parents Colette et Guy. Je veux aussi adresser toute mon affection à ma sœur Sophie ainsi qu'aux deux hommes qui partagent maintenant sa vie, Noah et Raphaël. Je n'oublie pas Louis, Paule et Jean-Yves qui savent toute ma sympathie pour eux. Viennent enfin Odile et Félix sans qui mes journées seraient moins denses mais aussi moins belles et Maud pour l'ineffable importance de tout ce qui compte vraiment.

Table des matières

Introduction	1
Curriculum vitæ	3
1 Sélection de modèle pour la régression	9
1.1 Cadre général	9
1.2 Régression Gaussienne hétéroscédastique	12
1.3 Courbes décalées	16
1.4 Régression additive	22
2 Autour de la covariance	29
2.1 Processus Gaussien bivarié	29
2.2 Décomposition de Cholesky régularisée	32
3 Applications aéronautiques	37
3.1 Tolérancement statistique sous contraintes industrielles	37
3.2 Comportement des systèmes électriques aéronautiques	42
4 Applications environnementales	51
4.1 Écologie forestière	51
4.2 Pollution de plastique des océans	56
4.3 Allométrie en cultures mixtes	59
Conclusion et perspectives	63
Liste des travaux	70
Bibliographie	75

Introduction

Présentation générale

Ce manuscrit représente une synthèse de mon activité de recherche depuis l'obtention de mon doctorat à l'Université de Nice Sophia Antipolis en 2009. Après avoir occupé un poste d'attaché temporaire d'enseignement et de recherche à l'Université de Provence, j'ai rejoint l'Université Toulouse III Paul Sabatier en qualité de maître de conférences en 2010. Depuis 2019, je suis détaché à l'Institut Supérieur de l'Aéronautique et de l'Espace en tant qu'enseignant-chercheur en statistique et science des données. Ces postes m'ont donné l'opportunité d'enseigner les mathématiques en général et la statistique en particulier à différents niveaux de l'enseignement supérieur français de la licence au master ainsi qu'en écoles d'ingénieurs et en formation continue. J'ai également eu la chance de dispenser 8 cours de niveau avancé à l'international.

Issu d'une formation orientée en statistique mathématique, mon activité s'est diversifiée au fil des années. Les rencontres scientifiques et les collaborations m'ont conduit à considérer différents domaines d'applications de la statistique. J'ai codirigé deux thèses CIFRE Airbus qui ont été soutenues en 2021, l'une avec Anne Ruiz-Gazen et l'autre avec Nicolas Couellan. Je codirige actuellement deux nouvelles thèses débutées en octobre 2021, l'une avec Sébastien Gadat et l'autre avec Thierry Klein. À ce jour, mes travaux ont donné lieu à 11 publications dans des journaux à comité de lecture et à 5 actes de conférences.

Organisation du document

Pour tous les travaux abordés dans ce document, le contexte scientifique est présenté ainsi que les principaux résultats obtenus. En particulier, les preuves et les simulations sont omises. Pour une présentation complète et des discussions plus avancées, les articles originaux pourront être consultés.

L'ensemble de mes publications est librement accessible depuis ma page professionnelle.

Le document est organisé de la façon suivante. Le chapitre 1 se concentre sur mes activités dans le cadre de la sélection de modèle et présente les résultats obtenus en termes de vitesse d'estimation. Le chapitre 2 fait la synthèse de mes travaux liés à des problèmes d'estimation de structures de covariance. Les chapitres suivants sont consacrés à la présentation de travaux plus appliqués. Le chapitre 3 décrit certaines applications développées dans le domaine aéronautique, principalement en lien avec les thèses CIFRE Airbus que j'ai codirigées. Enfin, le chapitre 4 aborde les applications dans le domaine environnemental.

Notations

\mathbb{R}	Ensemble des nombres réels
\mathbb{N}	Ensemble des entiers naturels
\mathbb{Z}	Ensemble des entiers relatifs
$ E $	Cardinal d'un ensemble E
$\mathbf{1}_E$	Fonction indicatrice de E
I_n	Matrice identité de taille $n \times n$
A^\top	Transposée de la matrice A
$\text{tr}(A)$	Trace de la matrice A
$\det(A)$	Déterminant de la matrice A
\mathbb{P}	Notation générique d'une probabilité
\mathbb{E}	Notation générique d'une espérance
Var	Notation générique de la variance
Cov	Notation générique de la covariance
$\ x\ $	Norme usuelle d'un vecteur $x \in \mathbb{R}^d$
$\lfloor t \rfloor$	Partie entière inférieure de $t \in \mathbb{R}$
∇f	Gradient de la fonction f

Curriculum vitæ

Xavier Gendre

Né le 17 août 1981 à Toulouse (France)

Membre de l'équipe *Mathématiques Appliquées* du *Département d'Ingénierie des Systèmes Complexes* de l'ISAE SUPAERO.

Chercheur associé à l'*Institut de Mathématiques de Toulouse*.

Adresse : ISAE SUPAERO, 10 avenue Édouard Belin, 31055 Toulouse, France

Email : xavier.gendre@isae-supaero.fr

Téléphone : 05 61 33 84 80

Page web : <https://personnel.isae-supaero.fr/xavier-gendre>

Situation administrative

- 2019 – Enseignant-chercheur en Statistique et Science des données à l'ISAE SUPAERO en détachement.
- 2010 – 2019 Maître de Conférences (CNU 26) à l'Université Toulouse III Paul Sabatier.
- 2009 – 2010 Attaché Temporaire d'Enseignement et de Recherche à l'Université de Provence.
- 2005 – 2009 Allocataire-moniteur puis Attaché Temporaire d'Enseignement et de Recherche à l'Université de Nice-Sophia Antipolis.

Cursus universitaire

- 2005 – 2009 Doctorat en Mathématiques à l'Université de Nice-Sophia Antipolis sous la direction de Y. Baraud.
- 2004 – 2005 DEA de Mathématiques, Université Paris-Sud XI, Orsay.
- 2002 – 2005 Magistère MMFAI, École Normale Supérieure, Paris.

Publications et communications

Voir la [liste des travaux](#).

Encadrement scientifique et enseignement

Étudiants en thèse

- 2021 – *Marelys Crespo Navas*, avec S. Gadat.
- 2021 – *Rémi Perrichon*, avec T. Klein.
- 2018 – 2021 *Fériel Boulfani*, CIFRE Airbus, avec A. Ruiz-Gazen.
- 2018 – 2021 *Ambre Diet*, CIFRE Airbus, avec N. Couellan.

Étudiants de Master

- 2021 *Marelys Crespo Navas* (M2), avec S. Gadat.
- 2018 *Nhân Đỗ Văn* (M2).
- 2018 *Rémi Mahmoud* (INSA), avec P. Casadebaig et N. Gaudio.
- 2015 *Yaya Hassan* (M1), avec A. ter Halle.

Cours à l'étranger

- 2019 *Métodos MCMC para estadísticas*, Universidad de La Habana, Cuba.
- 2018 *Introduction to Stochastic Optimization for Statistics*, Ho Chi Minh City University of Science, Vietnam.
- 2017 *Introducción a la selección de modelos*, Universidad Nacional Autónoma de México, México.
- 2016 *Introduction to Data Mining*, Ho Chi Minh City University of Science, Vietnam.
- 2014 *Estadística no paramétrica*, Universidad de La Habana, Cuba.
- 2014 *Mathematical statistics*, Ho Chi Minh City University of Science, Vietnam.
- 2011 *Introducción a la selección de modelo*, Universidad de La Habana, Cuba.
- 2004 *Analysis II* (avec R. Richard), Chennai Mathematical Institute, India.

Enseignement

ISAE SUPAERO

- 2020 – Probabilités et Statistique (1A) : cours, TD, TP.
- 2020 – Fondements théoriques de la décision statistique (3A) : cours, TP.
- 2020 – Introduction à l'apprentissage supervisé (Certificat Big Data) : cours, TP.
- 2019 – Méthodes de Monte Carlo (2A) : cours en salle informatique.
- 2019 – Algorithmes stochastiques (3A) : cours, TP.

Université Toulouse III Paul Sabatier

- 2017 – 2018 Méthodes numériques (L2) : TP.
- 2015 – 2016 Physique mathématique (L3) : TP.
- 2015 – 2017 Probabilités (L3) : cours.
- 2015 – 2019 Outils statistiques avec R et Python (M2) : TP.
- 2014 – 2016 Probabilités et Statistique (IUT) : TD.
- 2014 – 2015 Méthodes numériques (L2) : TP.
- 2013 – 2016 Mathématiques discrètes (L1) : TP.
- 2013 – 2018 Algorithmique et programmation (L2) : TP.
- 2012 – 2019 Data mining (M2) : cours, TP.
- 2010 – 2015 Statistique exploratoire (L3) : cours, TD.
- 2010 – 2014 Statistique inférentielle (L3) : cours, TD, TP.
- 2010 – 2011 Data mining (M2) : cours, TP.

Université de Provence

- 2009 – 2010 Statistique (L3) : TD.
- 2009 – 2010 Statistique pour la biologie (L3) : cours, TD.
- 2009 – 2010 Théorie des sondages (M2) : cours, TD.

ENSAE-ENSAI Formation continue (Cepe)

- Data science : bases de données (SQL et NoSQL), machine learning, pipeline de traitement de données, visualisation de données, web scraping.
- Méthodes statistiques : introduction, analyse de données, classification supervisée (analyse discriminante, régression logistique, arbres de décision, ...), méthodes d'agrégation d'estimateurs, modèles de régression généralisés.
- Logiciels statistiques / Python : initiation, intermédiaire.
- Logiciels statistiques / R : initiation, intermédiaire, Shiny, Tidyverse.

Responsabilités pédagogiques

- 2020 – Correspondant ISAE-SUPAERO pour le parcours Aléatoire du *Master 2 Recherche et Innovation* co-accrédité avec l'Université Toulouse III – Paul Sabatier.
- 2020 – Co-responsable des enseignements de *Probabilités et Statistique* dans le tronc commun 1A de l'ISAE SUPAERO.
- 2019 – Correspondant ISAE-SUPAERO pour les enseignements de mathématiques du cursus Ingénieur par Apprentissage (FISA).
- 2015 – 2018 Co-responsable UT3 du *Master 2 Statistics and Econometrics* (co-accrédité avec la *Toulouse School of Economics* de l'Université Toulouse I Capitole).
- 2011 – 2015 Co-responsable UT3 de la *L3 Statistique et Informatique Décisionnelle*.
- 2013 – 2015 Membre du comité de pilotage Coursus Master en Ingénierie pour la filière *Statistique et Informatique Décisionnelle*.

Vie scientifique et responsabilités collectives

Activités d'intérêt collectif

- 2022 – Représentant ISAE à la Fédération MA2N .
- 2015 Membre du comité de sélection MCF 0819, section 26.
- 2015 Membre du groupe de réflexion sur les « Conditions de travail des collègues féminines de l'Institut de Mathématiques de Toulouse ».
- 2014 – 2015 Membre nommé du conseil de l'Institut de Mathématiques de Toulouse.
- 2013 – 2014 Suppléant élu au Comité Technique d'Établissement de l'Université Toulouse III Paul Sabatier.
- 2011 – 2014 Membre du comité informatique de l'Institut de Mathématiques de Toulouse.

Événements scientifiques

- 2022 – Membre du comité scientifique des *Journées Statistiques du Sud*.
- 2017 – 2019 Membre du comité local d'organisation de *UseR! 2019*.
- 2015 – 2016 Membre du comité local d'organisation des *5èmes Rencontres R*.
- 2012 Membre du comité local d'organisation des *6èmes Journées Statistiques du Sud*.

Événements locaux

- 2017 Organisation (avec A. Brault et C. Pellegrini) de la journée de rentrée des doctorants de l'équipe *Statistique et Probabilités* de l'*Institut de Mathématiques de Toulouse*.
- 2014 Organisation (avec B. Huou et T. Madaule) de la journée de rentrée des doctorants de l'équipe Statistique et Probabilités de l'Institut de Mathématiques de Toulouse.
- 2011 – 2013 Co-responsable du séminaire de Statistique de l'*Institut de Mathématiques de Toulouse*.

Invitations scientifiques

- École thématique *Traitement des données* dans le cadre du programme ECOS Nord, Universidad Nacional Autónoma de México, México, 2017.
- Séjour scientifique de 3 mois au sein de l'unité *Mathématique, Informatique et Génome* de l'INRA de Jouy-en-Josas, 2009.

Projets scientifiques

- 2017 Partenaire du projet INRA de Pari Scientifique *IDEA (Intra- and interspecific Diversity mixturE in Agriculture)*.

Jury du baccalauréat : présidence en 2011, 2012, 2013 et 2016.

Chapitre 1

Sélection de modèle pour la régression

La sélection de modèle est le domaine de la statistique mathématique où j'ai produit mes premiers travaux de recherche. Ce chapitre présente ainsi des résultats développés dans le cadre de ma thèse [XG-Thesis] et dans son prolongement qui ont fait l'objet des publications [XG-Article01], [XG-Article02] et [XG-Article03]. Outre les questions de sélection de modèle, nous présentons également le rôle de la concentration de la mesure pour cette théorie et les propriétés qui découlent des inégalités oracles, en particulier le caractère adaptatif des estimateurs considérés.

1.1 Cadre général

À partir de l'observation d'un échantillon aléatoire de loi P inconnue, la problématique générale de l'apprentissage statistique consiste à estimer un objet $s^* \in \mathcal{S}$ relatif à cette loi. Pour cela, nous considérons une fonction de contraste $\gamma : \mathcal{S} \rightarrow \mathbb{R}$ qui ne dépend que de l'échantillon telle que son espérance sous la loi P ,

$$s \in \mathcal{S} \mapsto \mathbb{E}[\gamma(s)],$$

est minimale en s^* . Pour tout contraste, il est ainsi possible de définir une fonction de perte ℓ par

$$\forall s \in \mathcal{S}, \ell(s, s^*) = \mathbb{E}[\gamma(s)] - \mathbb{E}[\gamma(s^*)] \geq 0.$$

Cette quantité quantifie la qualité d'un candidat $s \in \mathcal{S}$ pour représenter s^* .

De nombreux problèmes statistiques se formulent de cette façon. L'apprentissage supervisé est un cas particulier où les observations sont des

couples $(X_1, Y_1), \dots, (X_n, Y_n)$ pour lesquels nous cherchons à prédire la variable de sortie Y à partir de la variable d'entrée X . Lorsque la variable de sortie est continue, nous parlons de régression et l'objectif devient l'estimation d'une fonction s^* reliant la variable d'entrée à celle de sortie. Le cadre statistique est alors souvent considéré sous la forme d'un signal perturbé par un bruit centré ε ,

$$Y_i = s^*(X_i) + \varepsilon_i, \quad i \in \{1, \dots, n\}.$$

Des fonctions de contrastes largement utilisées dans cette situation sont les moindres carrés et la vraisemblance.

Pour répondre au problème de la régression, nous devons construire un estimateur $\hat{s} \in \mathcal{S}$ de la cible s^* uniquement à partir de l'échantillon. Une approche classique consiste à minimiser l'erreur commise sur les observations au sens du risque empirique donné par la fonction de contraste. En effet, en minimisant γ , il est loisible de penser que nous pouvons obtenir un objet proche du minimiseur de ℓ puisque cette fonction est essentiellement donnée par l'espérance de γ sous la loi P inconnue. Cependant, l'espace \mathcal{S} est généralement si grand qu'il est très souvent possible d'y trouver un élément $\hat{s}_{\mathcal{S}}$ capable de ne commettre presque aucune erreur sur les observations malgré leur nature aléatoire. Ce phénomène de surapprentissage n'est pas désirable en pratique car la solution est alors si propre aux données qu'elle produirait de piètres prédictions sur de nouvelles observations, ce qui se traduit par un risque $\mathbb{E}[\ell(\hat{s}_{\mathcal{S}}, s^*)]$ important.

Pour pallier ce problème, il faut se restreindre à un modèle S , *i.e.* un sous-ensemble de \mathcal{S} , afin de considérer un estimateur par minimum de contraste \hat{s}_S tel que

$$\hat{s}_S \in \operatorname{argmin}_{s \in S} \gamma(s).$$

La difficulté avec cette approche consiste à trouver un modèle adéquat pour assurer de bonnes propriétés à l'estimateur par minimum de contraste associé. L'idée de la sélection de modèle est de considérer une collection de modèles $\{S_m, m \in \mathcal{M}\}$ ainsi que les estimateurs par minimum de contraste $\hat{s}_m \in S_m$, $m \in \mathcal{M}$, et de construire une procédure basée sur les données pour choisir le meilleur estimateur de s^* parmi eux.

Au sein de \mathcal{M} , il existe un indice m^* dont le risque est minimal,

$$m^* \in \operatorname{argmin}_{m \in \mathcal{M}} \mathbb{E}[\ell(\hat{s}_m, s^*)].$$

L'objet \hat{s}_{m^*} est appelé l'oracle en référence à [DJ94] et il est important de noter qu'il ne s'agit pas d'un estimateur puisque sa construction dépend de la loi P inconnue au travers de l'espérance. Cependant, son risque étant minimal

parmi ceux des estimateurs \hat{s}_m , il sert de référence et l'objectif devient de choisir un estimateur dont le risque soit du même ordre de grandeur que celui de l'oracle.

Pour faire ce choix dans la suite, nous procédons par pénalisation. Étant donnée une fonction de pénalité $\text{pen} : \mathcal{M} \rightarrow \mathbb{R}_+$, nous définissons l'indice $\hat{m} \in \mathcal{M}$ comme un minimiseur du contraste pénalisé,

$$\hat{m} \in \operatorname{argmin}_{m \in \mathcal{M}} \{ \gamma(\hat{s}_m) + \text{pen}(m) \},$$

et nous posons $\tilde{s} = \hat{s}_{\hat{m}}$. Les premiers résultats obtenus par contraste pénalisé sont dus à Akaike [Aka70] pour l'estimation de densité par vraisemblance pénalisée (critère AIC) et à Mallows [Mal73] pour l'estimation de la fonction de régression dans un cadre Gaussien homoscédastique à variance connue. Les modèles considérés dans ces travaux historiques sont des espaces linéaires de dimensions finies et les pénalités y sont proportionnelles à la dimension du modèle. Birgé et Massart [BM01] ont généralisé l'approche de Mallows et ils ont obtenu un contrôle non asymptotique du risque de \tilde{s} en lien avec la taille de la collection de modèles. C'est dans cette perspective que s'inscrivent les travaux présentés dans la suite de ce chapitre.

Afin de valider théoriquement la qualité de l'estimateur sélectionné \tilde{s} , nous cherchons généralement à établir des inégalités de la forme

$$\mathbb{E}[\ell(\tilde{s}, s^*)] \leq C_1 \inf_{m \in \mathcal{M}} \left\{ \inf_{s \in S_m} \ell(s, s^*) + \text{pen}(m) \right\} + C_2 \quad (1.1)$$

où C_1 et C_2 sont des quantités à déterminer. Lorsque la somme dans l'infimum est de l'ordre du risque $\mathbb{E}[\ell(\hat{s}_m, s^*)]$ de l'estimateur \hat{s}_m et sous certaines hypothèses sur la collection de modèles, de tels résultats permettent de comparer le risque de \tilde{s} à celui de l'oracle en donnant lieu à des inégalités dites oracles,

$$\mathbb{E}[\ell(\tilde{s}, s^*)] \leq C \inf_{m \in \mathcal{M}} \mathbb{E}[\ell(\hat{s}_m, s^*)] = C \mathbb{E}[\ell(\hat{s}_{m^*}, s^*)].$$

Le choix de la fonction de pénalité est donc crucial pour permettre ce raisonnement. Par définition, nous savons que, pour tout $m \in \mathcal{M}$ et tout $s_m \in S_m$,

$$\gamma(\tilde{s}) + \text{pen}(\hat{m}) \leq \gamma(\hat{s}_m) + \text{pen}(m) \leq \gamma(s_m) + \text{pen}(m).$$

Il vient alors la majoration fondamentale suivante,

$$\ell(\tilde{s}, s^*) \leq \ell(s_m, s^*) + \text{pen}(m) + Z(s_m) - Z(\tilde{s}) - \text{pen}(\hat{m})$$

où, pour tout $s \in \mathcal{S}$, $Z(s) = \gamma(s) - \mathbb{E}[\gamma(s)]$. Nous voyons ainsi que la pénalité doit être suffisamment grande pour annihiler les variations de $Z(s_m) - Z(\tilde{s})$ mais aussi assez petite pour satisfaire une inégalité de la forme

$$\ell(s_m, s^*) + \text{pen}(m) \leq \mathbb{E}[\ell(\hat{s}_m, s^*)].$$

L'intérêt de la concentration de la mesure pour la sélection de modèle est précisément de permettre un contrôle fin des variations de $Z(s) - Z(s')$ lorsque s et s' sont des éléments proches dans un modèle donné.

Obtenir une inégalité oracle pour une procédure de sélection de modèle assure que l'estimateur choisi admet bien un risque du même ordre de grandeur que le choix optimal que nous aurions pu faire si la loi P était connue. Cependant, de telles inégalités ont l'inconvénient de ne comparer le risque de \tilde{s} qu'avec les risques des estimateurs \hat{s}_m , $m \in \mathcal{M}$. D'un point de vue statistique, il est souhaitable de pouvoir discuter de ce risque par rapport à des classes d'estimateurs plus générales. Pour ce faire, une approche usuelle consiste à considérer le risque maximal sur un certain espace \mathcal{T}_α caractérisé par une propriété dépendante d'un paramètre $\alpha \in A$ (e.g. la régularité de la fonction s^*). Le point de vue minimax revient à dire qu'un estimateur est bon si son risque maximal sur \mathcal{T}_α est proche du risque dit minimax défini par

$$R(\mathcal{T}_\alpha, \ell) = \inf_T \sup_{s^* \in \mathcal{T}_\alpha} \mathbb{E}[\ell(T, s^*)]$$

où l'infimum porte sur tous les estimateurs de s^* dont ceux qui utilisent la connaissance de α . L'estimateur \tilde{s} est dit adaptatif au sens du minimax pour le paramètre α si

$$\forall \alpha \in A, \exists C_\alpha > 1 \text{ tel que } \sup_{s^* \in \mathcal{T}_\alpha} \mathbb{E}[\ell(\tilde{s}, s^*)] \leq C_\alpha R(\mathcal{T}_\alpha, \ell)$$

où C_α est un facteur autorisé à dépendre de la loi P et de α mais pas du nombre n des observations. Autrement dit, le risque d'un estimateur adaptatif est du même ordre que le risque optimal sur \mathcal{T}_α mais sa construction n'utilise pas la connaissance de α et cette propriété demeure vraie quelque soit la valeur de ce paramètre. Au-delà du choix d'un estimateur, les procédures de sélection de modèle qui vérifient une inégalité oracle sont intéressantes car elles permettent de construire des estimateurs adaptatifs au sens du minimax (voir [BM97]).

Les sections suivantes présentent mes travaux autour de la sélection de modèle. Pour chacun d'entre eux, les fonctions de contraste, de perte et de pénalité sont explicitées ainsi que les résultats obtenus. Les preuves et les études numériques sont omises et pourront être trouvées dans le corps des articles.

1.2 Régression Gaussienne hétéroscédastique

Cette section présente des résultats publiés dans [XG-Article01].

Cadre statistique

Le cadre non paramétrique de la régression Gaussienne hétéroscédastique correspond à l'observation de couples $(X_1, Y_1), \dots, (X_n, Y_n) \in \mathbb{R}^2$ tels que

$$Y_i = s^*(X_i) + \sqrt{\sigma^*(X_i)}\varepsilon_i, \quad i \in \{1, \dots, n\}, \quad (1.2)$$

où les fonctions $s^* : \mathbb{R} \rightarrow \mathbb{R}$ et $\sigma^* : \mathbb{R} \rightarrow \mathbb{R}_+^*$ sont inconnues et les variables aléatoires $\varepsilon_1, \dots, \varepsilon_n$ sont indépendantes et de même loi Gaussienne standard. La situation où les entrées X_1, \dots, X_n sont déterministes est dite à support fixé par opposition au cas aléatoire. Nous renvoyons à [Bar00] et [Bar02] pour une discussion sur l'approche de la sélection de modèle selon la nature du support. Dans la suite de cette section, nous considérons le problème à support fixé et nous introduisons les vecteurs inconnus $s = (s_1, \dots, s_n)^\top$ et $\sigma = (\sigma_1, \dots, \sigma_n)^\top$ définis par

$$s_i = s^*(X_i) \in \mathbb{R} \text{ et } \sigma_i = \sigma^*(X_i) > 0, \quad i \in \{1, \dots, n\}.$$

En notant $P_{s,\sigma}$ la loi du vecteur $Y = (Y_1, \dots, Y_n)^\top$ avec

$$Y_i = s_i + \sqrt{\sigma_i}\varepsilon_i, \quad i \in \{1, \dots, n\},$$

nous introduisons la fonction de perte donnée par la divergence de Kullback-Leibler entre les lois $P_{s,\sigma}$ et $P_{t,\tau}$, pour tous vecteurs $t = (t_1, \dots, t_n)^\top \in \mathbb{R}^n$ et $\tau = (\tau_1, \dots, \tau_n)^\top \in (0, +\infty)^n$,

$$\mathcal{K}(P_{s,\sigma}, P_{t,\tau}) = \frac{1}{2} \sum_{i=1}^n \frac{(s_i - t_i)^2}{\tau_i} + \phi\left(\frac{\tau_i}{\sigma_i}\right)$$

où, pour tout $u > 0$, $\phi(u) = \log u + 1/u - 1$.

Considérons deux copies indépendantes $Y^{[1]}$ et $Y^{[2]}$ de loi $P_{s,\sigma}$, notre objectif est l'estimation de la paire de vecteurs (s, σ) par vraisemblance pénalisée. Le contraste sous-jacent est donc la fonction de vraisemblance qui peut être définie pour chacun des deux vecteurs observés,

$$\gamma_n^{[k]}(t, \tau) = \frac{1}{2} \sum_{i=1}^n \frac{(Y_i^{[k]} - t_i)^2}{\tau_i} + \log \tau_i, \quad k \in \{1, 2\}.$$

L'estimation de s se fait en minimisant $\gamma_n^{[1]}$ et celle de σ en minimisant $\gamma_n^{[2]}$. La nécessité de deux copies indépendantes est essentiellement motivée par le fait d'avoir des estimateurs indépendants pour la moyenne et la variance comme ce serait le cas à variance constante. En pratique, une alternative consiste à considérer un point sur deux dans (1.2) pour le premier échantillon et les autres pour le second.

Modèles et estimateurs

Au-delà du problème d'estimation, notre objectif est d'identifier des segments du vecteur de loi $P_{s,\sigma}$ où la variance est à peu près constante. Pour ce faire, nous introduisons des modèles de paires de vecteurs définis par morceaux de la façon suivante. Par souci de simplicité, nous supposons que $n = 2^N$ et, pour tout m dans un ensemble \mathcal{M} au plus dénombrable, nous associons une partition p_m de $\{1, \dots, n\}$ donnée par les $|p_m| = 2^{k_m}$ blocs consécutifs

$$\{(i-1)2^{N-k_m} + 1, \dots, i2^{N-k_m}\}, \quad i \in \{1, \dots, |p_m|\}.$$

Pour chaque bloc $I \in p_m$ et tout $x \in \mathbb{R}^n$, nous écrivons $x|_I$ pour désigner le vecteur de $\mathbb{R}^{|p_m|}$ donné par les coordonnées $(x_i)_{i \in I}$. Nous pouvons également associer à m un sous-espace linéaire $E_m \subset \mathbb{R}^{|p_m|}$ de dimension $d_m \in \{1, \dots, 2^{N-k_m}\}$. Cet espace E_m est laissé au choix du statisticien et il peut, par exemple, être engendré par d_m vecteurs orthnormaux. Notons également que plusieurs $m \in \mathcal{M}$ peuvent être construits sur les mêmes morceaux mais avec des espaces E_m distincts. Nous pouvons maintenant considérer l'espace $S_m \subset \mathbb{R}^n$ d'estimation de la moyenne par des vecteurs définis par morceaux,

$$S_m = \{x \in \mathbb{R}^n \text{ tels que } \forall I \in p_m, x|_I \in E_m\},$$

et l'espace $\Sigma_m \subset \mathbb{R}^n$ d'estimation de la variance par des vecteurs constants par morceaux,

$$\Sigma_m = \left\{ \sum_{I \in p_m} g_I \mathbf{1}_I \text{ avec } \forall I \in p_m, g_I > 0 \right\}.$$

La dimension de $S_m \times \Sigma_m$ est notée $D_m = |p_m|(d_m + 1)$ et la collection de modèles que nous retenons pour l'estimation de (s, σ) est $\{S_m \times \Sigma_m, m \in \mathcal{M}\}$.

Pour chaque $m \in \mathcal{M}$, la paire d'estimateurs par minimum de contraste $(\hat{s}_m, \hat{\sigma}_m) \in S_m \times \Sigma_m$ est définie par

$$\hat{s}_m = \pi_m Y^{[1]}$$

et

$$\hat{\sigma}_m = \sum_{I \in p_m} \hat{\sigma}_{m,I} \mathbf{1} \text{ avec } \forall I \in p_m, \hat{\sigma}_{m,I} = \frac{1}{|I|} \sum_{i \in I} (Y^{[2]} - \pi_m Y^{[2]})_i^2$$

où π_m désigne la projection orthogonale dans E_m . Comme discuté précédemment, les estimateurs \hat{s}_m et $\hat{\sigma}_m$ sont indépendants par construction.

Majoration du risque

Notre procédure ne fait aucune hypothèse sur le vecteur s mais demande la connaissance d'un majorant $\rho \geq 1$ pour le rapport

$$\frac{\max\{\sigma_1, \dots, \sigma_n\}}{\min\{\sigma_1, \dots, \sigma_n\}} \leq \rho.$$

Cette borne quantifie le degré d'hétéroscédasticité au sens où « $\rho = 1$ » correspond au cas de la variance constante tel que traité dans [BGH09] et « $\rho > 1$ » à celui où la variance est autorisée à fluctuer.

Étant donnée une fonction de pénalité $\text{pen} : \mathcal{M} \rightarrow \mathbb{R}_+$, nous choisissons un modèle $\hat{m} \in \mathcal{M}$ par vraisemblance pénalisée,

$$\hat{m} \in \operatorname{argmin}_{m \in \mathcal{M}} \left\{ \gamma_n^{[1]}(\hat{s}_m, \hat{\sigma}_m) + \text{pen}(m) \right\},$$

et nous notons $(\tilde{s}, \tilde{\sigma}) = (\hat{s}_{\hat{m}}, \hat{\sigma}_{\hat{m}})$ la paire d'estimateurs sélectionnée.

Afin d'obtenir une inégalité de la forme (1.1), nous devons introduire deux hypothèses sur la taille de la collection de modèles. La première hypothèse est une majoration de la dimension de chaque modèle en fonction de la taille n des données de façon à assurer que chaque bloc contienne assez de points pour que les estimateurs soient correctement définis :

$$(\mathbf{H}_{\text{Dim}}) \exists \theta > 1, \exists \delta, \epsilon > 0, \max_{m \in \mathcal{M}} D_m \leq \min \left\{ \frac{(\theta - 1)n}{\theta(\gamma + 2)}, \frac{5\delta\gamma n}{\log^{1+\epsilon} n} \right\}.$$

La seconde hypothèse limite le nombre de modèles ayant la même dimension au sein de la collection :

$$(\mathbf{H}_{\text{Card}}) \exists A, B > 0, \forall (k, d) \in \mathbb{N}^2, \left| \left\{ m \in \mathcal{M}, D_m = 2^k(d+1) \right\} \right| \leq A(d+1)^B.$$

Théorème 1.1 *Sous les hypothèses $(\mathbf{H}_{\text{Dim}})$ et $(\mathbf{H}_{\text{Card}})$ et pour la fonction de pénalité*

$$\forall m \in \mathcal{M}, \text{pen}(m) = \left(\gamma\theta + \log^{1+\epsilon} D_m \right) D_m,$$

il existe des constantes $C_1, C_2 > 0$ telles que

$$\mathbb{E}[\mathcal{K}(P_{s,\sigma}, P_{\tilde{s},\tilde{\sigma}})] \leq C_1 \inf_{m \in \mathcal{M}} \left\{ \inf_{(t,\tau) \in S_m \times \Sigma_m} \mathcal{K}(P_{s,\sigma}, P_{t,\tau}) + D_m \log^{1+\epsilon} D_m \right\} + C_2.$$

Le résultat de ce théorème est proche d'une inégalité oracle à un facteur logarithmique près et, comme nous l'avons évoqué dans la première section, cela permet de déduire des propriétés d'adaptativité pour la paire d'estimateurs $(\tilde{s}, \tilde{\sigma})$.

Adaptativité

Nous revenons maintenant au problème non paramétrique de l'estimation de la paire de fonctions (s^*, σ^*) dans (1.2). Pour $\alpha \in (0, 1)$ et $L > 0$, la classe de fonctions que nous considérons est la boule de Hölder d'ordre α et de rayon L sur $[0, 1]$,

$$\mathcal{H}_\alpha(L) = \{f : [0, 1] \rightarrow \mathbb{R} : \forall x, y \in [0, 1], |f(x) - f(y)| \leq L|x - y|^\alpha\}. \quad (1.3)$$

Pour la collection de modèles, nous nous limitons à des histogrammes en ne considérant que des espaces E_m de vecteurs constants par morceaux. En particulier, l'hypothèse $(\mathbf{H}_{\text{Card}})$ est satisfaite avec $A = 1$ et $B = 0$. Les propriétés d'approximation de tels espaces sont suffisantes pour déduire le résultat suivant sur la paire d'estimateurs $(\tilde{s}_H, \tilde{\sigma}_H)$ sélectionnée.

Proposition 1.2 *Soient $\alpha_1, \alpha_2 \in (0, 1)$ et $L_1, L_2 > 0$. Sous l'hypothèse $(\mathbf{H}_{\text{Dim}})$, si la taille des données vérifie*

$$n \geq \max \left\{ \frac{4(\inf \sigma^*)^4}{(L_1^2 \inf \sigma^* + L_2^2)^2}, e^{4(1+\epsilon)^2} \right\},$$

alors il existe une constante $C > 0$ telle que

$$\sup_{(s^*, \sigma^*) \in \mathcal{H}_{\alpha_1}(L_1) \times \mathcal{H}_{\alpha_2}(L_2)} \mathbb{E} \left[\frac{\mathcal{K}(P_{s, \sigma}, P_{\tilde{s}_H, \tilde{\sigma}_H})}{n} \right] \leq C \left(\frac{n}{\log^{1+\epsilon} n} \right)^{-2\alpha/(2\alpha+1)}$$

où $\alpha = \min\{\alpha_1, \alpha_2\}$.

À partir d'une observation du vecteur Y et pour une perte quadratique, le risque minimax de l'estimation de la moyenne $s^* \in \mathcal{H}_{\alpha_1}(L_1)$ est de l'ordre de $n^{-2\alpha_1/(2\alpha_1+1)}$ [GP05] et celui de la variance $\sigma^* \in \mathcal{H}_{\alpha_2}(L_2)$ est de l'ordre de $\max\{n^{-4\alpha_1}, n^{-2\alpha_2/(2\alpha_2+1)}\}$ [WBCL08]. Le maximum de ces risques est donc de l'ordre de $n^{-2\alpha/(2\alpha+1)}$ pour la plus mauvaise régularité $\alpha = \min\{\alpha_1, \alpha_2\}$. À un facteur logarithmique près, il s'agit bien de ce que nous obtenons pour la divergence de Kullback-Leibler.

1.3 Courbes décalées

Cette section présente des résultats publiés dans [XG-Article02].

Cadre statistique

Nous considérons les perturbations aléatoires de J courbes $s_1, \dots, s_J : [0, 1] \rightarrow \mathbb{R}$ échantillonnées sur n points équirépartis $t_i = i/n$ pour $i \in \{1, \dots, n\}$,

$$Y_{i,j} = s_j(t_i) + \varepsilon_{i,j}, \quad i \in \{1, \dots, n\}, \quad j \in \{1, \dots, J\},$$

où les variables $\varepsilon_{i,j}$ sont supposées indépendantes et de même loi Gaussienne centrée de variance $\sigma^2 > 0$ connue. Dans de nombreuses applications, les courbes observées varient autour d'un même motif $s^* : [0, 1] \rightarrow \mathbb{R}$ que nous souhaitons estimer.

Deux variabilités sont ainsi à l'œuvre dans ce cadre : un bruit additif ε et une déformation géométrique de s^* . Dans l'approche répandue de Grenander [GM07], cette variabilité géométrique est modélisée par l'action d'un groupe de Lie sur l'espace des courbes. Lorsque la déformation affecte le support des observations, cela complique la construction de bons estimateurs du motif moyen s^* .

Dans cette section, nous considérons la déformation simple du décalage aléatoire et nous renvoyons à [RS02] pour une motivation complète de ce modèle,

$$Y_{i,j} = s^*(t_i - \theta_j^*) + \varepsilon_{i,j}, \quad i \in \{1, \dots, n\}, \quad j \in \{1, \dots, J\}, \quad (1.4)$$

où les variables θ_j^* sont indépendantes, de même loi sur \mathbb{R} de densité g inconnue et indépendantes des $\varepsilon_{i,j}$. Dans la suite, le motif moyen inconnu $s^* : [0, 1] \rightarrow \mathbb{R}$ est supposé appartenir à l'espace $L_{\text{per}}^2([0, 1])$ des fonctions de période 1 et de carré intégrable.

Bien que cela ne soit pas explicite dans cette présentation des résultats, le caractère périodique des fonctions conduit naturellement à travailler avec les coefficients de Fourier,

$$\forall f \in L_{\text{per}}^2([0, 1]), \quad \forall k \in \mathbb{Z}, \quad c_k(f) = \int_0^1 f(t) e^{-i2\pi kt} dt.$$

En particulier, nous considérerons dans la suite les espaces de fonctions non constantes de régularité $\alpha > 1/2$,

$$\tilde{W}_\alpha(A, c_*) = \left\{ f \in L_{\text{per}}^2([0, 1]) : \sum_{k \in \mathbb{Z}} (1 + |k|^{2\alpha}) |c_k(f)|^2 \leq A^2 \text{ avec } |c_1(f)| \geq c_* \right\}$$

où $A, c_* > 0$. La définition de ces espaces $\tilde{W}_\alpha(A, c_*)$ est motivée par celle des boules de Sobolev. La condition additionnelle « $|c_1(f)| \geq c_*$ » est nécessaire pour des raisons d'identifiabilité.

Moyenne de Fréchet

La moyenne empirique des courbes observées ne fournit évidemment pas une estimation consistante de s^* et les courbes doivent être alignées au préalable. Notons qu'il n'y a pas unicité de cet alignement, ce qui pose un problème d'identifiabilité pour estimer s^* . De plus, par périodicité, les décalages peuvent être supposés à valeurs dans $[-1/2, 1/2]$. Ces remarques conduisent à considérer le problème au sens des orbites sous l'action du décalage,

$$\forall f \in L^2_{\text{per}}([0, 1]), [f] = \{t \mapsto f(t - \theta) \text{ pour } \theta \in [-1/2, 1/2]\}.$$

La fonction de perte que nous considérons est ainsi donnée par le plus petit écart quadratique entre deux représentants des orbites,

$$\forall f_1, f_2 \in L^2_{\text{per}}([0, 1]), d^2([f_1], [f_2]) = \inf_{\theta \in [-1/2, 1/2]} \int_0^1 |f_1(t - \theta) - f_2(t)|^2 dt.$$

La moyenne de Fréchet est une généralisation de la moyenne Euclidienne au sens du minimiseur d'un critère de variabilité. Si $f_1, \dots, f_J \in L^2_{\text{per}}([0, 1])$, la moyenne de Fréchet $[f]$ des orbites $[f_1], \dots, [f_J]$ est définie par

$$[f] \in \operatorname{argmin}_{[f]} \frac{1}{J} \sum_{j=1}^J d^2([f], [f_j]).$$

Il est simple de voir qu'un représentant $\bar{f} \in ([f_1], [f_2])$ de l'orbite $[f]$ peut être construit en deux étapes :

1. Calcul des décalages :

$$(\tilde{\theta}_1, \dots, \tilde{\theta}_J) \in \operatorname{argmin}_{(\theta_1, \dots, \theta_J) \in [-1/2, 1/2]^J} \frac{1}{J} \sum_{j=1}^J \int_0^1 \left| f_j(t + \theta_j) - \frac{1}{J} \sum_{j'=1}^J f_{j'}(t + \theta_{j'}) \right|^2 dt.$$

2. Alignement des courbes et moyenne :

$$\bar{f}(t) = \frac{1}{J} \sum_{j=1}^J f_j(t + \tilde{\theta}_j), \quad t \in [0, 1].$$

Estimation des décalages

Pour des raisons d'identifiabilité, certaines hypothèses sont nécessaires pour estimer le vecteur des décalages $\theta^* = (\theta_1^*, \dots, \theta_J^*)^\top$. Plusieurs conditions sont envisageables et celles que nous retenons dans cette section sont :

(**H_{supp}**) Il existe $0 < \kappa < 1/8$ tel que la densité g des décalages aléatoires est à support compact dans $[-\kappa/2, \kappa/2]$.

(**H_{nc}**) Le motif moyen s^* est tel que $c_1(s^*) = \int_0^1 s^*(x)e^{-i2\pi x} dx \neq 0$.

L'hypothèse (**H_{supp}**) impose que les décalages demeurent petits et que les courbes observées soient relativement concentrées autour du motif moyen. Une telle hypothèse est commune dans la littérature pour déduire l'unicité et la consistance des moyennes de Fréchet bien qu'elle puisse être légèrement relâcher. L'hypothèse (**H_{nc}**) permet d'éviter le cas d'une fonction s^* constante sur $[0, 1]$ qui rendrait l'estimation des décalages impossible.

L'estimation des décalages a vocation à être utilisée pour recaler les courbes observées et estimer le motif moyen s^* . Afin d'étudier cette procédure en deux temps, il est théoriquement plus confortable de se placer dans le cas d'estimateurs indépendants. Si la taille $n = 2N$ du support est paire, le cadre (1.4) permet facilement de le faire en séparant les données une observation sur deux,

$$Y_{i,j}^{(k)} = Y_{2i-k,j}, \quad i \in \{1, \dots, N\}, \quad j \in \{1, \dots, J\}, \quad k \in \{0, 1\}.$$

Pour chaque courbe, les vecteurs $Y^{(0)}$ et $Y^{(1)}$ sont indépendants et l'un est utilisé pour estimer les décalages tandis que l'autre le sera pour estimer le motif moyen. Dans ce qui suit, pour $k \in \{0, 1\}$, nous notons $\mathbb{E}^{(k)}$ pour désigner l'espérance sous la loi de $Y^{(k)}$.

Pour $0 < \kappa < 1/8$, nous introduisons l'espace des décalages centrés

$$\Theta_\kappa = \left\{ (\theta_1, \dots, \theta_J)^\top \in [-\kappa/2, \kappa/2]^J \text{ avec } \sum_{j=1}^J \theta_j = 0 \right\}$$

et la cible identifiable $\theta^0 = (\theta_1^0, \dots, \theta_J^0)^\top \in \Theta_\kappa$ donnée par

$$\theta_j^0 = \theta_j^* - \frac{1}{J} \sum_{j'=1}^J \theta_{j'}^*, \quad j \in \{1, \dots, J\}.$$

En minimisant le critère de variabilité de la première étape du calcul de la moyenne de Fréchet à partir des $k_0 \geq 1$ premiers coefficients de Fourier des courbes observées $Y_{\cdot,1}^{(0)}, \dots, Y_{\cdot,J}^{(0)}$, nous obtenons un estimateur des décalages $\hat{\theta}^0 = (\hat{\theta}_1^0, \dots, \hat{\theta}_J^0)^\top \in \Theta_\kappa$ qui vérifie le résultat suivant.

Théorème 1.3 *Sous les hypothèses (**H_{supp}**) et (**H_{nc}**) et pour tout $J \geq 2$, $\alpha \geq 2$ et $A, c_* > 0$, il existe une constante $C > 0$ telle que si $s^* \in \tilde{W}_\alpha(A, c_*)$, alors*

$$\frac{1}{J} \mathbb{E}^{(0)} [\|\hat{\theta}^0 - \theta^0\|^2] \leq \frac{C}{n} \left(1 + \frac{k_0^5}{n^{1/2}} \right) \left(1 + \frac{k_0^{3/2} J^3}{n^{1/2}} \right). \quad (1.5)$$

Si le nombre J de courbes est fixe et que la taille n du support tend vers l'infini, le problème d'estimation des décalages est paramétrique et nous retrouvons par ce résultat la vitesse d'estimation usuelle d'ordre n^{-1} . Cependant, si n et J augmentent tous les deux, nous obtenons que la vitesse de convergence donnée par la borne (1.5) dépend d'un rapport entre des puissances de n et J . Cela montre que l'estimation des décalages n'est pas un simple problème paramétrique dans le cadre doublement asymptotique.

Le corollaire immédiat du Théorème 1.3 est que si le nombre de courbes J est de l'ordre de n^ρ pour $\rho \in (0, 1/6]$, alors il existe une constante $C_\rho > 0$ telle que

$$\frac{1}{J} \mathbb{E}^{(0)} [\|\hat{\theta}^0 - \theta^0\|^2] \leq \frac{C_\rho}{n}.$$

Une bonne estimation des décalages est nécessaire pour obtenir des propriétés sur la moyenne de Fréchet. Pour cela, nous voyons que le nombre de courbes observées ne doit pas être trop grand par rapport à la taille n du support.

Estimation du motif moyen

Soit $j \in \{1, \dots, J\}$, les coefficients de Fourier empiriques $(\hat{c}_{k,j}^{(1)})_{k \in \mathbb{Z}}$ peuvent être estimés à partir des observations $Y_{1,j}^{(1)}, \dots, Y_{n,j}^{(1)}$. Par construction, ces quantités sont indépendantes du vecteur d'estimateurs $\hat{\theta}^0$ utilisé pour recaler les courbes. Étant donné un nombre $m \in \mathcal{M} = \{1, \dots, N\}$ de coefficients de Fourier, nous pouvons définir la moyenne de Fréchet régularisée,

$$\hat{s}_m(t) = \sum_{|k| \leq m} \left(\frac{1}{J} \sum_{j=1}^J \hat{c}_{k,j}^{(1)} e^{i2\pi k \hat{\theta}_j^0} \right) e^{i2\pi kt}, \quad t \in [0, 1].$$

En tant que projection sur l'espace S_m engendré par les m premières fonctions de la base de Fourier, l'estimateur \hat{s}_m est l'estimateur par minimum de contraste donné par les moindres carrés.

L'approche de la sélection de modèle permet ainsi de choisir un estimateur parmi la collection $\{\hat{s}_m, m \in \mathcal{M}\}$ uniquement à partir des données,

$$\hat{m} \in \operatorname{argmin}_{m \in \mathcal{M}} \left\{ \int_0^1 |\hat{s}_N(t) - \hat{s}_m(t)|^2 dt + \eta \frac{(2m+1)\sigma^2}{NJ} \right\}$$

où $\eta > 1$ est un paramètre de régularisation. La pénalité proportionnelle à la dimension est classique pour une telle procédure et nous renvoyons au chapitre 4 de [Mas07] pour plus de détails. En particulier, l'estimateur sélec-

tionné $\tilde{s} = \hat{s}_{\tilde{m}}$ vérifie une inégalité oracle,

$$\begin{aligned} & \mathbb{E}^{(1)} \left[\int_0^1 \left| \mathbb{E}^{(1)}[\hat{s}_N(t)] - \tilde{s}(t) \right|^2 dt \right] \\ & \leq C_\eta \inf_{m \in \mathcal{M}} \left\{ \inf_{f \in \mathcal{S}_m} \int_0^1 \left| \mathbb{E}^{(1)}[\hat{s}_N(t)] - f(t) \right|^2 dt + \frac{2(m+1)\sigma^2}{NJ} \right\} \end{aligned} \quad (1.6)$$

où $C_\eta > 0$ est une constante.

Vitesse de convergence

En exploitant conjointement la vitesse d'estimation des décalages (1.5) et l'inégalité (1.6), il est possible d'obtenir une borne supérieure sur le risque de l'estimateur \tilde{s} pour la fonction de perte d^2 .

Théorème 1.4 *Nous nous plaçons sous les hypothèses du Théorème 1.3 et supposons qu'il existe $c_1, c_2 \in (0, 1)$, $\rho \in (0, 1/6]$ et $\alpha > 3/2$ tels que*

$$c_1 n^\rho \leq J \leq c_2 n^\rho \quad \text{et} \quad nJ \geq \max \left\{ 21J, \frac{(4\sigma^2)^{2\alpha+1}}{c_2^{2\alpha}} \right\}.$$

Soit $A, c_* > 0$, il existe une constante $C > 0$ telle que

$$\sup_{g \in \mathcal{G}_\kappa} \sup_{s^* \in \tilde{W}_\alpha(A, c_*)} \mathbb{E} \left[d^2([\tilde{s}], [s^*]) \right] \leq C \left(n^{-1} + (nJ)^{-2\alpha/(2\alpha+1)} \right) \quad (1.7)$$

où \mathcal{G}_κ désigne l'ensemble des densités à support dans $[-\kappa/2, \kappa/2]$.

À travers la somme dans la borne supérieures (1.7), nous retrouvons les deux sources de variabilité du cadre statistique (1.4). En effet, le terme n^{-1} peut être interprété comme le coût du recalage des courbes bruitées tandis que le terme $(nJ)^{-2\alpha/(2\alpha+1)}$ correspond à la vitesse minimax d'estimation en l'absence de décalage mais avec bruit additif.

Il n'existe pas de résultat général concernant la vitesse minimax de convergence d'une moyenne de Fréchet. Cependant, nous avons pu démontrer la borne inférieure suivante dans le cas des courbes décalées.

Théorème 1.5 *Prenons $\alpha > 1/2$ et $A, c_* > 0$, alors il existe une constante $C > 0$ qui ne dépend que de α, A, c_* et σ^2 telle que*

$$\liminf_{\min\{n, J\} \rightarrow \infty} (nJ)^{2\alpha/(2\alpha+1)} \inf_T \sup_{g \in \mathcal{G}_\kappa} \sup_{s^* \in \tilde{W}_\alpha(A, c_*)} \mathbb{E} \left[d^2([T], [s^*]) \right] \geq C$$

où l'infimum porte sur tous les estimateurs de s^* .

Sous les hypothèses du Théorème 1.4, nous avons

$$(nJ)^{2\alpha/(2\alpha+1)} \leq c_2^{2\alpha/(2\alpha+1)} n^{2(1+\rho)\alpha/(2\alpha+1)} \leq c_2^{2\alpha/(2\alpha+1)} n$$

dès lors que $2\rho\alpha \leq 1$. Ainsi, notre moyenne de Fréchet régularisée \tilde{s} converge à une vitesse d'ordre $(nJ)^{-2\alpha/(2\alpha+1)}$. Grâce au théorème précédent, cela implique que l'estimateur \tilde{s} est adaptatif au sens du minimax pour la régularité α du motif moyen s^* .

1.4 Régression additive

Cette section présente des résultats publiés dans [XG-Article03].

Cadre statistique

Le modèle statistique de la régression additive a été introduit par Leontief [Leo47] et Scheffé [Sch59], nous renvoyons au chapitre 8 de [HMSW04] pour une motivation détaillée. Ce modèle met en relation des variables réelles d'entrées $X^{(1)}, \dots, X^{(k)}$ et une variable réelle de sortie Z de la façon suivante,

$$Z = \mu + \sum_{j=1}^k f_j(X^{(j)}) + \sigma\varepsilon$$

où le coefficient $\mu \in \mathbb{R}$ est inconnu et les fonctions inconnues $f_j : \mathbb{R} \rightarrow \mathbb{R}$ sont appelées des composantes. Le bruit additif ε est centré et de variance unitaire. Dans la suite de cette section, le facteur de variance σ^2 sera supposé connu par souci de simplicité mais il peut être estimé sans altérer la forme des résultats présentés. Nous renvoyons à la section 3 de [XG-Article03] pour plus de détails sur le cas à variance inconnue.

La régression linéaire est un cas particulier de régression additive pour lequel les composantes sont supposées linéaires. Considérer des composantes plus générales autorise plus de souplesse dans la modélisation des phénomènes étudiés. De plus, l'aspect sommatoire du cadre de la régression additive permet de garder l'avantage d'une analyse composante par composante comme dans le cas linéaire. Notre objectif dans la suite est l'estimation d'une composante cible sans hypothèse de régularité.

Afin de présenter notre méthode d'estimation, nous devons préciser certains aspects du cadre statistique. En limitant notre étude à des composantes définies sur $[0, 1]$, nous notons $s^* : [0, 1] \rightarrow \mathbb{R}$ la composante à estimer et $t^1, \dots, t^K : [0, 1] \rightarrow \mathbb{R}$ les $K \geq 1$ autres composantes. Étant données des mesures de probabilités ν, ν_1, \dots, ν_K sur $[0, 1]$, le support est donné par

n tirages indépendants $(x_1, y_1^1, \dots, y_1^K)^\top, \dots, (x_n, y_n^1, \dots, y_n^K)^\top \in [0, 1]^K$ de même loi $\nu \otimes \nu_1 \otimes \dots \otimes \nu_K$. Les observations s'écrivent sous la forme

$$Z_i = s^*(x_i) + \mu + \sum_{j=1}^K t^j(y_i^j) + \sigma \varepsilon_i, \quad i \in \{1, \dots, n\}. \quad (1.8)$$

Pour des raisons d'identifiabilité, les composantes doivent être centrées et nous introduisons l'espace des fonctions centrées de carré intégrable par rapport à une mesure de probabilité ν sur $[0, 1]$,

$$L_0^2([0, 1], \nu) = \left\{ f : [0, 1] \rightarrow \mathbb{R} : \int_0^1 f(t) \nu(dt) = 0 \text{ et } \int_0^1 f(t)^2 \nu(dt) < +\infty \right\}.$$

Nous supposons dans la suite que les composantes sont telles que

$$s^* \in L_0^2([0, 1], \nu) \quad \text{et} \quad t^j \in L_0^2([0, 1], \nu_j), \quad j \in \{1, \dots, K\}.$$

Les résultats présentés dans cette section sont obtenus sous deux hypothèses différentes sur le bruit additif $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^\top$:

(H_{Gau}) Le vecteur ε est Gaussien standard dans \mathbb{R}^n .

(H_{Mom}) Les variables ε_i sont i.i.d. centrées, de variance unitaire et il existe $p > 2$ tel que $\tau_p = \mathbb{E}[|\varepsilon_i|^p] < \infty$, $i \in \{1, \dots, n\}$.

L'hypothèse **(H_{Mom})** est plus faible que **(H_{Gau})**. L'avantage de considérer ces deux cas distincts est de pouvoir illustrer à quel point les résultats sont meilleurs sous l'hypothèse Gaussienne et quelle souplesse cela apporte pour la procédure de sélection de modèle.

Par souci de lisibilité, nous introduisons les vecteurs $s = (s_1, \dots, s_n)^\top$ et $t = (t_1, \dots, t_n)^\top$ comme les valeurs respectives de la composante à estimer et de la somme des autres composantes sur le support des observations,

$$s_i = s^*(x_i) \quad \text{et} \quad t_i = \mu + \sum_{j=1}^K t^j(y_i^j), \quad i \in \{1, \dots, n\}.$$

Si nous supposons qu'il existe deux espaces $E, F \subset \mathbb{R}^n$ tels que $s \in E$, $t \in F$ et $E \oplus F = \mathbb{R}^n$, alors la stratégie pour estimer le vecteur s consiste à prendre la projection P_n sur E parallèlement à F et à considérer le cadre statistique

$$Y = P_n Z = s + \sigma P_n \varepsilon \quad (1.9)$$

où $Y = (Y_1, \dots, Y_n)^\top \in E = \text{Im}(P_n)$. En pratique, de tels espaces E et F n'existent évidemment pas et il nous faudra préciser comment construire une matrice P_n qui imite cette approche. Cependant, le cadre (1.9) pour une matrice P_n connue quelconque permet de formuler les résultats de sélection de modèles nécessaires à l'estimation de la composante s^* .

Modèles et estimateurs

Dans le cadre statistique (1.9) avec une matrice P_n connue, nous pouvons supposer $s \in \text{Im}(P_n)$ sans perte de généralité. Pour la même raison, les modèles de la collection au plus dénombrable $\{S_m, m \in \mathcal{M}\}$ sont supposés être des sous-espaces de $\text{Im}(P_n)$. En considérant le contraste des moindres carrés normalisé,

$$\gamma(z) = \|Y - z\|_n^2 = \frac{1}{n} \sum_{i=1}^n (Y_i - z_i)^2, \quad z = (z_1, \dots, z_n)^\top \in \mathbb{R}^n,$$

l'estimateur par minimum de contraste dans S_m est donné par la projection orthogonale des données $\hat{s}_m = \pi_m Y$.

Contrairement au cas simple où P_n est la matrice identité, le terme de variance qui apparaît dans le risque de l'estimateur \hat{s}_m n'est plus proportionnel à la dimension du modèle,

$$\mathbb{E}[\|s - \hat{s}_m\|_n^2] = \|s - \pi_m s\|_n^2 + \frac{\text{tr}(P_n^\top \pi_m P_n)}{n} \sigma^2.$$

Pour obtenir des bornes supérieures intéressantes sur le risque de l'estimateur sélectionné $\tilde{s} = \hat{s}_{\hat{m}}$, cette non linéarité en la dimension devra être prise en compte dans la forme de la fonction de pénalité $\text{pen} : \mathcal{M} \rightarrow \mathbb{R}_+$ utilisée pour le choix de $\hat{m} \in \mathcal{M}$,

$$\hat{m} \in \underset{m \in \mathcal{M}}{\text{argmin}} \{ \gamma(\hat{s}_m) + \text{pen}(m) \}.$$

Les résultats ci-dessous montrent que la taille de la collection de modèles est plus ou moins contrainte selon que nous considérons l'hypothèse (\mathbf{H}_{Mom}) ou (\mathbf{H}_{Gau}). Pour cela, nous introduisons la notation suivante pour le nombre de modèles de dimension donnée dans notre collection,

$$\forall d \in \mathbb{N}, \quad N_d = |\{m \in \mathcal{M} \text{ tels que } \dim(S_m) = d\}|.$$

De plus, pour toute matrice A de taille $n \times n$, nous notons $\rho(A)$ la norme spectrale

$$\rho(A) = \sup_{x \in \mathbb{R}^n \setminus \{0\}} \frac{\|Ax\|_n}{\|x\|_n}.$$

Majoration du risque

Les majorations du risque de l'estimateur \tilde{s} que nous présentons sont données pour une fonction de pénalité telle que

$$\forall m \in \mathcal{M}, \quad \text{pen}(m) \geq (1 + \theta + L_m) \frac{\text{tr}(P_n^\top \pi_m P_n)}{n} \sigma^2 \quad (1.10)$$

où $\theta > 0$ est un paramètre de régularisation et, pour tout $m \in \mathcal{M}$, L_m est un poids positif dont le rôle sera précisé dans ce qui suit.

Théorème 1.6 *Nous supposons que la pénalité vérifie (1.10). Sous $(\mathbf{H}_{\text{Gau}})$, il existe des constantes $C_G, C'_G > 0$ qui ne dépendent que de θ telles que*

$$\mathbb{E}[\|s - \tilde{s}\|_n^2] \leq (1 + C_G) \inf_{m \in \mathcal{M}} \left\{ \|s - \pi_m s\|_n^2 + \text{pen}(m) \right\} + \frac{\rho^2(P_n)\sigma^2}{n} R_G$$

avec R_G donné par

$$C'_G \sum_{m \in \mathcal{M}} \exp \left(- \frac{C'_G L_m \text{tr}(P_n^\top \pi_m P_n)}{\rho^2(P_n)} \right).$$

Sous $(\mathbf{H}_{\text{Mom}})$, pour tout $q > 0$ tel que $2(q+1) < p$, il existe des constantes $C_M, C'_M > 0$ qui ne dépendent que de p, q et θ telles que

$$\mathbb{E}[\|s - \tilde{s}\|_n^{2q}]^{1/q} \leq (1 + C_M) \inf_{m \in \mathcal{M}} \left\{ \|s - \pi_m s\|_n^2 + \text{pen}(m) \right\} + \frac{\rho^2(P_n)\sigma^2}{n} R_M^{1/q}$$

avec R_M donné par

$$C'_M \tau_p \left[N_0 + \sum_{m \in \mathcal{M}: S_m \neq \{0\}} \left(1 + \frac{\text{tr}(P_n^\top \pi_m P_n)}{\rho^2(\pi_m P_n)} \right) \left(\frac{L_m \text{tr}(P_n^\top \pi_m P_n)}{\rho^2(P_n)} \right)^{q-p/2} \right].$$

La conséquence du choix de l'hypothèse sur le bruit ε apparaît principalement dans les termes de reste R_G et R_M . En effet, ces quantités sont des sommes de probabilités contrôlées par des inégalités de concentration. Le cas Gaussien $(\mathbf{H}_{\text{Gau}})$ est plus favorable car il conduit à des termes exponentiels plus facilement sommables que les termes polynômiaux sous $(\mathbf{H}_{\text{Mom}})$. Ainsi, pour pouvoir négliger le terme de reste et obtenir des inégalités oracle, la taille de la collection de modèles pourra être plus grande sous l'hypothèse Gaussienne que sous l'hypothèse de moment comme le montre le corollaire suivant.

Corollaire 1.7 *Considérons $\theta, L > 0$ et la fonction de pénalité donnée par*

$$\forall m \in \mathcal{M}, \text{pen}(m) = (1 + \theta + L) \frac{\text{tr}(P_n^\top \pi_m P_n)}{n} \sigma^2.$$

Nous supposons qu'il existe $c \in (0, 1)$ tel que

$$\forall m \in \mathcal{M}, c\rho^2(P_n) \dim(S_m) \leq \text{tr}(P_n^\top \pi_m P_n) \quad (1.11)$$

et que, selon l'hypothèse retenue, il existe $A, \omega > 0$ tels que la collection de modèles vérifie :

- Sous $(\mathbf{H}_{\text{Gau}})$,

$$\sup_{d \in \mathbb{N}: N_d > 0} \frac{\log N_d}{d} \leq A \quad \text{et} \quad L \geq \frac{2(1+\theta)^3}{c\theta^2}(A + \omega).$$

- Sous $(\mathbf{H}_{\text{Mom}})$ avec $p > 6$ et $N_0 \leq 1$,

$$\sup_{d > 0: N_d > 0} \frac{N_d}{(1+d)^{p/2-3-\omega}} \leq A \quad \text{et} \quad L \geq \omega A^{2/(p-2)}.$$

Alors, l'estimateur \tilde{s} est tel que

$$\mathbb{E}[\|s - \tilde{s}\|_n^2] \leq C \inf_{m \in \mathcal{M}} \left\{ \|s - \pi_m s\|_n^2 + \max\{L, 1\} \frac{\max\{\text{tr}(P_n^\top \pi_m P_n), c\rho^2(P_n)\}}{n} \sigma^2 \right\}$$

où $C > 1$ ne dépend que de p (sous l'hypothèse de moment), θ , ω et c .

Ce résultat correspond au cas des poids $L_m = L$ constants et donc d'une pénalité du même ordre de grandeur que le terme de variance dans le risque des estimateurs. De cette façon, la majoration du risque obtenue permet facilement de déduire une inégalité oracle pour de nombreuses matrices P_n . De plus, la condition sur N_d illustre que nous pouvons obtenir cette borne supérieure sur le risque avec des collections de modèles bien plus riches sous l'hypothèse Gaussienne que sous l'hypothèse de moment.

Si la taille de la collection de modèles dépasse cette borne, il devient nécessaire de prendre des poids L_m non constant et souvent proportionnel à $\log n$ pour contrôler les termes de reste dans le Théorème 1.6. Dans un tel cas, l'inégalité obtenue demeure presque oracle mais à un facteur logarithmique près qui dégrade les propriétés que nous pouvons en déduire.

Estimation adaptative d'une composante

Nous revenons maintenant au problème initial d'estimation non paramétrique d'une composante en régression additive à partir des observations du modèle statistique (1.8),

$$Z_i = s_i + t_i + \sigma \varepsilon_i, \quad i \in \{1, \dots, n\}.$$

Pour cela, nous considérons $D_n, D_n^{(1)}, \dots, D_n^{(K)} \in \mathbb{N}$ tels que

$$D_n + D_n^{(1)} + \dots + D_n^{(K)} < n$$

et des fonctions orthonormales $\phi_1, \dots, \phi_{D_n} \in L_0^2([0, 1], \nu)$ et $\psi_1^{(j)}, \dots, \psi_{D_n^{(j)}}^{(j)} \in L_0^2([0, 1], \nu_j)$, $j \in \{1, \dots, K\}$. Nous définissons ensuite $E \subset \mathbb{R}^n$ comme l'espace engendré par les vecteurs $(\phi_i(x_1), \dots, \phi_i(x_n))^\top$ pour $i \in \{1, \dots, D_n\}$ et,

pour tout $j \in \{1, \dots, K\}$, $F^j \subset \mathbb{R}^n$ comme l'espace engendré par les vecteurs $(\psi_i^{(j)}(y_1^j), \dots, \psi_i^{(j)}(y_n^j))^\top$ pour $i \in \{1, \dots, D_n^{(j)}\}$. En incorporant la constante, nous introduisons

$$F = \mathbb{R} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} + F^1 + \dots + F^K$$

et $G = (E + F)^\perp$.

Pour des mesures de probabilités ν, ν_1, \dots, ν_K raisonnables, nous avons presque sûrement $\dim(E) = D_n$ et $E \cap F = \{0\}$. En particulier, $E \oplus F \oplus G = \mathbb{R}^n$ et nous pouvons considérer la matrice P_n de projection sur E parallèlement à $F + G$ pour définir le vecteur

$$Y = \bar{s} + \sigma P_n \varepsilon$$

où nous avons posé

$$\bar{s} = P_n s + P_n t = s + (P_n - I_n)(s - \pi_E s) + P_n(t - \pi_{F+G} t)$$

avec respectivement π_E et π_{F+G} les projections orthogonales dans E et $F + G$. Étant donné que nous considérons des modèles $S_m \subset E$, la quantité $\|s - \pi_E s\|^2$ est majorée par tous les termes de biais $\|s - \pi_m s\|^2$. De plus, l'espace $F + G$ est construit comme un grand espace d'approximation et la quantité $\|t - \pi_{F+G} t\|^2$ pourra être négligée sous certaines hypothèses simples.

À l'aide d'algèbre linéaire et de concentration de la mesure, il est possible de montrer que la condition (1.11) est satisfaite pour notre matrice P_n avec grande probabilité. Ainsi, pour une collection de modèles $\{S_m, m \in \mathcal{M}\}$ et un bruit additif ε vérifiant $(\mathbf{H}_{\text{Gau}})$ ou $(\mathbf{H}_{\text{Mom}})$, nous pouvons appliquer le Corollaire 1.7 pour obtenir l'inégalité oracle

$$\begin{aligned} \mathbb{E}[\|s - \tilde{s}\|_n^2] &\leq C \inf_{m \in \mathcal{M}} \left\{ \|s - \pi_m s\|_n^2 + \frac{\text{tr}(P_n^\top \pi_m P_n)}{n} \sigma^2 \right\} \\ &\quad + C' \left(\|t - \pi_{F+G} t\|_n^2 + \frac{R\sigma^2}{n} \right) \end{aligned}$$

où C, C' et R sont des constantes qui ne dépendent pas de n .

Afin d'illustrer le caractère adaptatif de l'estimateur \tilde{s} , nous considérons la boule de Hölder $\mathcal{H}_\alpha(L)$ d'ordre $\alpha \in (0, 1)$ et de rayon $L > 0$ comme dans (1.3). Pour $K \leq D_n < n/2$ et $D_n^{(1)} = \dots = D_n^{(K)} = \lfloor D_n/K \rfloor$, les exemples de [BM00] offrent un large choix de fonctions $\phi_1, \dots, \phi_{D_n}$ et $\psi_1^{(j)}, \dots, \psi_{D_n^{(j)}}^{(j)}$, $j \in \{1, \dots, K\}$, avec de bonnes propriétés d'approximation pour définir la matrice de projection $P_{H,n}$ et une collection de modèles $\{S_m, m \in \mathcal{M}\}$ telle que $N_d \leq 1$ pour tout $d \in \mathbb{N}$.

Proposition 1.8 *Soit $\eta > 0$, nous considérons l'estimateur \tilde{s} construit par la procédure précédente avec la fonction de pénalité*

$$\forall m \in \mathcal{M}, \text{pen}(m) = (1 + \eta) \frac{\text{tr}(P_{H,n}^\top \pi_m P_{H,n})}{n} \sigma^2.$$

Sous les hypothèses du Corollaire 1.7 et pour

$$\alpha > \frac{1}{2} \left(\frac{\log n}{\log D_n} - 1 \right) > 0 \quad \text{et} \quad R > 0,$$

l'estimateur \tilde{s} vérifie

$$\sup_{s^*, t^1, \dots, t^K \in \mathcal{H}_\alpha(L)} \mathbb{E}_{\varepsilon, d}[\|s - \tilde{s}\|_n^2] \leq C_\alpha n^{-2\alpha/(2\alpha+1)}$$

où $\mathbb{E}_{\varepsilon, d}$ désigne l'espérance sous la loi du bruit ε et du support aléatoire et C_α est une constante qui ne dépend pas de n .

Nous renvoyons à [Sto85] pour une preuve que le risque minimax sur $\mathcal{H}_\alpha(L)$ pour l'estimation d'une composante en régression additive est de l'ordre de $n^{-2\alpha/(2\alpha+1)}$. En particulier, il est remarquable que ce risque ne soit pas dépendant du nombre $K + 1$ de composantes du modèle (1.8). Le résultat de la proposition précédente montre que l'estimateur \tilde{s} construit par notre procédure admet un risque du même ordre de grandeur et que cet estimateur est donc adaptatif au sens du minimax pour le paramètre α .

Chapitre 2

Autour de la covariance

À l'image du cadre statistique présenté dans la dernière section du chapitre précédent, je me suis intéressé à différents problèmes statistiques faisant intervenir des structures de covariance non triviales. Cet intérêt a donné lieu à des encadrements de stages de Master 2 et à des collaborations telles que [\[XG-Article07\]](#) et [\[XG-Article08\]](#).

2.1 Processus Gaussien bivarié

Cette section présente des résultats publiés dans [\[XG-Article07\]](#).

Introduction

En fin d'année 2015, Daira Velandia est venue faire un séjour scientifique à l'Institut de Mathématiques de Toulouse. C'est dans ce cadre que nos échanges au sujet de l'estimation de la covariance de processus Gaussiens ont débuté avec François Bachoc et Jean-Michel Loubes. Son intérêt se portait en particulier sur le cas des processus Gaussiens bivariés dans le cadre asymptotique par remplissage que nous présentons dans la suite de cette section.

Nous considérons un processus Gaussien bivarié centré défini sur le domaine compact $[0, 1]$,

$$Z = \left\{ Z(s) = \begin{pmatrix} Z_1(s) \\ Z_2(s) \end{pmatrix} \in \mathbb{R}^2, s \in [0, 1] \right\}.$$

Notre objectif est l'estimation de la fonction de covariance de Z ,

$$(s, s') \in [0, 1]^2 \mapsto \text{Cov}(Z_s, Z_{s'}) = \begin{pmatrix} \text{Cov}(Z_1(s), Z_1(s')) & \text{Cov}(Z_1(s), Z_2(s')) \\ \text{Cov}(Z_1(s), Z_2(s')) & \text{Cov}(Z_2(s), Z_2(s')) \end{pmatrix},$$

à partir des observations du processus Z sur n points du domaine. Les deux cadres asymptotiques largement étudiés pour aborder ce problème sont le cadre par expansion et le cadre par remplissage. Nous ne considérons pas ici le cadre par expansion qui consiste à faire grandir la taille du domaine avec le nombre d'observations n et nous renvoyons à [BVV15] pour une étude de la covariance dans ce cas. Dans le cadre par remplissage qui nous intéresse, le processus Z est observé sur un nombre n croissant de points distincts supposés denses dans $[0, 1]$ quand n tend vers l'infini. Afin de formaliser le cadre par remplissage, nous pouvons considérer une suite $(s'_k)_{k \geq 1}$ dense dans $[0, 1]$ et, pour tout $n \geq 1$, prendre $s_1 = s'_{(1)}, \dots, s_n = s'_{(n)}$ comme la version ordonnée des points s'_1, \dots, s'_n ,

$$0 \leq s'_{(1)} < \dots < s'_{(n)} \leq 1.$$

Un modèle de covariance est généralement choisi pour représenter la fonction de covariance et le problème d'estimation devient paramétrique. Dans ce travail, nous considérerons le modèle exponentiel pour lequel la covariance s'écrit

$$\text{Cov}_\psi(Z_s, Z_{s'}) = \begin{pmatrix} \sigma_1^2 e^{-\theta|s-s'|} & \rho\sigma_1\sigma_2 e^{-\theta|s-s'|} \\ \rho\sigma_1\sigma_2 e^{-\theta|s-s'|} & \sigma_2^2 e^{-\theta|s-s'|} \end{pmatrix} \quad (2.1)$$

où $\psi = (\theta, \sigma_1^2, \sigma_2^2, \rho)^\top \in \Psi = (0, +\infty)^3 \times (-1, 1)$ et nous notons P_ψ la mesure Gaussienne centrée associée. Nos résultats étendent au cas bivarié ceux obtenus par [Yin91] dans le cas univarié où ce modèle de covariance $\text{Cov}_\psi(Z_1(s), Z_1(s')) = \sigma^2 e^{-\theta|s-s'|}$ correspond au processus de Ornstein-Uhlenbeck.

Avec le modèle de covariance (2.1), les observations se mettent sous la forme d'un vecteur $Z_n = (Z_1(s_1), \dots, Z_1(s_n), Z_2(s_1), \dots, Z_2(s_n))^\top$ de loi Gaussienne centrée et de matrice de covariance Σ_ψ qui s'écrit comme le produit de Kronecker $A \otimes R$ avec

$$A = \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho \\ \sigma_1\sigma_2\rho & \sigma_2^2 \end{pmatrix} \quad \text{et} \quad R = \left(e^{-\theta|s_i-s_j|} \right)_{i,j \in \{1, \dots, n\}}.$$

Cette forme produit est dite séparable et cela motive notre étude dans le cas bivariée car le nombre de paramètres du modèle est raisonnable. De plus, le caractère explicite de la fonction de vraisemblance

$$f_n(\psi) = (2\pi)^{-n/2} \det(\Sigma_\psi)^{-1} \exp\left(-\frac{1}{2} Z_n^\top \Sigma_\psi^{-1} Z_n\right)$$

permet de considérer certains estimateurs par maximum de vraisemblance.

Équivalence de mesures Gaussiennes

Les paramètres d'un modèle de covariance se distinguent en deux catégories (voir [IR78] et [Ste99]) :

- un paramètre est dit microergodique si, pour deux valeurs distinctes de ce paramètre, les mesures Gaussiennes sont orthogonales,
- un paramètre est dit non-microergodique si, même pour deux valeurs distinctes de ce paramètre, les mesures Gaussiennes sont équivalentes.

Une conséquence importante est que seuls les paramètres microergodiques peuvent être estimés de façon consistante.

Étant donné $\psi_1, \psi_2 \in \Psi$, une quantité pertinente pour étudier l'équivalence ou l'orthogonalité de P_{ψ_1} et P_{ψ_2} à partir des n observations est la divergence de Kullback-Leibler symétrisée,

$$I_n(P_{\psi_1}, P_{\psi_2}) = \mathbb{E}_{\psi_1} \left[\log \frac{f_n(\psi_1)}{f_n(\psi_2)} \right] + \mathbb{E}_{\psi_2} \left[\log \frac{f_n(\psi_2)}{f_n(\psi_1)} \right]$$

où \mathbb{E}_{ψ_k} désigne l'espérance sous Z_n de loi $\mathcal{N}(0, \Sigma_{\psi_k})$, $k \in \{1, 2\}$. Avec la construction des points d'observation s_1, \dots, s_n présentée précédemment, de nouveaux points sont ajoutés au support quand n augmente mais aucun n'est supprimé. Cette remarque assure que $I_n(P_{\psi_1}, P_{\psi_2})$ est croissante et nous pouvons définir la limite potentiellement infinie

$$I(P_{\psi_1}, P_{\psi_2}) = \lim_{n \rightarrow \infty} I_n(P_{\psi_1}, P_{\psi_2}).$$

Par [IR78], nous savons que P_{ψ_1} et P_{ψ_2} sont équivalentes si et seulement si $I(P_{\psi_1}, P_{\psi_2}) < \infty$ et cette propriété permet de prouver le lemme suivant.

Lemme 2.1 *Pour tout $\psi_1 = (\theta_1, \sigma_{1,1}^2, \sigma_{1,2}^2, \rho_1)$, $\psi_2 = (\theta_2, \sigma_{2,1}^2, \sigma_{2,2}^2, \rho_2) \in \Psi$, les mesures Gaussiennes P_{ψ_1} et P_{ψ_2} sont équivalentes sur la tribu engendrée par Z si et seulement si*

$$\sigma_{1,1}^2 \theta_1 = \sigma_{1,2}^2 \theta_2 \quad , \quad \sigma_{2,1}^2 \theta_1 = \sigma_{2,2}^2 \theta_2 \quad \text{et} \quad \rho_1 = \rho_2.$$

La conséquence de ce lemme est qu'il n'est pas possible d'estimer de façon consistante tous les paramètres du modèle de covariance séparable exponentiel (2.1) à partir de données observées sur $[0, 1]$. Cependant, les paramètres microergodiques $\sigma_1^2 \theta$, $\sigma_2^2 \theta$ et ρ peuvent être estimés de façon consistante.

Estimateurs du maximum de vraisemblance

Étant donné un pavé borné non vide $\Psi_0 = (a_\theta, b_\theta) \times (a_{\sigma_1^2}, b_{\sigma_1^2}) \times (a_{\sigma_2^2}, b_{\sigma_2^2}) \times (a_\rho, b_\rho) \subset \Psi$, l'estimateur du maximum de vraisemblance $\hat{\psi}_n = (\hat{\theta}_n, \hat{\sigma}_{n,1}^2, \hat{\sigma}_{n,2}^2, \hat{\rho}_n)$

est défini par

$$\hat{\psi}_n = \operatorname{argmax}_{\psi \in \Psi_0} f_n(\psi).$$

Le résultat suivant donne la consistance forte des estimateurs des paramètres microergodiques.

Théorème 2.2 *Soit $\psi = (\theta, \sigma_1^2, \sigma_2^2, \rho) \in \Psi_0$, sous P_ψ , l'estimateur $\hat{\psi}_n$ est bien défini avec probabilité un pour n assez grand et nous avons*

$$\hat{\theta}_n \hat{\sigma}_{n,1}^2 \xrightarrow[n \rightarrow \infty]{p.s.} \theta \sigma_1^2, \quad \hat{\theta}_n \hat{\sigma}_{n,2}^2 \xrightarrow[n \rightarrow \infty]{p.s.} \theta \sigma_2^2 \quad \text{et} \quad \hat{\rho}_n \xrightarrow[n \rightarrow \infty]{p.s.} \rho.$$

La normalité asymptotique a aussi été obtenue pour les estimateurs des paramètres microergodiques sous la forme du théorème suivant.

Théorème 2.3 *Si ψ appartient à un compact d'intérieur non vide inclus dans Ψ_0 , alors*

$$\sqrt{n} \begin{pmatrix} \hat{\theta}_n \hat{\sigma}_{n,1}^2 - \theta \sigma_1^2 \\ \hat{\theta}_n \hat{\sigma}_{n,2}^2 - \theta \sigma_2^2 \\ \hat{\rho} - \rho \end{pmatrix} \xrightarrow[n \rightarrow \infty]{\mathcal{L}} \mathcal{N}(0, \Sigma)$$

où la matrice de covariance vaut

$$\Sigma = \begin{pmatrix} 2\theta^2 \sigma_1^4 & 2\theta^2 \rho^2 \sigma_1^2 \sigma_2^2 & \theta \rho \sigma_1^2 (1 - \rho^2) \\ 2\theta^2 \rho^2 \sigma_1^2 \sigma_2^2 & 2\theta^2 \sigma_2^4 & \theta \rho \sigma_2^2 (1 - \rho^2) \\ \theta \rho \sigma_1^2 (1 - \rho^2) & \theta \rho \sigma_2^2 (1 - \rho^2) & (1 - \rho^2)^2 \end{pmatrix}.$$

Il est en particulier intéressant de remarquer que la matrice de covariance asymptotique ne dépend pas de la construction de la suite dense des points du support.

2.2 Décomposition de Cholesky régularisée

Cette section présente des résultats publiés dans [\[XG-Article08\]](#).

Introduction

Lorsque j'ai rejoint l'ISAE SUPAERO en 2019, j'ai fait la rencontre de Olivier Besson et François Vincent qui travaillent tous deux sur des thématiques liées au traitement du signal. En particulier, leur intérêt pour les données multicanal les amène à considérer des problèmes d'estimation de matrices de covariance pour des mesures de risque spécifiques. Nous avons

ainsi collaboré pour étudier une régularisation de la décomposition de Cholesky de la matrice de covariance empirique afin de minimiser la métrique log-Cholesky récemment introduite par [Lin19].

Le cadre statistique que nous considérons est donné par des vecteurs Gaussiens $X_1, \dots, X_n \in \mathbb{R}^p$ indépendants et de même loi $\mathcal{N}(0, \Sigma)$ avec $n \geq p$ et une matrice symétrique positive et inversible Σ de taille $p \times p$. L'estimateur du maximum de vraisemblance de la matrice Σ est donné par S/n avec

$$S = \sum_{i=1}^n X_i X_i^\top.$$

Par construction, S suit la loi de Wishart $\mathcal{W}(n, \Sigma)$ à n degrés de liberté et de matrice Σ . Lorsque le nombre d'observations n n'est pas significativement plus grand que le nombre de variables p , il est connu que S/n est moins bien conditionnée que Σ . En effet, du point de vue spectral, les grandes valeurs propres de Σ ont tendance à être surestimées et les petites valeurs propres à être sous-estimées. Ainsi, il est naturel de considérer des techniques de régularisation de S pour bien estimer Σ .

Ce sujet a fait l'objet d'un très grand nombre de travaux et nous renvoyons à [MJS12] et [Tsu16] ainsi qu'aux références qu'ils citent pour une présentation plus complète. L'approche de Stein est historiquement importante. Pour un estimateur $\hat{\Sigma}$, cette approche considère la fonction de perte proposée dans [JS61],

$$\mathcal{L}_1(\Sigma, \hat{\Sigma}) = \text{tr}(\hat{\Sigma}\Sigma^{-1}) - \log \det(\hat{\Sigma}\Sigma^{-1}) - p$$

et cherche à minimiser le risque associé $\mathbb{E}[\mathcal{L}_1(\Sigma, \hat{\Sigma})]$.

Estimateurs régularisés

Une classe usuelle d'estimateurs obtenus par régularisation de S est construite à partir de la décomposition spectrale $S = U \text{diag}(\lambda) U^\top$ où U est une matrice orthogonale et $\text{diag}(\lambda)$ désigne la matrice diagonale donnée par le vecteur des valeurs propres $\lambda = (\lambda_1, \dots, \lambda_p)^\top \in \mathbb{R}_+^p$. Les estimateurs considérés sont alors de la forme $\hat{\Sigma} = U \text{diag}(\varphi(\lambda)) U^\top$ pour certaines fonctions non linéaires $\varphi : \mathbb{R}^p \rightarrow \mathbb{R}^p$ de λ (voir [She95]). Les inconvénients de cette procédure sont que certaines valeurs propres de $\hat{\Sigma}$ peuvent être négatives et que l'ordre des valeurs propres n'est pas préservé. Des techniques pour corriger ces problèmes existent mais elles dépassent le cadre de cette introduction.

Dans la suite, nous nous intéressons à une autre classe d'estimateurs construits par régularisation de la décomposition de Cholesky de S ,

$$S = G_S G_S^\top$$

où G_S est une matrice triangulaire inférieure dont les termes diagonaux sont positifs. Une telle matrice G_S est unique et s'appelle le facteur de Cholesky de S . La décomposition de Cholesky est bien adaptée aux lois de Wishart car si $W \sim \mathcal{W}(n, I_p)$, alors le facteur de Cholesky $G_W = (G_{W,ij})_{1 \leq i, j \leq p}$ admet des entrées indépendantes de lois données par

$$\forall p \geq i > j \geq 1, G_{W,ij} \sim \mathcal{N}(0, 1) \text{ et } G_{W,jj}^2 \sim \chi^2(n - j + 1).$$

De plus, il est utile de rappeler que si $S \sim \mathcal{W}(n, \Sigma)$ alors la matrice $\Sigma^{-1/2} S \Sigma^{-1/2}$ suit la loi $\mathcal{W}(n, I_p)$ (voir [GN00]). Ainsi, le facteur de Cholesky de $\Sigma^{-1/2} S \Sigma^{-1/2}$ suit la loi de G_W et ne dépend pas de la matrice Σ .

Pour un vecteur $d = (d_1, \dots, d_p)^\top \in \mathbb{R}^p$, nous considérons l'estimateur régularisé $\hat{\Sigma}_d$ défini par

$$\hat{\Sigma}_d = G_S \text{diag}(d) G_S^\top. \quad (2.2)$$

Pour la perte $\mathcal{L}_1(\Sigma, \hat{\Sigma}_d)$, la qualité de cet estimateur est mesurée par le risque $\mathcal{R}_1(\Sigma, d) = \mathbb{E}[\mathcal{L}_1(\Sigma, \hat{\Sigma}_d)]$. Grâce à la remarque précédente, un simple calcul permet de retrouver le résultat de Stein : le minimiseur de $\mathcal{R}_1(\Sigma, d)$ est donné par $d_1 = (d_{1,1}, \dots, d_{1,p})'$ avec

$$d_{1,j} = (n + p - 2j + 1)^{-1}.$$

Le fait que le minimiseur du risque $\mathcal{R}_1(\Sigma, d)$ ne dépende pas de Σ est une propriété importante des estimateurs contruits par régularisation de la décomposition de Cholesky. Cela demeure vrai pour des fonctions de perte autres que celle de Stein, par exemple $\mathcal{L}_2(\Sigma, \hat{\Sigma}_d) = \text{tr}((\hat{\Sigma}_d \Sigma^{-1} - I_p)^2)$. Plus généralement, l'utilisation du facteur de Cholesky est intéressante en pratique car il s'agit d'un objet simple à calculer et qui a de nombreuses utilisations en traitement du signal.

Au-delà de la construction d'estimateurs, le facteur de Cholesky est aussi un objet naturel à considérer pour construire des fonctions de pertes telles que celles proposées par [EO87],

$$\begin{aligned} \mathcal{L}_3(\Sigma, \hat{\Sigma}) &= \text{tr} \left((G_\Sigma^{-1} G_{\hat{\Sigma}} - I_p)(G_\Sigma^{-1} G_{\hat{\Sigma}} - I_p)^\top \right), \\ \mathcal{L}_4(\Sigma, \hat{\Sigma}) &= \text{tr} \left((G_{\hat{\Sigma}}^{-1} G_\Sigma - I_p)(G_{\hat{\Sigma}}^{-1} G_\Sigma - I_p)^\top \right). \end{aligned}$$

Pour des estimateurs régularisés comme (2.2), les mêmes arguments permettent de faire le lien avec le facteur de Cholesky de $W \sim \mathcal{W}(n, I_p)$,

$$G_{\hat{\Sigma}_d} = G_\Sigma G_W \text{diag}(d)^{1/2}.$$

Ainsi, les minimiseurs des risques associés $\mathcal{R}_3(\Sigma, d)$ et $\mathcal{R}_4(\Sigma, d)$ sont encore indépendants de la matrice Σ .

Risque log-Cholesky

La métrique log-Cholesky introduite par [Lin19] est définie sur l'ensemble des matrices triangulaires inférieures de diagonale positive. Elle permet de considérer la perte suivante construite à partir des facteurs de Cholesky comme \mathcal{L}_3 et \mathcal{L}_4 ,

$$\begin{aligned} \mathcal{L}_5(\Sigma, \hat{\Sigma}) &= \|G_{\hat{\Sigma}} - G_{\Sigma}\|_F^2 - \|\text{diag}(G_{\hat{\Sigma}}) - \text{diag}(G_{\Sigma})\|^2 \\ &\quad + \|\log \text{diag}(G_{\hat{\Sigma}}) - \log \text{diag}(G_{\Sigma})\|^2 \end{aligned}$$

où $\|\cdot\|_F$ désigne la norme de Frobenius et $\text{diag}(A)$ est le vecteur formé par les éléments diagonaux de la matrice A .

Comme précédemment, des arguments relatifs à la loi du facteur de Cholesky d'une matrice de Wishart permettent de déduire un expression du risque associé pour des estimateurs de la forme (2.2),

$$\begin{aligned} \mathcal{R}_5(\Sigma, d) &= \mathbb{E} \left[\mathcal{L}_5(\Sigma, \hat{\Sigma}) \right] \\ &= \sum_{j=1}^p a_j(\Sigma) d_j - 2 \sum_{j=1}^p b_j(\Sigma) \sqrt{d_j} + \frac{1}{4} \sum_{j=1}^p \log^2(d_j) + \sum_{j=1}^p c_j \log(d_j) \\ &\quad + \text{tr}(\Sigma) - \|\text{diag}(G_{\Sigma})\|^2 + \mathbb{E} \left[\|\log \text{diag}(G_W)\|^2 \right] \end{aligned}$$

où nous avons posé, pour tout $j \in \{1, \dots, p\}$,

$$\begin{aligned} a_j(\Sigma) &= \mathbb{E} \left[\chi_{n-j+1}^2 \right] \left((G_{\Sigma}^{\top} G_{\Sigma})_{jj} - G_{\Sigma, jj}^2 \right) + \sum_{i=j+1}^p (G_{\Sigma}^{\top} G_{\Sigma})_{ii}, \\ b_j(\Sigma) &= \mathbb{E} \left[\sqrt{\chi_{n-j+1}^2} \right] \left((G_{\Sigma}^{\top} G_{\Sigma})_{jj} - G_{\Sigma, jj}^2 \right), \\ c_j &= \mathbb{E} \left[\log \sqrt{\chi_{n-j+1}^2} \right]. \end{aligned}$$

Nous avons montré dans [XG-Article08] que le risque $\mathcal{R}_5(\Sigma, d)$ admet un unique minimum atteint pour un vecteur $d_5 = d_5(\Sigma)$ qui dépend de Σ contrairement aux cadres présentés précédemment. Puisque Σ est inconnue en pratique, il n'est pas possible d'utiliser le paramètre de régularisation optimal $d_5(\Sigma)$ d'un point de vue statistique. De plus, aucun estimateur sans biais de $\mathcal{R}_5(\Sigma, d)$ ne semble disponible et des approches alternatives doivent être considérées.

Une première proposition naïve consiste à remarquer que, puisque les paramètres optimaux d_1, \dots, d_4 ne dépendent pas de Σ , ils peuvent être calculés dans le cas $\Sigma = I_p$. En imitant cette remarque, le vecteur $d_5(I_p)$ peut être proposé comme paramètre de régularisation. Cette solution peut être sous-optimale mais elle présente l'avantage de la simplicité en pratique.

Afin de remédier au problème de perte de performance de l'estimateur calculé avec $d_5(I_p)$ plutôt que $d_5(\Sigma)$, nous proposons d'exploiter le fait que le paramètre de régularisation d_5 soit défini comme le minimiseur d'une fonction dont la dépendance en Σ est explicite. Ainsi, à partir d'un premier estimateur $\hat{\Sigma}$, il est possible de considérer le minimiseur de $d \mapsto \mathcal{R}_5(\hat{\Sigma}, d)$. Par exemple, $\hat{\Sigma} = n^{-1}S$ peut être utilisé pour obtenir

$$\hat{d}_5 = \operatorname{argmin}_{d \in \mathbb{R}_+^p} \mathcal{R}_5(\hat{\Sigma}, d)$$

et définir l'estimateur régularisé $G_S \operatorname{diag}(\hat{d}_5) G_S^\top$.

La procédure précédente peut également être itérée pour rechercher une certaine stabilité du résultat. Considérons une première décomposition de Cholesky régularisée $\hat{\Sigma}^{(0)} = G_S \operatorname{diag}(\hat{d}_5^{(0)}) G_S^\top$ pour un certain paramètre initial $\hat{d}_5^{(0)}$, par exemple $d_5(I_p)$ ou le vecteur constant $(n^{-1/2}, \dots, n^{-1/2})^\top$. Pour tout $n \geq 1$, nous posons

$$\hat{d}_5^{(n)} = \operatorname{argmin}_{d \in \mathbb{R}_+^p} \mathcal{R}_5(\hat{\Sigma}^{(n-1)}, d)$$

et $\hat{\Sigma}^{(n-1)} = G_S \operatorname{diag}(\hat{d}_5^{(n-1)}) G_S^\top$. Nos études numériques illustrent que cette procédure converge rapidement en pratique et donne de bonnes performances sur des données simulées.

Chapitre 3

Applications aéronautiques

Entre 2018 et 2021, j'ai codirigé les thèses CIFRE Airbus de Ambre Diet [Die21] avec Nicolas Couellan et de Fériel Boulfani [Bou21] avec Anne Ruiz-Gazen. Dans le cadre de ces travaux, nous avons développé plusieurs applications mathématiques au domaine industriel aéronautique dont certaines ont fait l'objet de publications et de communications dans des conférences internationales. Ce chapitre présente certaines problématiques abordées et les résultats publiés.

3.1 Tolérancement statistique sous contraintes industrielles

Cette section présente des résultats publiés dans [XG-Article09] et [XG-Conf3].

Contexte industriel

Le tolérancement gère les incertitudes liées à la géométrie et aux dimensions des différentes pièces à assembler lors des étapes de la fabrication d'un produit. Le besoin de gestion des tolérances peut être lié à des exigences de production, de performance ou d'esthétique du produit. Le cadre qui nous intéresse ici est celui de la fabrication d'un avion. Dans ce contexte aéronautique, les problèmes d'incertitude sont reliés à des questions de sécurité, de qualité et de performance de l'avion mais aussi à des contraintes industrielles (technologie, outillage, ...).

La notion d'assemblage se définit comme un ensemble de pièces ou de sous-assemblages combinés pour remplir une fonction donnée. Chaque élément d'un assemblage dispose de caractéristiques dimensionnelles (hauteur, largeur, inclinaison, ...). Il convient de distinguer les caractéristiques finales

de l'assemblage dont dépendent les performances de l'avion et les caractéristiques des éléments impliqués dans l'assemblage. Les premières jouent le rôle de sorties dans un contexte statistique tandis que les autres sont des entrées du problème.

Le lien entre les entrées et les sorties d'un assemblage se représente par un modèle de tolérance. Ce concept permet de mettre en relation les contributions des caractéristiques des éléments constitutifs et les exigences de haut niveau sur l'assemblage. Pour ce faire, nous considérons une chaîne de côte associée à chaque exigence comme une liste d'éléments et leurs limites de tolérance. Chaque assemblage, dans la fabrication d'un produit, est relié à une chaîne de côte théorique qui sert de base à la démarche du tolérancement.

Dans un contexte industriel, un produit doit vérifier des spécifications de dimension précises. La dimension théorique d'une caractéristique est appelée la valeur nominale. Bien entendu, les dimensions d'un produit ne peuvent généralement pas être exactement égales aux valeurs nominales en pratique. Les sources de variabilité sont multiples : précision des outils utilisés, environnement, conditions de transport, ... Dans la suite, nous ne considérons qu'une variabilité globale qui mêle les effets extérieurs et la précision de la mesure physique des caractéristiques. Pour la chaîne de côte, cette variabilité se représente comme un intervalle de tolérance autour d'une valeur nominale.

Le fait qu'un assemblage puisse être constitué de sous-assemblages induit un effet de cascade dans le modèle de tolérance. En effet, les variations d'une caractéristique d'une entrée d'un assemblage influent sur la sortie qui devient une caractéristique d'entrée pour l'assemblage suivant et ainsi de suite. Depuis les pièces élémentaires jusqu'à l'avion complet, chaque écart par rapport à la valeur nominale doit être pris en compte pour analyser son impact et définir un intervalle de tolérance raisonnable pour la faisabilité de l'assemblage. La cascade des tolérances apparaît ainsi comme un problème fonctionnel complexe motivé par un besoin industriel exigeant.

Une première approche consiste à spécifier en premier les exigences de haut niveau de l'assemblage. La méthodologie du tolérancement permet ensuite d'analyser l'incertitude sur les dimensions des éléments impliqués dans l'assemblage. Cette démarche est dite descendante car elle propage les exigences dans la cascade des tolérances du haut niveau vers les pièces élémentaires. En particulier, l'approche descendante permet de décider quelles doivent être les tolérances sur les entrées initiales du modèle de tolérance pour que les exigences de haut niveau soient respectées.

Une autre façon de procéder, dite ascendante, suppose que les intervalles de tolérance pour les dimensions des pièces élémentaires sont connus. L'objectif est alors de décrire l'incertitude sur les exigences de haut niveau. Autrement dit, il s'agit de prédire l'intervalle dans lequel évolue une dimension

liée à l'exigence de haut niveau d'un assemblage à partir des tolérances sur les dimensions des éléments constitutifs. Dans la suite de cette section, nous présentons des travaux qui s'inscrivent dans cette approche ascendante.

La chaîne de côte est l'objet central des méthodologies de tolérancement. Pour chaque caractéristique contribuant à l'assemblage, un intervalle de tolérance est donné et il correspond à la variabilité dimensionnelle de la caractéristique par rapport à sa valeur nominale. L'approche statistique du tolérancement correspond à modéliser cette variabilité par une loi de probabilité. Ainsi, pour une chaîne de côte donnée, une des problématiques du tolérancement statistique est de déterminer la variation de sortie d'une exigence de haut niveau en connaissant la loi des caractéristiques d'entrée de l'assemblage.

Modèle de tolérance en phase de design

Dans le contexte de la fabrication des avions, il n'est pas envisageable de faire des pré-séries. L'information disponible sur la distribution statistique des caractéristiques d'entrée est donc limitée en phase de design. La suite de cette section concerne cette phase de design et l'approche ascendante. Pour un modèle de tolérance linéaire, nous présentons le problème du tolérancement comme une question de concentration de la mesure pour des sommes de variables indépendantes. Les outils théoriques utilisés sont similaires aux inégalités de concentration utiles pour prouver les résultats énoncés dans le chapitre 1.

La mise en place d'une chaîne de côte est un compromis entre des exigences de dimension et la faisabilité du procédé industriel. Si les tolérances sont trop étroites, le processus de production devient plus difficile à mettre en œuvre à cause du coût nécessaire pour satisfaire les tolérances ou du grand nombre d'éléments mis au rebut pour non respect des tolérances. D'un autre côté, des tolérances trop grandes augmentent le risque de non conformité des exigences de haut niveau ou celui de la baisse des performances.

Dans ce qui suit, nous ne considérons que des intervalles de tolérance centrés autour d'une valeur nominale pour les entrées et la sortie du modèle. D'autres façons de procéder existent et nous renvoyons à [HOS+19] pour une présentation plus générale. L'approche traditionnelle du tolérancement statistique consiste à supposer que les caractéristiques d'entrée sont des variables indépendantes Gaussiennes centrées autour de leurs valeurs nominales. Cette hypothèse permet généralement de construire des intervalles de tolérance resserrés mais ne couvre pas le cas où les entrées ne sont pas raisonnablement distribuées autour des valeurs nominales. En phase de design et sans information additionnelle, une telle hypothèse de normalité des

entrées ne peut pas être vérifiée et nous souhaitons construire des intervalles de tolérance plus robustes à la loi des entrées.

Nous nous plaçons dans le cadre statistique donné par n variables d'entrées indépendantes $Z_1, \dots, Z_n \in \mathbb{R}$, des valeurs nominales $z_1, \dots, z_n \in \mathbb{R}$ et une sortie $Y \in \mathbb{R}$. Le modèle de tolérance doit être spécifié pour représenter la relation entre les Z_i et Y . Dans un contexte raisonnable en aéronautique où la variabilité des entrées est petite par rapport à l'échelle de l'assemblage, le modèle linéaire est une approche largement utilisée pour le tolérancement (voir [LH97]). Pour des coefficients $\alpha_1, \dots, \alpha_n \in \mathbb{R}$, cela conduit à considérer le modèle

$$Y = \sum_{i=1}^n \alpha_i (Z_i - z_i).$$

Il est important de noter que les valeurs nominales z_i et les coefficients α_i de ce modèle sont des valeurs fixées pour un assemblage et ne sont pas des paramètres à estimer. Ainsi, il est équivalent de considérer la variabilité de Z_i autour de z_i ou celle de $X_i = \alpha_i (Z_i - z_i)$ autour de zéro. L'écriture du modèle de tolérance linéaire se simplifie,

$$Y = \sum_{i=1}^n X_i$$

où les variables réelles X_1, \dots, X_n sont supposées centrées et indépendantes. La chaîne de côte correspond alors à des bornes de tolérance $v_1, \dots, v_n > 0$ imposant à chaque X_i d'être dans l'intervalle $[-v_i, v_i]$ pour être considéré comme conforme.

Étant donné un niveau de confiance $\rho \in (0, 1)$, l'objectif est de déterminer un intervalle de tolérance $[-t_\rho, t_\rho]$ correspondant à une exigence de haut niveau que la sortie Y doit satisfaire avec probabilité supérieure à $1 - \rho$,

$$\mathbb{P}(|Y| \geq t_\rho) \leq \rho.$$

Dans un cadre Gaussien où $N_i \sim \mathcal{N}(0, v_i^2/3)$, $i \in \{1, \dots, n\}$, l'inégalité de concentration Gaussienne (voir [BLM13]) donne

$$\mathbb{P}\left(\left|\sum_{i=1}^n N_i\right| \geq \ell_\rho \tau\right) \leq \rho \tag{3.1}$$

où nous avons posé

$$\ell_\rho = \frac{1}{3} \sqrt{2 \log\left(\frac{2}{\rho}\right)} \quad \text{et} \quad \tau = \sqrt{\sum_{i=1}^n v_i^2}.$$

Comme discuté précédemment, l'hypothèse de normalité n'est pas acceptable en phase de design et nous ne pouvons considérer que les bornes de tolérance v_1, \dots, v_n . Pour se prémunir de tout biais de production dans la dimension des pièces élémentaires, la loi uniforme sur $[-v_i, v_i]$ apparaît comme la loi la moins informative pour la variable d'entrée X_i . Le modèle de tolérance linéaire revient alors à considérer des sommes de variables uniformes indépendantes. Cela correspond au cadre étudié par [KVC01] où les auteurs calculent explicitement la loi de Y mais dont le résultat est numériquement inutilisable et n'est donc pas adapté à notre problème de tolérancement en pratique.

Avec $X_i \sim \mathcal{U}([-v_i, v_i])$, l'écart-type de Y vaut $\tau/\sqrt{3}$ comme dans le cas Gaussien mais l'inégalité (3.1) n'est plus valable. Les variables d'entrée étant bornées, il est possible d'utiliser l'inégalité de Hoeffding (voir [BLM13]) pour obtenir

$$\mathbb{P} \left(\left| \sum_{i=1}^n X_i \right| \geq 3\ell_\rho \tau \right) \leq \rho.$$

Au facteur 3 près, nous retrouvons la borne du cas Gaussien qui correspond, pour n assez grand, au cas où les X_i ont la même variance, *i.e.* $v_1 = \dots = v_n$. Lorsque les valeurs v_i ne sont plus équilibrées, cette borne de Hoeffding est pessimiste comme le montre la figure 3.1. Lorsque un des v_i domine les autres, l'intervalle de tolérance sur Y se reserre autour de la valeur nominale. Il est donc pertinent d'introduire une métrique de la dispersion des v_i afin d'en tirer parti dans la borne de tolérance de la sortie lorsque les contributions d'entrée sont déséquilibrées.

L'inégalité de Hoeffding n'utilise que les bornes des entrées et pas le fait qu'elles suivent des lois uniformes. Pour intégrer cette information et faire apparaître la dépendance en la dispersion des v_i , nous utilisons la méthode de Cramér-Chernoff (voir [BLM13]) qui conduit à

$$\forall t > 0, \mathbb{P} \left(\left| \sum_{i=1}^n X_i \right| \geq t \right) \leq 2 \exp \left(\inf_{\lambda > 0} \phi(\lambda, t) \right)$$

où la fonction ϕ est donnée par

$$\phi(\lambda, t) = S_\lambda(v) + n \log \left(\frac{1 - e^{-2\lambda\bar{v}}}{2\lambda\bar{v}} \right) + n\lambda\bar{v} - \lambda t.$$

avec

$$\bar{v} = \frac{1}{n} \sum_{i=1}^n v_i \quad \text{et} \quad S_\lambda(v) = \sum_{i=1}^n \left(\log \left(\frac{1 - e^{-2\lambda v_i}}{2\lambda v_i} \right) - \log \left(\frac{1 - e^{-2\lambda\bar{v}}}{2\lambda\bar{v}} \right) \right)$$

À t fixé, la minimisation de $\lambda \mapsto \phi(\lambda, t)$ dépend de l'équilibre entre les valeurs v_i mesuré par la quantité $S_\lambda(v)$. Des bornes supérieures simples sur $S_\lambda(v)$ permettent de se ramener à des mesures de dispersion plus classiques,

$$S_\lambda(v) \leq \lambda \sum_{i=1}^n |v_i - \bar{v}| \quad \text{ou} \quad S_\lambda(v) \leq \frac{\lambda^2}{2} \sum_{i=1}^n (v_i - \bar{v})^2.$$

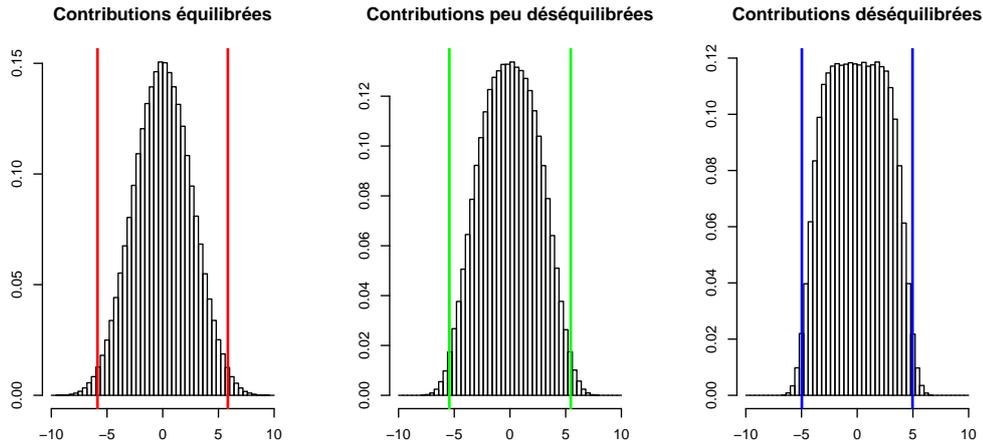


FIGURE 3.1 – Distribution de 10^6 observations de la somme de 5 variables uniformes avec $\tau^2 = 20$ pour différents équilibres des v_i et quantiles à 1% et 99%.

Cette méthode de calcul des tolérances en phase de design a été validée sur des données simulées et des données réelles issues de chaînes de côte de Airbus. Cette application et la métrique $S_\lambda(v)$ ont permis aux équipes de tolérancement de mieux comprendre le rôle de certains paramètres de forme utilisés en pratique. De plus, l'avantage des inégalités présentées est leur nature non asymptotique qui les rend directement utilisables en pratique et peu coûteuses à calculer contrairement à des méthodes de type Monte Carlo souvent considérées dans la littérature du tolérancement statistique.

3.2 Comportement des systèmes électriques aéronautiques

Cette section présente des résultats publiés dans [XG-Article11], [XG-Conf2] et [XG-Conf4].

Dimensionnement de générateurs électrique d'avions

Les générateurs électriques embarqués à bord des avions sont dimensionnés en fonction des charges électriques installées dans l'appareil. La méthode usuelle consiste à faire la somme de toutes les charges possibles, ce qui a tendance à surestimer la capacité nécessaire des générateurs. Pour dimensionner de futurs générateurs, nous proposons une méthode d'estimation du rapport entre la consommation électrique observée à bord d'un avion et la valeur donnée par le fabriquant.

Lorsqu'un avion est livré à une compagnie aérienne, il est accompagné d'un ELA (*Electrical Load Analysis*) qui décrit son réseau électrique et la charge totale par générateur embarqué. Ce document sert de référence à la compagnie, par exemple, en cas d'installation de nouveaux appareils électriques dans le réseau de l'avion pour éviter toute surcharge. Dans ce contexte, il est naturel de surestimer les générateurs pour parer à des situations critiques. Cependant, une surestimation induit un coût plus élevé, un encombrement plus important et une plus grande masse à transporter.

Dans ce qui suit, nous utilisons la théorie des valeurs extrêmes pour quantifier la surestimation de la taille d'un générateur électrique. Cette théorie statistique permet d'estimer des quantiles extrêmes et potentiellement des points terminaux mais également de donner des intervalles de confiance sur ces valeurs. La nature de ces résultats est asymptotique au sens où le nombre d'observations est supposé tendre vers l'infini.

Étant donné que chaque avion possède son propre ELA, il n'est pas pertinent d'estimer la charge maximale par appareil. L'étude que nous proposons estime le rapport entre la charge observée et la charge maximale prédite par l'ELA. De cette façon, la valeur de ce rapport peut être utilisée pour corriger le dimensionnement de futurs générateurs indépendamment du type de navigation de l'appareil.

Les données dont nous disposons sont des séries temporelles de valeurs de charge de générateurs de courant alternatif échantillonnées à la seconde. Le volume total représente 60000 heures de vol de 18 appareils répartis en 8 groupes selon les conditions de navigation. Sur le plan technique, un générateur peut supporter une surcharge dite intermittente sur une faible durée. Nous ne nous intéressons pas à ces phénomènes dans la suite et nous ne retenons que les charges dites permanentes en mode nominal durant la phase de vol et la phase de roulage prises séparément.

Pour une série temporelle donnée, nous définissons la série $(Y_i)_{i \in \{1, \dots, N\}}$ des rapports entre la charge observée et la charge maximale. Ces rapports peuvent être plus grands que 1 si la charge dépasse celle donnée par l'ELA. Nous éliminons les charges intermittentes par une moyenne mobile de largeur

T définie par

$$X_k = \frac{1}{T} \sum_{i=1}^T Y_{(k-1)T+i}, \quad 1 \leq k \leq n = \lfloor N/T \rfloor.$$

En régime stationnaire durant les phases de vol considérées, les variables X_k peuvent raisonnablement être considérées comme indépendantes et de même loi qu'une variable X . Pour une probabilité $p \in (0, 1)$ petite, nous sommes intéressés par l'estimation du quantile q_p tel que

$$\mathbb{P}(X > q_p) = p$$

et par celle du point terminal x^* potentiellement infini défini par

$$x^* = \sup \{x > 0 : \mathbb{P}(X \leq x) < 1\}.$$

La théorie des valeurs extrêmes est utilisée dans un très grand nombre de domaines appliqués et nous renvoyons à [CHBS05] pour une introduction au sujet. La première étape élémentaire consiste à noter que nous avons par indépendance

$$\mathbb{P}(\max\{X_1, \dots, X_n\} \leq x) = F_X^n(x)$$

où F_X désigne la fonction de répartition de X . Pour considérer le maximum des X_i dans un cadre asymptotique où cette probabilité ne tend pas vers une limite triviale, il convient de le normaliser en introduisant des suites $(a_k)_{k \geq 1}$ de $(0, +\infty)$ et $(b_k)_{k \geq 1}$ de \mathbb{R} telles que

$$\frac{\max\{X_1, \dots, X_n\} - b_n}{a_n}$$

converge en loi au sens d'une fonction de répartition limite G ,

$$\lim_{n \rightarrow \infty} F_X^n(a_n x + b_n) = G(x).$$

L'existence de telles suites est assurée pour une large classe de fonctions de répartition F_X . Les candidates à la fonction de répartition limite G sont connues et leurs paramètres peuvent être estimés par maximum de vraisemblance.

Nous considérons l'approche dite de Pareto généralisée pour traiter nos données. Cela consiste à fixer un seuil $u > 0$ et à considérer la loi de $X - u$ conditionnellement à l'événement $X > u$. Ce seuil est autorisé à dépendre de n et, pour une valeur $u_n > 0$ assez grande, la fonction de répartition empirique des variables $X_1 - u_n, \dots, X_n - u_n$ telles que $X_i > u_n$ peut être

approchée par la fonction de répartition de la loi de Pareto généralisée de paramètres $\mu, \sigma > 0$ et $\xi \in \mathbb{R}$,

$$\forall x > 0, H(x) = \begin{cases} 1 - (1 + \xi x/\beta)^{-1/\xi} & \text{si } \xi \neq 0, \\ 1 - \exp(-x/\beta) & \text{si } \xi = 0, \end{cases}$$

avec un paramètre d'échelle $\beta = \sigma + \xi(u - \mu)$. Pour que ce résultat soit théoriquement valide il faut un nombre d'observations n grand et un rapport K_n/n petit avec

$$K_n = |\{k \in \{1, \dots, n\} : X_k > u_n\}|.$$

La recherche d'un bon seuil u_n s'apparente à un compromis biais-variance : une valeur élevée assure un biais petit au sens où la loi de Pareto généralisée est une bonne approximation mais une variance d'autant plus élevée que le nombre K_n d'observations retenues est faible (voir [dHF06]). En pratique, des outils graphiques d'aide à la décision permettent de choisir un seuil raisonnable.

Les paramètres ξ et β de la loi de Pareto généralisée peuvent être estimés par les estimateurs du maximum de vraisemblance $\hat{\xi}$ et $\hat{\beta}$. Il s'ensuit l'estimation du quantile q_p par

$$\hat{q}_p = \begin{cases} u_n + \frac{\hat{\beta}}{\hat{\xi}} \left(\left(\frac{K_n}{np} \right)^{\hat{\xi}} - 1 \right) & \text{si } \hat{\xi} \neq 0 \\ u_n + \hat{\beta} \log \left(\frac{K_n}{np} \right) & \text{si } \hat{\xi} = 0. \end{cases}$$

Lorsque $\hat{\xi} < 0$, il est possible de faire tendre p vers 0 pour obtenir un estimateur du point terminal x^* ,

$$\hat{x}^* = u_n - \hat{\beta}/\hat{\xi}.$$

Des intervalles de confiance asymptotiques pour \hat{q}_p et \hat{x}^* sont disponibles dans la littérature (voir [dHF06]).

Nous avons utilisé les outils présentés dans les paragraphes précédents pour étudier la charge des générateurs alternatifs à bord des avions. La même procédure d'estimation des quantiles et des points terminaux a été appliquée aux 8 groupes de données, les estimations de ξ étant toutes significativement négatives. La procédure statistique de test développée par [EEdH19] a été utilisée pour valider l'égalité des points terminaux obtenus sur chacun des groupes de données avec un niveau de confiance de 95%. Ce résultat conduit à considérer que le rapport maximal entre la charge électrique observée et celle donnée par l'ELA est estimé à 80%, indépendamment de la compagnie aérienne ou des conditions de navigation.

Apprentissage avec données fonctionnelles

Les nombreux capteurs embarqués dans les avions fournissent un volume important de données en vol. Pour un système électrique, nous avons présenté ci-dessus des données temporelles échantillonnées à la seconde pour la charge d'un générateur. À chaque seconde, d'autres mesures liées aux générateurs électriques sont disponibles telles que, par exemple, la température de l'huile utile pour détecter un fonctionnement anormal du générateur. Il est aussi possible de coupler ces informations avec d'autres sources de données comme la vitesse moteur de l'avion, la température de l'air extérieur, l'altitude, ... Le point commun de ces données est qu'elles se représentent comme des fonctions discrétisées sur un même support. Nous présentons ci-dessous des méthodes d'apprentissage supervisé dans le cadre fonctionnel pour aborder des problèmes de régression dans le domaine aéronautique.

Le cadre statistique multivarié que nous considérons est donné par q variables fonctionnelles X^1, \dots, X^q et une variable fonctionnelle de sortie Y observées à chaque seconde durant N vols de longueurs différentes. Pour le vol $\ell \in \{1, \dots, N\}$, nous notons n_ℓ sa durée en secondes. Il n'est pas simple de manipuler des données fonctionnelles définies sur des domaines de longueurs différentes. Normaliser les durées introduit un biais méthodologique entre les vols et ne permet pas de construire des méthodes utilisables en vol puisque la durée totale n'est connue qu'une fois le vol terminé. L'alternative que nous retenons est de prédire la valeur de la variable de sortie Y toutes les T secondes à partir des q fonctions d'entrée observées sur cette fenêtre temporelle. En posant $m_\ell = \lfloor n_\ell/T \rfloor$, les entrées du vol ℓ se mettent sous la forme d'une matrice X_ℓ de taille $m_\ell \times qT$ où chaque ligne correspond à la concaténation des valeurs des fonctions d'entrée observées durant une fenêtre de temps T ,

$$X_\ell = \begin{pmatrix} \dots & X^j(1) & \dots & X^j(T) & \dots \\ \dots & X^j(T+1) & \dots & X^j(2T) & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \dots & X^j((m_\ell-1)T+1) & \dots & X^j(m_\ell T) & \dots \end{pmatrix}.$$

De même, la sortie est donnée par $y_\ell = (Y(T), \dots, Y(m_\ell T))^\top \in \mathbb{R}^{m_\ell}$.

Les éléments de la matrice X_ℓ s'identifient aux observations de fonctions $x_k^j : [0, 1] \rightarrow \mathbb{R}$, $j \in \{1, \dots, q\}$ et $k \in \{1, \dots, m_\ell\}$,

$$x_k^j(t_i) = X^j((k-1)T + i)$$

sur le support des points $t_i = i/T$, $i \in \{1, \dots, T\}$. Dans la suite, nous identifions la fonction x_k^j au vecteur $(x_k^j(t_1), \dots, x_k^j(t_T))^\top \in \mathbb{R}^T$. Puisque les unités

des variables d'entrée peuvent être différentes, nous normalisons les observations x_k^j à l'aide des moyennes

$$\bar{x}^j = \frac{1}{m_\ell} \sum_{k=1}^{m_\ell} x_k^j \in \mathbb{R}^T, \quad j \in \{1, \dots, q\},$$

et des termes d'inertie

$$\sigma_j^2 = \frac{1}{m_\ell} \sum_{k=1}^{m_\ell} \|x_k^j - \bar{x}^j\|^2, \quad j \in \{1, \dots, q\}.$$

Ainsi, le jeu de données des entrées fonctionnelles normalisées sur $[0, 1]$ correspond à des fonctions $f_k^j : [0, 1] \rightarrow \mathbb{R}$, $j \in \{1, \dots, q\}$ et $k \in \{1, \dots, m_\ell\}$, identifiées aux vecteurs $f_k^j = (f_k^j(t_1), \dots, f_k^j(t_T))^\top \in \mathbb{R}^T$ donnés par

$$f_k^j = \frac{x_k^j - \bar{x}^j}{\sqrt{\sigma_j^2}} \in \mathbb{R}^T, \quad j \in \{1, \dots, q\}, \quad k \in \{1, \dots, m_\ell\}.$$

Les fonctions f_k^j sont supposées de carré intégrable sur $[0, 1]$ et nous considérons une base $\{\xi_d\}_{d \in \mathbb{N}}$ de $L^2([0, 1])$ pour les représenter. Comme dans [RS05], les ξ_d peuvent correspondre à la base de Fourier, à une base d'ondelettes, ... Pour un entier $D < T$ à calibrer, nous réduisons la dimension du problème en ne retenant que les D premiers coefficients pour chaque fonction

$$c_{k,d}^j = \frac{1}{T} \sum_{i=1}^T f_k^j(t_i) \xi_d(t_i), \quad d \in \{1, \dots, D\}.$$

Le problème de régression que nous considérons se ramène ainsi à prédire le vecteur de sortie $y \in \mathbb{R}^{m_\ell}$ à partir de la matrice d'entrée C_ℓ de taille $m_\ell \times qD$ donnée par les coefficients calculés pour le vol ℓ ,

$$C_\ell = \begin{pmatrix} \cdots & c_{1,1}^j & \cdots & c_{1,D}^j & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \cdots & c_{m_\ell,1}^j & \cdots & c_{m_\ell,D}^j & \cdots \end{pmatrix}.$$

Plus généralement, la même procédure peut être appliquée à chacun des N vols. La concaténation des lignes des matrices C_ℓ donne une matrice d'entrée C de taille $n \times p$, $n = m_1 + \dots + m_N$ et $p = qD$, et celle des vecteurs y_ℓ donne un vecteur de sortie $y \in \mathbb{R}^n$

Différentes méthodes de régression classiques ont été appliquées à ces données dans le cadre aéronautique : régression ridge, forêts aléatoires, réseaux de neurones, ... et nous renvoyons à [HTF09] pour plus de détails sur

ces procédures. Ces études sont décrites dans [Bou21] et certaines ont fait l'objet de la présentation [XG-Conf4] qui traite de la détection de fonctionnement anormal de générateur électrique à bord des avions. Dans la suite, la présentation se limite à la méthode de régularisation dropout, introduite dans [HSK⁺12] et [SHK⁺14], qui correspond à une de mes perspectives de recherche.

Étant donnée une classe de prédicteurs $\{f_w : \mathbb{R}^p \rightarrow \mathbb{R}, w \in W\}$ et pour la fonction de perte des moindres carrés, le problème de la régression décrit ci-dessus revient à minimiser

$$w \in W \mapsto L(w) = \frac{1}{n} \sum_{i=1}^n (y_i - f_w(c_i))^2$$

où $c_i \in \mathbb{R}^p$ désigne la $i^{\text{ème}}$ ligne de C . Lorsque cela est numériquement possible, une solution à ce problème d'optimisation consiste à construire une suite $(w_t)_{t \in \mathbb{N}}$ par descente de gradient,

$$w_{t+1} = w_t - \eta \nabla L(w_t), \quad t \geq 0,$$

où $w_0 \in W$ est une valeur initiale donnée et $\eta > 0$ est un pas de descente fixe à déterminer. Sous certaines hypothèses (voir [Nes04]), w_t converge vers un minimiseur de L quand t tend vers l'infini. Cette méthode est cependant limitée en pratique car elle nécessite un calcul explicite du gradient de L à chaque itération qui s'avère coûteux en temps si la taille des données est grande. De plus, la minimisation de L se fait sur les données observées sans tenir compte de l'erreur relative à la variabilité de ces données. Le prédicteur associé au minimiseur est sujet au surapprentissage et une forme de régularisation est nécessaire pour l'éviter. La régularisation dropout offre une façon de contourner ces difficultés en pratique.

Dans le contexte de la régression linéaire associée aux prédicteurs $f_w(c) = w^\top c$, $w \in \mathbb{R}^p$, [MAV18] présente la régularisation dropout comme la minimisation de la fonction de perte définie pour un paramètre $\rho \in [0, 1]$ par

$$w \in \mathbb{R}^p \mapsto L_\rho(w) = \mathbb{E}_b \left[\frac{1}{n} \sum_{i=1}^n (y_i - w^\top \text{diag}(b)c_i)^2 \right]$$

où $\text{diag}(b)$ désigne la matrice diagonale dont les éléments diagonaux sont donnés par le vecteur aléatoire $b = (b_1, \dots, b_p)^\top$ avec b_j indépendantes de loi $(1 - \rho)^{-1} \text{Ber}(1 - \rho)$. Un simple calcul conduit à la décomposition

$$L_\rho(w) = L(w) + \frac{\rho}{1 - \rho} \sum_{j=1}^p \gamma_j^2 w_j^2$$

où γ_j^2 est la variance empirique de la $j^{\text{ème}}$ colonne de C . Ainsi, pour la régression linéaire, le dropout apparaît comme une régularisation de Tikhonov des moindres carrés et préserve le minimiseur de L_ρ du phénomène de surapprentissage. L'expression de L_ρ comme une espérance d'une moyenne empirique permet de traiter ce problème d'optimisation par une descente de gradient stochastique $(w_{\rho,t})_{t \in \mathbb{N}}$,

$$w_{\rho,t+1} = w_{\rho,t} - \eta_{t+1} \nabla L_\rho^{(t+1)}(w_{\rho,t}), \quad t \geq 0,$$

avec un pas $\eta_{t+1} > 0$ autorisé à dépendre de l'indice d'itération et des fonctions de perte aléatoires $(L_\rho^{(t)})_{t \in \mathbb{N}}$ données par

$$w \in \mathbb{R}^p \longmapsto L_\rho^{(t)}(w) = \frac{1}{|B_t|} \sum_{i \in B_t} (y_i - w^\top \text{diag}(b_t) c_i)^2$$

où les vecteurs b_t sont indépendants de même loi que b et les B_t sont des sous-ensembles aléatoires de $\{1, \dots, n\}$ de même taille appelés batches. Il convient en particulier de noter que les batches B_t limitent le coût de calcul du gradient en le restreignant à un ensemble de données généralement petit par rapport à n . Sous certaines conditions, la suite $(w_{\rho,t})_{t \in \mathbb{N}}$ converge vers un minimiseur de L_ρ et conduit à un prédicteur régularisé plus simple à calculer que celui construit par descente de gradient déterministe.

La régularisation dropout s'étend à des classes d'estimateurs bien plus générales que la régression linéaire. Le principe reste le même : à chaque itération d'une descente de gradient stochastique, l'optimisation ne porte que sur une partie aléatoire des paramètres choisie par tirage de variables de Bernoulli de paramètre $1 - \rho$. Plus ρ est proche de 1, plus le nombre de paramètres à utiliser les données d'un batch a tendance à être faible. Puisque l'entraînement de chaque paramètre ne se fait pas sur l'ensemble des données, il est impossible de surapprendre et le prédicteur obtenu se retrouve régularisé. Malheureusement, la forme de la régularisation n'est pas toujours aussi simple que dans le cas de la régression linéaire et peu de travaux théoriques ont été proposés sur le dropout. Pour les réseaux de neurones, le dropout est largement utilisé en pratique (voir [GBC16]) et quelques résultats ont par exemple été obtenus dans [ABMS21] mais les propriétés statistiques des prédicteurs associés restent encore largement à explorer dans ce contexte.

Chapitre 4

Applications environnementales

Plusieurs rencontres scientifiques m'ont amené à collaborer sur des sujets environnementaux. Ces travaux se situent à la limite de la statistique et de la science des données. Au-delà des mathématiques appliquées, cette implication est importante pour moi car elle donne un sens au fait de faire de la statistique dans un monde bien plus grand.

4.1 Écologie forestière

Cette section présente des résultats publiés dans [\[XG-Article04\]](#) et [\[XG-Article06\]](#).

Retombées atmosphériques d'azote et climat

Les activités humaines ont significativement contribué à une augmentation des émissions d'azote et de soufre depuis la fin du 19^{ème} siècle. Il est connu que les retombées atmosphériques ont un impact fort sur le fonctionnement des écosystèmes forestiers en modifiant la chimie du sol et l'équilibre des nutriments, perturbant de fait la croissance des arbres et la biodiversité. Depuis les années 1980, un effort commun des pays d'Europe vise à réduire les émissions atmosphériques. Si les émissions de soufre ont diminué de près de 80% en France, la diminution des émissions d'azote est moins évidente avec environ 35% de moins pour NO_x et 5% pour NH_y . Le cycle de l'azote est plus complexe que celui du soufre car l'azote interagit avec les écosystèmes à de nombreux niveaux.

Les effets des émissions et des retombées d'azote sur les écosystèmes ont fait l'objet de nombreux travaux depuis les années 2000. Plusieurs expérimentations ont permis de montrer que la concentration d'azote dans le sol est une mesure importante pour évaluer l'impact des retombées atmosphé-

riques sur un écosystème donné. Cependant, ces travaux ne permettent pas de faire des prédictions sur le long terme et une étape de modélisation est nécessaire, en particulier pour prédire la chimie du sol. Des modèles dynamiques de simulation ont été développés afin de tester différents scénarios pour les retombées atmosphériques d’azote dans le temps.

La chimie du sol est fortement affectée par le climat et par les conditions de température et d’humidité. Ainsi, les impacts du changement climatique et les retombées d’azote doivent être considérées conjointement pour l’étude de la chimie du sol sur le long terme. C’est dans ce contexte que nous avons utilisé avec Noémie Gaudio le modèle ForSAFE [WSSB05] qui est développé pour la modélisation du milieu forestier. Ce modèle permet de simuler l’évolution dans le temps d’un écosystème forestier en fonction des caractéristiques du sol, du climat, des retombées atmosphériques et des propriétés forestières. Deux sites forestiers français du réseau RENECOFOR ont été utilisés comme références et leurs évolutions ont été simulées jusqu’en 2100 sous deux scénarios de retombées atmosphériques (la référence pour la législation européenne et celui de plus grande réduction faisable) et sous trois scénarios climatiques (poursuite de la croissance des émissions actuelles, commun effort de diminution avec forte prise de conscience environnementale et un scénario fictif de référence où le climat demeure tel qu’il est aujourd’hui).

Une première étape statistique importante consistait à valider les prédictions du modèle ForSAFE. Pour ce faire, des données simulées ont été comparées à des mesures réelles sur la période 1993-2009. Une fois paramétré, le modèle ForSAFE retourne des séries temporelles pour le pH du sol ainsi que la concentration de certains éléments chimiques (carbone, azote, ...). En notant $Y \in \mathbb{R}^n$ la différence entre une série simulée et les mesures observées, la validation du modèle correspond à un test statistique de nullité de la moyenne de Y . Sous l’hypothèse que Y soit un vecteur Gaussien de matrice de covariance $\sigma^2 I_n$, nous avons développé un test multiple basé sur des arguments de sélection de modèle à la façon de [BHL03]. Ce type de test d’hypothèse nulle « $\mathbb{E}[Y] = 0$ » est adaptatif, il n’impose pas la connaissance du facteur de variance σ^2 et il se construit à partir d’une collection dénombrable $\{S_m, m \in \mathcal{M}\}$ de sous-espaces de \mathbb{R}^n . Pour $m \in \mathcal{M}$, π_m désigne la projection orthogonale sur l’espace S_m de dimension D_m et nous considérons un niveau de confiance $\alpha \in (0, 1)$ ainsi que des réels $\{\alpha_m, m \in \mathcal{M}\}$ dans $(0, 1)$ tels que

$$\alpha = \sum_{m \in \mathcal{M}} \alpha_m.$$

La statistique de test est donnée par

$$T_\alpha = \sup_{m \in \mathcal{M}} \left\{ \frac{(n - D_m) \|\pi_m Y\|^2}{D_m \|Y - \pi_m Y\|^2} - F_{D_m, n - D_m}^{-1}(\alpha_m) \right\}$$

où $F_{D,N}^{-1}$ est la fonction quantile d'une loi de Fisher de paramètres (D, N) . L'hypothèse nulle est rejetée dès lors que T_α est positif. Appliqué aux données générées par ForSAFE, ce test a permis de valider la pertinence des simulations sur la période d'observation considérée.

Des analyses de variance ont été faites sur les données générées par le modèle ForSAFE sur le long terme pour déterminer les effets du scénario climatique et ceux des retombées atmosphériques d'azote. Pour les 6 combinaisons de scénarios retenues, les moyennes des variables simulées ont été comparées à l'aide de tests de Tukey. Les résultats ont mis en évidence que l'alcalinité du sol dépend bien plus fortement du changement climatique que du niveau des retombées d'azote à venir. D'un autre côté, les retombées d'azote devraient être un facteur plus décisif que le changement climatique pour le phénomène d'eutrophisation du sol, *i.e.* l'accumulation de nutriments. Ces conclusions illustrent la pertinence de considérer simultanément les effets du changement climatique et les retombées atmosphériques d'azote pour étudier la dynamique d'un écosystème forestier.

Température et lumière sous la canopée

L'abri forestier héberge des microclimats qui offrent des conditions de température et de luminosité différentes des champs ouverts. La majorité des études concernant ces microclimats ne prennent en compte que des valeurs maximales, minimales ou moyennes de la température et de la luminosité pour comparer les paramètres climatiques en sous-bois et en champs ouverts. Ce faisant, ces approches négligent les dynamiques temporelles liées au degré d'ouverture de la canopée et à la saisonnalité. Dans le contexte global du changement climatique, la durabilité de tels écosystèmes fait l'objet de nombreuses études. Les approches macroclimatiques et microclimatiques doivent être considérées conjointement pour comprendre les conditions de fraîcheur et d'humidité auxquelles les plantes des sous-bois sont particulièrement sensibles en lien avec d'autres variables telles que la luminosité.

Pour des raisons pratiques, les capteurs de températures sont généralement intégrés à des stations météorologiques en champs ouverts qui servent de références pour estimer la température forestière. Cependant, il est reconnu que la température seule n'est pas une mesure suffisante pour expliquer le développement des plantes dans ces écosystèmes, que la densité de la canopée influe sur ces microclimats de sous-bois et qu'il existe des différences

systematiques entre les températures en sous-bois et en champs ouverts. De plus, l'échelle des systèmes considérés dans les études écologiques basées sur des données climatiques est généralement de l'ordre du kilomètre alors que les dimensions de ces microclimats sont de quelques mètres. Les variations de température relatives à celles de la luminosité y sont donc différentes des observations effectuées en champs ouverts. Afin de décrire l'évolution corrélée de la température et de la luminosité dans de tels microclimats, nous avons proposé une étude exploratoire de données issues de mesures horaires effectuées en sous-bois avec la prise en compte du degré d'ouverture de la canopée quantifié à partir de photographies prises simultanément.

Pour chaque jour d'observation $i \in \{1, \dots, n\}$, nous disposons des 24 mesures horaires de température $Y_i^{(1)} = (Y_{i,1}^{(1)}, \dots, Y_{i,24}^{(1)})^\top$ et de luminosité $Y_i^{(2)} = (Y_{i,1}^{(2)}, \dots, Y_{i,24}^{(2)})^\top$. Les unités de ces variables étant incompatibles, les vecteurs sont centrés et réduits,

$$\tilde{Y}_i^{(\ell)} = \frac{Y_i^{(\ell)} - \bar{Y}_i^{(\ell)}}{\sqrt{\sigma_{i,\ell}^2}} \text{ avec } \bar{Y}_i^{(\ell)} = \frac{1}{24} \sum_{j=1}^{24} Y_{i,j}^{(\ell)} \text{ et } \sigma_{i,\ell}^2 = \frac{1}{24} \sum_{i=1}^{24} (Y_{i,j}^{(\ell)} - \bar{Y}_i^{(\ell)})^2$$

pour $\ell \in \{1, 2\}$. Ces vecteurs normalisés sont concaténés pour définir le jeu des données du jour i ,

$$Y_i = (\tilde{Y}_{i,1}^{(1)}, \dots, \tilde{Y}_{i,24}^{(1)}, \tilde{Y}_{i,1}^{(2)}, \dots, \tilde{Y}_{i,24}^{(2)})^\top \in \mathbb{R}^{48}.$$

Les observations ont été réalisées pendant 335 jours sur un site de référence en champs ouvert et 8 sites en sous-bois mesurés sur 5 emplacements distincts. Le volume total de données est ainsi de $n = 13735$ vecteurs définis comme ci-dessus.

Compte tenu de la nature temporelle des données, nous avons mis en œuvre une analyse en composantes principales fonctionnelles (voir [RS05]). Les deux premières composantes expliquent 79% de l'inertie mais conduisent à une représentation graphique dans le plan principal qui s'avère peu pratique à cause du nombre important de points et de la difficile interprétation des axes. Pour aider à l'analyse de ces résultats, nous avons développé des outils de visualisation de données spécifiques. Le premier outil est basé sur des estimateurs à noyaux de la densité (voir [Pol95] et [Bai03]) de sous-groupes de données et représente les enveloppes contenant 75% des observations de chaque groupe. Un exemple de résultat est donné par la figure 4.1 pour visualiser l'effet de l'ouverture de la canopée sur les profils de température et de luminosité. En plus des ensembles de niveaux, ce graphique donne aussi les quantiles à 10%, 25%, 50%, 75% et 90% pour l'ensemble des données le long de chaque axe, définissant ainsi une grille de lecture du plan principal. Lorsque elle existe, la donnée la plus proche de chaque intersection

est sélectionnée et le couple correspondant des courbes de température et de luminosité est présenté à la manière de la figure 4.2. Cette représentation permet ainsi d'interpréter visuellement les axes du plan principal.

La grille donnée par la figure 4.2 montre que la première composante principale correspond à l'effet attendu de corrélation positive entre température et luminosité. Plus un point a une abscisse importante, plus la température et la luminosité du jour sont simultanément élevées. Verticalement, le phénomène qui apparaît est celui de l'inertie de température dans les sous-bois avec une corrélation négative entre luminosité et température. Plus un point a une ordonnée importante, plus la température du jour a tendance à être élevée malgré une luminosité faible. Aidé par cette interprétation des axes, la figure 4.1 permet de mettre en évidence le rôle de l'ouverture de la canopée sur cet effet de température en sous-bois. Cela confirme l'intérêt d'étudier les microclimats de sous-bois en tenant compte de ces phénomènes inertiels qui induisent du stress thermique pour la végétation non expliqué par la température mesurée seule en champs ouvert.

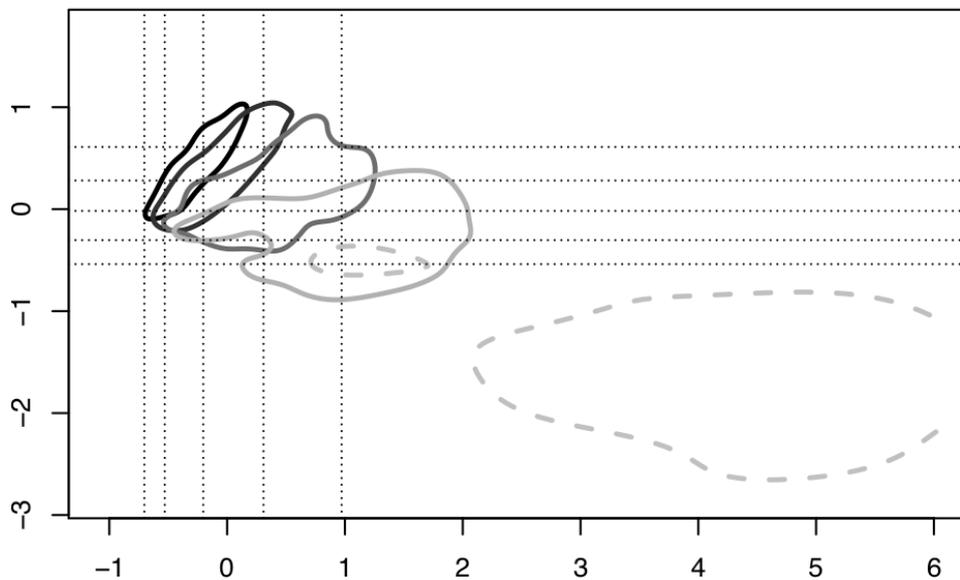


FIGURE 4.1 – Ensembles de niveaux contenant 75% des couples de courbes de température et de luminosité dans le plan principal selon le degré d'ouverture de la canopée. L'enveloppe grise en tiret représente les données en champs ouvert. Les enveloppes en trait plein vont d'une canopée dense (noir) à des situations plus ouvertes (gradient de gris).

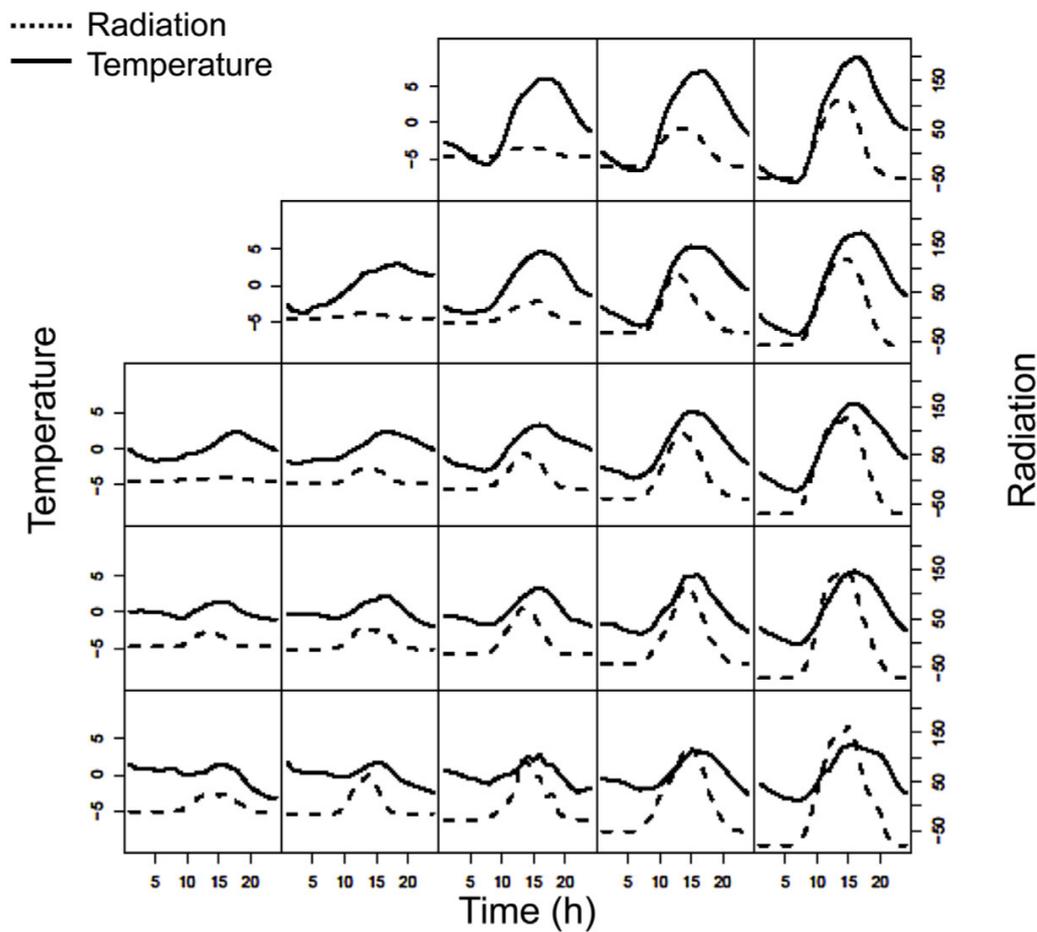


FIGURE 4.2 – Représentations centrées des paires de courbes de température et de luminosité quotidiennes correspondant aux intersections des quantiles du plan principal.

4.2 Pollution de plastique des océans

Cette section présente des résultats publiés dans [\[XG-Article05\]](#).

Expéditions 7ème Continent

L'objectif scientifique des expéditions 7ème Continent est la compréhension des effets de la pollution de plastique sur l'équilibre des écosystèmes marins. Ces expéditions permettent de collecter des observations de cette pollution et de ramener des échantillons de débris plastiques dérivant à proximité de la surface. Ces débris se concentrent dans d'immenses zones d'accumulation de la taille d'un continent. Sous les effets des courants marins, les déchets plastiques se retrouvent piégés dans des gyres océaniques qui sont l'objet des explorations du 7ème Continent. Les travaux présentés ci-dessous sont le fruit d'une collaboration avec Alexandra ter Halle, responsable du volet scientifique de ces expéditions.

La pollution de plastique est majoritairement constituée de polyéthylène qui est un matériau très utilisé dans les activités humaines : emballages, sacs plastiques, ... Une fois dans le milieu marin, les débris se fragmentent sous forme de paillettes de quelques millimètres appelées microplastiques. L'expédition 7ème Continent de 2014 a recueilli de tels débris dans le gyre Atlantique Nord. Les collectes ont été effectuées à l'aide d'un filet de type manta à maille fine de $300 \mu\text{m}$ plongé dans les 30 cm de la surface par temps calme. De nombreux débris ont ainsi été ramassés et triés manuellement afin de mener notre étude statistique. En dehors des fils de pêche qui n'ont pas été considérés, 1093 débris plastiques ont été débarassés de la matière organique, lavés et pesés à 10^{-2} mg près.

Dans la littérature, les débris plastiques sont généralement catégorisés par taille en fonction de leur diamètre. Cette métrique est peu pertinente pour considérer leur fragmentation car elle n'est pas conservative au cours du processus. En particulier, cela conduit à des erreurs de raisonnement qui tendent à mal évaluer la quantité de microplastiques. Dans notre étude, nous avons considéré la masse des débris qui est une grandeur conservative puisque elle est préservée lors de la fragmentation. Contrairement aux catégories de taille, les masses des débris plastiques semblent bien distribuées selon une loi de puissance que nous avons étudiée en proposant un modèle de fragmentation des plastiques marins.

Fragmentation des débris plastiques

L'objectif de notre étude est d'étudier la distribution des masses des fragments de microplastiques. Nous ne considérons pas le temps de fragmentation dans la suite car il est impossible avec les connaissances actuelles de dater précisément l'âge d'un débris plastique. Nous ne nous intéressons donc que à la distribution moyenne de la masse des débris issus d'une fragmentation qui

serait régulièrement alimentée par de nouveaux microplastiques comme l'est le phénomène à l'œuvre dans les gyres océaniques. Un modèle de fragmentation théorique neutre peut se construire comme un processus auto-similaire (voir [Ber06]). Étant donnés n débris plastiques, nous introduisons la liste de leurs masses

$$\mathcal{M}_n = \{x_1, \dots, x_n\}.$$

Une itération du processus de fragmentation est donnée par les étapes suivantes :

1. Choisir aléatoirement un débris $K \sim \mathcal{U}(\{1, \dots, n\})$ à fragmenter.
2. Tirer $U \sim \mathcal{U}([0, 1])$ pour définir les masses de deux nouveaux fragments Ux_K et $(1 - U)x_K$.
3. Remplacer le fragment de masse x_K pour définir la nouvelle liste des $n + 1$ masses

$$\mathcal{M}_{n+1} = \{x_1, \dots, x_{K-1}, Ux_K, (1 - U)x_K, x_{K+1}, \dots, x_n\}.$$

La loi uniforme de la variable U correspond à l'hypothèse que la fragmentation d'un débris ne dépend pas de sa géométrie.

Afin d'être utilisé, ce processus de fragmentation aléatoire doit être initialisé. Pour cela, nous considérons les 10% des masses observées les plus lourdes qui correspondent à 110 fragments allant de 4.93 mg à 13.81 mg. Le processus de fragmentation est simulé un grand nombre de fois à partir de ces données initiales pour estimer la distribution théorique des masses de fragments devant être observée pour le modèle de fragmentation neutre. Cette distribution théorique est ensuite comparée à la distribution des observations pour les fragments de masses comprises entre m_{\min} et 4.93 mg pour m_{\min} allant de 0.01 mg à 3.43 mg de sorte que les données utilisées représentent toujours au moins 10% de la taille de l'échantillon.

L'objectif de cette approche consiste à détecter à partir de quelle masse le modèle de fragmentation neutre des débris ne correspond plus à ce qui est observé. L'adéquation des masses observées à la distribution théorique fait l'objet d'un test de Kolmogorov-Smirnov à 5% pour chaque valeur de m_{\min} . Le principe consiste à accepter l'adéquation lorsque la statistique de test est inférieure à un seuil théorique qui dépend du nombre d'observations retenues et du niveau de confiance. Les résultats sont présentés dans la figure 4.3.

Le modèle de fragmentation neutre est accepté pour les débris de plus de 1 mg et nettement rejeté à partir d'une masse minimale inférieure à 0.8 mg. Ce rejet est la conséquence de la faible masse observée pour les fragments de petite taille. La somme des masses des fragments de moins de 1 mg représente 240.08 mg dans l'échantillon alors que la valeur théorique attendue est autour

de 4800 mg, ce qui est 20 fois plus grand. Ce fort déficit de débris légers indique que d'autres phénomènes sont à l'œuvre pour les microplastiques de moins de 1 mg. Certaines études suggèrent une fragmentation plus rapide pour les débris les plus légers mais, compte tenu de leur taille, l'ingestion par des animaux marins ou la sédimentation peuvent aussi expliquer le manque de tels microplastiques dans la surface océanique.

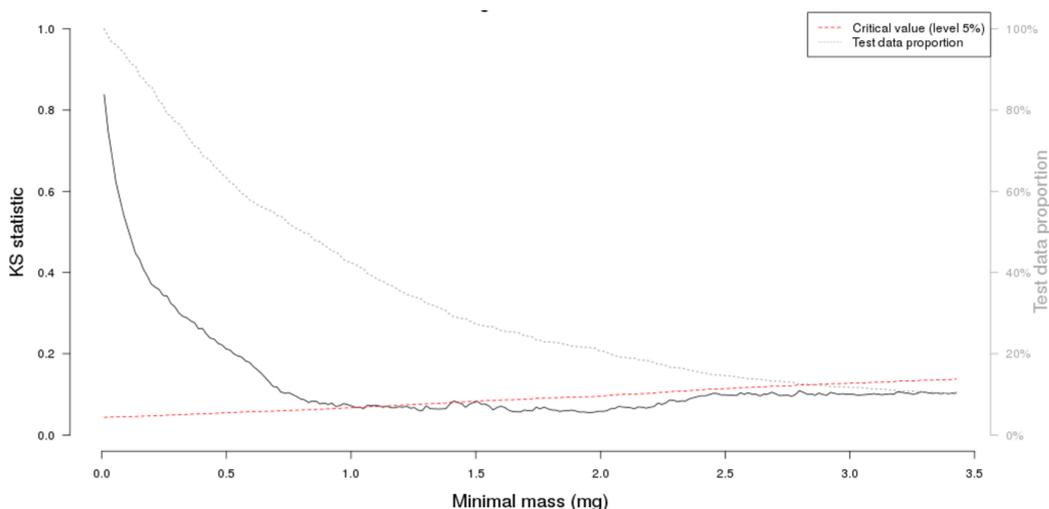


FIGURE 4.3 – Statistique de test de Kolmogorov-Smirnov (noir) et valeur critique de rejet (rouge) en fonction de la masse m_{\min} . Le volume total de données impliqué pour chaque valeur de m_{\min} est tracé en gris.

4.3 Allométrie en cultures mixtes

Cette section présente des résultats présentés dans [XG-Conf1] et publiés dans [XG-Article10] à partir des données [XG-Data1].

Introduction

Les cultures mixtes sont des combinaisons de deux espèces ou plus dans un même champs. En exploitant les ressources différemment et plus efficacement, ces cultures ouvrent des possibilités prometteuses dans le contexte d'une agriculture durable. La plupart des cultures mixtes mêlent une céréale et une légumineuse car l'azote atmosphérique fixé par la légumineuse bénéficie à la céréale. Ce bénéfice permet de diminuer le besoin d'engrais azotés et

l'intérêt agronomique pour les cultures mixtes est motivé par la possibilité d'un meilleur rendement.

En écologie, l'allométrie correspond à l'étude de la vitesse de croissance des organismes. En particulier, la taille d'une plante est connue pour être un indicateur principal du rendement annuel en grain. Cette relation est à la base de relations macroécologiques intéressantes pour concevoir et gérer des systèmes de cultures mixtes. Dans le cadre de ce travail, nous avons étudié le lien entre l'allométrie et les performances des cultures mixtes céréale-légumineuse et l'influence des conditions de culture sur cette relation. Les données sont issues d'expérimentations effectuées dans 28 champs d'Europe de l'ouest dans différentes conditions de culture (mixte ou non, avec et sans apports azotés) et de climat.

Analyse des données

La variable d'intérêt pour mesurer la performance est le rapport du rendement en grain y sur la biomasse aérienne totale à maturité x . Nous avons analysé le lien allométrique entre ces variables par la méthode de l'axe majeur normalisé (SMA pour *Standardized Major Axis*) telle que introduite dans [WWF06] qui considère la relation

$$y = ax^b$$

pour des paramètres réels a et b à déterminer. Cette méthode populaire dans les études d'allométrie a une interprétation géométrique d'alignement de droites dites allométriques proche de ce qui peut être fait en régression linéaire. En particulier, des tests statistiques pour comparer les pentes, les ordonnées à l'origine et le décalage ont été proposés par [TW13]. Ces outils nous permettent d'évaluer l'effet du type de plante, de la culture mixte et de la fertilisation azotée sur la performance.

Lorsque deux expérimentations induisent des pentes de droites allométriques significativement différentes, nous considérons la biomasse aérienne x_0 correspondante à leur intersection. La taille des plantes pour cette biomasse permet alors de définir un seuil à partir duquel le rendement d'une expérimentation est meilleur que l'autre. À l'aide des résultats présentés dans [FM72], il est possible de donner un intervalle de confiance sur l'abscisse x_0 et donc sur le seuil de la taille. Ces résultats sont importants pour argumenter en faveur des cultures mixtes et leur robustesse a été validée par bootstrap sur le jeu de données complet.

Nous avons également introduit un indice de différence de biomasse pour évaluer la domination d'une espèce dans une culture mixte. Cet indice varie

de -1 (espèce dominante) à $+1$ (espèce dominée) et une relation linéaire avec le rendement a été proposée. Cet aspect de notre étude offre une aide à la décision en direction des agriculteurs désireux de concevoir des cultures mixtes.

Les résultats de notre étude indiquent que la légumineuse a une plus grande biomasse et un rendement en grain plus important que la céréale. Les pentes des droites allométriques sont plus raides en culture mixte qu'en monoculture, ce qui correspond à un effet plus important en particulier en l'absence d'engrais azotés. En culture mixte, la droite allométrique de la céréale n'est pas significativement affectée par la fertilisation mais le rendement de la légumineuse est meilleur sans fertilisation par rapport au rendement avec des engrais azotés. Cependant, avec la fertilisation azotée du champs, la céréale a un avantage compétitif par rapport à la légumineuse et la complémentarité de la culture mixte s'efface devant la domination de la céréale. Ainsi, la culture mixte semble améliorer le rendement de la céréale et de la légumineuse en particulier en l'absence de fertilisation.

Les relations allométriques apparaissent comme un nouveau cadre prometteur pour organiser et améliorer des cultures mixtes. Les outils développés dans notre étude permettent d'identifier des compromis entre des objectifs agronomiques selon l'espèce qui doit dominer. En effet, l'agriculteur peut être plus intéressé par le rendement de la légumineuse que par celui de la céréale impliquée dans la culture mixte pour limiter les maladies présentes dans le champs. La recherche de ces compromis peut être aidée par les seuils de taille proposés et cette connaissance est un levier important pour argumenter en faveur d'une transition vers des cultures plus durables.

Conclusion et perspectives

Par le présent document, j'ai proposé une synthèse de mes activités de recherche depuis l'obtention de mon doctorat. Les relations entre certains de mes travaux en statistique mathématique et les applications que j'ai pu en faire illustrent l'intérêt que je porte aux problèmes statistiques théoriques et pratiques dans leur globalité. La capacité à mettre en œuvre des méthodologies statistiques de façon numérique et concrète est l'aspect qui motive mon travail depuis plusieurs années. C'est dans cette perspective que j'ai codirigé deux thèses CIFRE Airbus, que je codirige deux thèses actuellement et que je postule pour être habilité à diriger d'autres travaux en mathématiques appliquées.

Algorithmes stochastiques Dans le contexte actuel d'une disponibilité toujours croissante de données, il est courant de considérer des sources d'information séquentielles (capteurs, web, ...). Les approches statistiques qui considèrent l'ensemble des données au moment des calculs ne sont pas adaptées à ce cadre et conduisent à des difficultés techniques rédhibitoires en pratique (accès à un grand volume de données, mise à jour en temps réel, ...). Les algorithmes stochastiques et la littérature riche sur leurs applications en statistique offrent de nombreuses pistes de travail pour développer de nouveaux outils statistiques qui répondent à ces enjeux.

Les approches du type descente de gradient stochastique sont largement utilisées en apprentissage automatique pour entraîner des prédicteurs. La régularisation dropout présentée dans la section 3.2 s'inscrit dans ce cadre. Très utilisée en pratique pour entraîner des réseaux de neurones et éviter le phénomène de surapprentissage, les propriétés de ce procédé de régularisation demeurent peu connues. Mon intérêt pour cette approche est également motivé par des applications pratiques des réseaux de neurones en physique pour des problèmes liés aux surfaces d'énergie potentielle (voir [Beh17] et [BvRMM19]). En pratique, l'entraînement de tels réseaux de neurones est souvent interrompu de façon précoce dans l'espoir de limiter une trop forte adéquation aux données. Cette méthode ne présente aucune garantie théo-

rique et le dropout représente une opportunité intéressante pour aborder le problème tant au niveau théorique que pratique.

Les algorithmes stochastiques sont également au cœur des travaux engagés pour la thèse de Marelys Crepo Navas que je codirige avec Sébastien Gadat depuis octobre 2021. L'objectif est le développement de méthodes d'estimation en ligne de paramètres d'un processus stochastique au fil des observations d'une trajectoire. Dans un cadre Bayésien, nous étudions des algorithmes inspirés de l'approche Langevin Monte Carlo pour estimer les paramètres. Les propriétés mathématiques de ces estimateurs font actuellement l'objet d'un travail actif et le sujet trouve naturellement de nombreuses issues applicatives concrètes en ingénierie financière, par exemple.

Applications en aéronautique et espace Les opportunités de collaboration offertes par l'environnement de travail à l'Institut Supérieur de l'Aéronautique et de l'Espace sont riches. En effet, l'approche statistique uniquement basée sur les données connaît un fort développement dans les domaines aéronautiques et spatiaux. Au-delà des travaux déjà entrepris, cela permet de nouvelles rencontres motivantes et des possibilités d'applications nombreuses.

Dans le prolongement de la thèse de Fériel Boulfani [Bou21], nous avons développé une méthodologie statistique de détection d'anomalie pour des données fonctionnelles multivariées. Le principe généralise celui de la méthode ICS (*Invariant Coordinate Selection*) telle que présentée dans [ANRG18]. En exploitant l'information contenue dans la matrice de covariance empirique et dans celle des moments supérieurs, nous proposons deux extensions au cadre fonctionnel dans [XG-Preprint1]. Des applications aux données aéronautiques issues de la thèse CIFRE font l'objet d'une application mais la méthode plus générale autorise son utilisation dans des contextes différents.

Depuis octobre 2021, je codirige également la thèse de Rémi Perrichon avec Thierry Klein. Les travaux en cours abordent la représentation et la classification de trajectoires d'avions. Plusieurs contraintes peuvent concerner de telles trajectoires comme les conditions météorologiques ou la gestion du trafic aérien, par exemple. Un des enjeux consiste à proposer des modèles statistiques de prédiction et de classification pour anticiper des phénomènes de retard ou d'engorgement de l'espace aérien.

Une récente collaboration avec Philippe Garnier de l'Institut de Recherche en Astrophysique et Planétologie nous amène à étudier le champ magnétique Martien. Les questions sous-jacentes ont été évoquées lors de [XG-Conf5]. Du point de vue statistique, l'objectif consiste à caractériser l'influence de certains facteurs physiques sur le champ magnétique de la planète Mars mesuré par plusieurs missions spatiales. Un premier travail (voir [XG-Preprint2])

et [XG-Preprint3]) porte en particulier sur la calibration et l'utilisation de techniques de sélection de variables telles que AIC ou le lasso en régression linéaire. D'autres approches statistiques sont également à l'étude et devraient faire l'objet de travaux à venir.

Applications environnementales Les enjeux écologiques représentent un domaine d'application important pour moi car leur importance dépasse le cadre du travail mathématique. Prendre part à des projets environnementaux est une grande source de motivation pour appliquer mes connaissances à des problématiques qui nous concernent tous mais également pour explorer de nouvelles approches statistiques motivées par les questions envisagées.

Le travail sur les cultures mixtes présenté dans la section 4.3 se poursuit avec la mise en place d'outils d'aide à la décision pour la conception et la gestion de cultures mixtes. L'expérimentation n'est pas simple à l'échelle agronomique et l'utilisation de modèles de prédiction permet de simuler les performances attendues en pratique. Le modèle STICS [BGJ+03] est largement utilisé en agronomie pour prévoir le rendement de cultures à partir de nombreux paramètres d'entrée. Une branche du logiciel a récemment été développée spécialement pour les cultures mixtes et offre donc un outils pertinent pour l'étude et la promotion de telles cultures. Cependant, l'utilisation de ces modèles de prédiction est délicate à cause du grand nombre de paramètres mis à disposition et du manque de connaissance sur l'impact de chacun d'entre eux sur le résultat. Une étude de sensibilité du modèle STICS dans le cadre des cultures mixtes s'avère nécessaire et doit faire l'objet d'une nouvelle collaboration à venir avec Noémie Gaudio.

La collaboration sur la pollution de plastique décrite dans la section 4.2 fait également partie de mes perspectives de travail à court terme. Afin de mieux comprendre le vieillissement des débris, il est nécessaire de comprendre comment évolue leurs propriétés physiques, en particulier pour le polyéthylène qui compose la majorité de la pollution. Une approche de chimie analytique pour cela consiste à considérer des chromatogrammes obtenus par pyrolyse de fragments de polyéthylène à différents stades de vieillissement. Le résultat correspond à des courbes positives dont la composition peut s'interpréter comme une combinaison de chromatogrammes élémentaires. Une approche statistique basée sur une factorisation matricielle non négative apparaît comme une piste prometteuse sur la base des premières expérimentations. Une régularisation adaptée doit être mise en œuvre afin de rendre les composantes chimiquement interprétables et cela fait l'objet d'une collaboration en cours avec Alexandra ter Halle.

Liste des travaux

- [XG-Article01] X. Gendre. Simultaneous estimation of the mean and the variance in heteroscedastic Gaussian regression. *Electronic Journal of Statistics*, 2, 1345–1372, 2008. 10.1214/08-EJS267.
- [XG-Article02] J. Bigot and X. Gendre. Minimax properties of Fréchet means of discretely sampled curves. *Annals of Statistics*, 41(2), 923–956, 2013. 10.1214/13-AOS1104.
- [XG-Article03] X. Gendre. Model selection and estimation of a component in additive regression. *ESAIM : Probability and Statistics*, 18, 77–116, 2014. 10.1051/PS/2012028.
- [XG-Article04] N. Gaudio, S. Belyazid, X. Gendre, A. Mansat, M. Nicolas, S. Rizzetto, H. Sverdrup, and A. Probst. Combined effect of atmospheric nitrogen deposition and climate change on temperate forest soil biogeochemistry : A modeling approach. *Ecological Modelling*, 306, 24–34, 2015. 10.1016/j.ecolmodel.2014.10.002.
- [XG-Article05] A. ter Halle, L. Ladirat, X. Gendre, D. Goudouneche, C. Pusineri, C. Routaboul, C. Tenailleau, B. Duployer, and E. Perez. Understanding the Fragmentation Pattern of Marine Plastic Debris. *Environmental Science & Technology*, 50(11), 5668–5675, 2016. 10.1021/acs.est.6b00594.
- [XG-Article06] N. Gaudio, X. Gendre, M. Saudreau, V. Seigner, and P. Balandier. Impact of tree canopy on thermal and radiative microclimates in a mixed temperate forest : A new statistical method to analyse hourly temporal dynamics. *Agricultural and Forest Meteorology*, 237, 71–79, 2017. 10.1016/j.agrformet.2017.02.010.

- [XG-Article07] D. Velandia, F. Bachoc, M. Bevilacqua, X. Gendre, and J.-M. Loubes. Maximum likelihood estimation for a bivariate Gaussian process under fixed domain asymptotics. *Electronic Journal of Statistics*, 11(2), 2978–3007, 2017. [10.1214/17-ejs1298](https://doi.org/10.1214/17-ejs1298).
- [XG-Article08] O. Besson, F. Vincent, and X. Gendre. A Stein’s approach to covariance matrix estimation using regularization of Cholesky factor and log-Cholesky metric. *Statistics and Probability Letters*, 167, 2020. [10.1016/j.spl.2020.108893](https://doi.org/10.1016/j.spl.2020.108893).
- [XG-Article09] A. Diet, N. Couellan, X. Gendre, and J. Martin. A Chernov bound for robust tolerance design and application. *The International Journal of Advanced Manufacturing Technology*, 111, 3571–3581, 2020. [10.1007/S00170-020-06231-8](https://doi.org/10.1007/S00170-020-06231-8).
- [XG-Article10] N. Gaudio, C. Violle, X. Gendre, F. Fort, E. Pelzer, S. Médiène, R. Mahmoud, H. Hauggaard-Nielsen, L. Bedoussac, C. Bonnet, G. Corre-Hellou, A. Couëdel, P. Hinsinger, E. Steen Jensen, E.-P. Journet, E. Justes, B. Kammoun, I. Litrico, N. Moutier, C. Naudin, and P. Casadebaig. Interspecific interactions regulate plant reproductive allometry in cereal-legume intercropping systems. *Journal of Applied Ecology*, 58(11), 2579–2589, 2021. [10.1111/1365-2664.13979](https://doi.org/10.1111/1365-2664.13979).
- [XG-Article11] F. Boulfani, X. Gendre, A. Ruiz-Gazen, and M. Salvignol. A statistical approach for sizing an aircraft electrical generator using extreme value theory. *CEAS Aeronautical Journal*, 2021. [10.1007/S13272-021-00540-8](https://doi.org/10.1007/S13272-021-00540-8).
- [XG-Conf1] R. Mahmoud, N. Gaudio, P. Casadebaig, X. Gendre, L. Bedoussac, G. Hellou, F. Fort, E.-P. Journet, I. Litrico, C. Naudin, and C. Violle. A trait-based approach to understand and predict the performance of arable annual mixed crops. In *International Conference on Ecological Sciences*, 2018.
- [XG-Conf2] F. Boulfani, X. Gendre, A. Ruiz-Gazen, and M. Salvignol. Comparing prediction procedures for functional data in aeronautic. In *CMStatistics*, 2019.
- [XG-Conf3] A. Diet, N. Couellan, X. Gendre, J. Martin, and J.-P. Navarro. A statistical approach for tolerancing from design

- stage to measurements analysis. In *Procedia CIRP*, 2020. 10.1016/j.procir.2020.05.171.
- [XG-Conf4] F. Boulfani, X. Gendre, A. Ruiz-Gazen, and M. Salvignol. Anomaly detection for aircraft electrical generator using machine learning in a functional data framework. In *Global Mosharaka Congress on Electrical Engineering*, 2020. 10.23919/GC-ElecEng48342.2020.9285984.
- [XG-Conf5] P. Garnier, C. Jacquy, V. Génot, B. Sanchez-Cano, X. Gendre, C. Mazelle, X. Fang, J. R. Gruesbeck, B. Hall, J. S. Halekas, and B. M. Jakosky. Dynamics of the Martian bow shock location. In *EGU General Assembly*, 2021. 10.5194/egusphere-egu21-9157.
- [XG-Data1] N. Gaudio, C. Violle, X. Gendre, F. Fort, E. Pelzer, S. Médiène, R. Mahmoud, H. Hauggaard-Nielsen, L. Bedousac, C. Bonnet, G. Corre-Hellou, A. Couëdel, P. Hinsinger, E. Steen Jensen, E.-P. Journet, E. Justes, B. Kammoun, I. Litrico, N. Moutier, C. Naudin, and P. Casadebaig. Interspecific interactions regulate plant reproductive allometry in cereal-legume intercropping systems. Dryad, Dataset, 2021. 10.5061/dryad.9ghx3ffhv.
- [XG-Preprint1] A. Archimbaud, F. Boulfani, X. Gendre, K. Nordhausen, A. Ruiz-Gazen, and J. Virta. ICS for multivariate functional anomaly detection with applications to predictivemaintenance and quality control. Soumis, 2021.
- [XG-Preprint2] P. Garnier, C. Jacquy, X. Gendre, V. Génot, C. Mazelle, X. Fang, J. Gruesbeck, B. Sánchez-Cano, and J. Halekas. The influence of crustal magnetic fields on the Martian bow shock location : a statistical analysis of Mars Atmosphere and Volatile EvolutioN and Mars Express observations. Soumis, 2021.
- [XG-Preprint3] P. Garnier, C. Jacquy, X. Gendre, V. Génot, C. Mazelle, X. Fang, J. Gruesbeck, B. Sánchez-Cano, and J. Halekas. The drivers of the Martian bow shock location : a statistical analysis of Mars Atmosphere and Volatile EvolutioN and Mars Express observations. Soumis, 2021.

[XG-Thesis] X. Gendre. *Estimation par sélection de modèle en régression hétéroscédastique*. PhD thesis, Université Nice – Sophia Antipolis, 2009. tel-00397608.

Bibliographie

- [ABMS21] R. Arora, P. Bartlett, P. Mianjy, and N. Srebro. Dropout : Explicit Forms and Capacity Control. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*. PMLR, 2021.
- [Aka70] H. Akaike. Statistical predictor identification. *Annals of the Institute for Statistical Mathematics*, 1970.
- [ANRG18] A. Archimbaud, K. Nordhausen, and A. Ruiz-Gazen. ICS for multivariate outlier detection with application to quality control. *Computational Statistics and Data Analysis*, 128, 2018.
- [Bai03] A. Baillo. Total error in a plug-in estimator of level sets. *Statistics and Probability Letters*, 65(4), 2003.
- [Bar00] Y. Baraud. Model selection for regression on a fixed design. *Probability Theory and Related Fields*, 117, 2000.
- [Bar02] Y. Baraud. Model selection for regression on a random design. *ESAIM : Probability and Statistics*, 6, 2002.
- [Beh17] J. Behler. First Principles Neural Network Potentials for Reactive Simulations of Large Molecular and Condensed Systems. *Angewandte Chemie International Edition*, 56(42), 2017.
- [Ber06] J. Bertoin. *Random Fragmentation and Coagulation Processes*. Cambridge University Press, 2006.
- [BGH09] Y. Baraud, C. Giraud, and S. Huet. Gaussian model selection with an unknown variance. *Annals of Statistics*, 37(2), 2009.
- [BGJ+03] N. Brisson, C. Gary, E. Justes, R. Roche, B. Mary, D. Ripoché, D. Zimmer, J. Sierra, P. Bertuzzi, P. Burger, F. Bussière, Y. M. Cabidoche, P. Cellier, P. Debaeke, J. P. Gaudillère, C. Hénault,

- F. Maraux, B. Seguin, and H. Sinoquet. An overview of the crop model stics. *European Journal of Agronomy*, 18(3), 2003.
- [BHL03] Y. Baraud, S. Huet, and B. Laurent. Adaptive tests of linear hypotheses by model selection. *Annals of Statistics*, 31(1), 2003.
- [BLM13] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities : A nonasymptotic theory of independence*. Oxford University Press, Oxford, 2013.
- [BM97] L. Birgé and P. Massart. From model selection to adaptive estimation. *Festschrift for Lucien Lecam : Research Papers in Probability and Statistics*, 1997.
- [BM00] L. Birgé and P. Massart. An adaptive compression algorithm in Besov spaces. *Constructive Approximation*, 16, 2000.
- [BM01] L. Birgé and P. Massart. Gaussian model selection. *Journal of the European Mathematical Society*, 3(3), 2001.
- [Bou21] F. Boulfani. *Caractérisation du comportement des systèmes électriques aéronautiques à partir d'analyses statistiques*. PhD thesis, Université de Toulouse, 2021.
- [BvRMM19] A. S. Bochkarev, A. van Rookeghem, S. Mossa, and N. Mingo. Anharmonic thermodynamics of vacancies using a neural network potential. *Physical Review Materials*, 3(9), 2019.
- [BVV15] M. Bevilacqua, R. Vallejos, and D. Velandia. Assessing the significance of the correlation between the components of a bivariate Gaussian random field. *Environmetrics*, 26, 2015.
- [CHBS05] E. Castillo, A. S. Hadi, N. Balakrishnan, and J. M. Sarabia. *Extreme Value and Related Models with Applications in Engineering and Science*. Wiley, 2005.
- [dHF06] L. de Haan and A. Ferreira. *Extreme Value Theory : An Introduction*. Springer, New York, 2006.
- [Die21] A. Diet. *An end-to-end approach of statistical tolerancing under industrial constraints : Contribution to the design/industrial virtual twin*. PhD thesis, Université de Toulouse, 2021.
- [DJ94] D. L. Donoho and I. M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81, 1994.

- [EEdH19] J.J. Einmahl, J.H. Einmahl, and L. de Haan. Limits to Human Life Span Through Extreme Value Theory. *Journal of the American Statistical Association*, 114(527), 2019.
- [EO87] M. L. Eaton and I. Olkin. Best Equivariant Estimators of a Cholesky Decomposition. *Annals of Statistics*, 15(4), 1987.
- [FM72] J. J. Filliben and J. E. McKinney. Confidence limits for the abscissa of intersection of two linear regressions. *Journal of Research of the National Bureau of Standards*, 76B, 1972.
- [GBC16] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.
- [GM07] U. Grenander and M. I. Miller. *Pattern Theory : From Representation to Inference*. Oxford University Press, Oxford, 2007.
- [GN00] A. K. Gupta and D. K. Nagar. *Matrix variate distributions*. Chapman and Hall/CRC, 2000.
- [GP05] L. Galtchouk and S. Pergamenshchikov. Efficient adaptive nonparametric estimation in heteroscedastic regression models. Université Louis Pasteur, IRMA, Preprint, 2005.
- [HMSW04] W. K. Härdle, M. Müller, S. Sperlich, and A. Werwatz. *Non-parametric and Semiparametric Models*. Springer, Berlin, Heidelberg, 2004.
- [HOS⁺19] B. Heling, T. Oberleiter, B. Schleich, K. Willner, and S. Wartzack. On the Selection of Sensitivity Analysis Methods in the Context of Tolerance Management. *Journal of Verification, Validation and Uncertainty Quantification*, 4(1), 2019.
- [HSK⁺12] G. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. arXiv :1207.0580, 2012.
- [HTF09] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning : Data Mining, Inference, and Prediction*. Springer, New York, 2009.
- [IR78] I. A. Ibragimov and Y. A. Rozanov. *Gaussian Random Processes*. Springer-Verlag, New York, 1978.

- [JS61] W. James and C. Stein. Estimation with Quadratic Loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1. University of California Press, Berkeley, 1961.
- [KVC01] F. Killmann and E. Von Collani. A note on the convolution of the uniform and related distributions and their use in quality control. *Economic Quality Control*, 16(1), 2001.
- [Leo47] W. Leontief. Introduction to a theory of the internal structure of functional relationships. *Econometrica*, 15, 1947.
- [LH97] S. C. Liu and S. J. Hu. Variation Simulation for Deformable Sheet Metal Assemblies Using Finite Element Methods. *Journal of Manufacturing Science and Engineering*, 119(3), 1997.
- [Lin19] Z. Lin. Riemannian Geometry of Symmetric Positive Definite Matrices via Cholesky Decomposition. *SIAM Journal on Matrix Analysis and Applications*, 40(4), 2019.
- [Mal73] C. L. Mallows. Some comments on c_p . *Technometrics*, 15, 1973.
- [Mas07] P. Massart. *Concentration Inequalities and Model Selection. Lecture Notes in Math*, volume 1896. Springer, Berlin, 2007.
- [MAV18] P. Mianjy, R. Arora, and R. Vidal. On the Implicit Bias of Dropout. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*. PMLR, 2018.
- [MJS12] T. Ma, L. Jia, and Y. Su. A new estimator of covariance matrix. *Journal of Statistical Planning and Inference*, 142(2), 2012.
- [Nes04] Y. Nesterov. *Introductory Lectures on Convex Optimization : A Basic Course*. Springer, Boston, 2004.
- [Pol95] W. Polonik. Measuring Mass Concentrations and Estimating Density Contour Clusters-An Excess Mass Approach. *Annals of Statistics*, 23(3), 1995.
- [RS02] J. O. Ramsay and B. W. Silverman. *Applied Functional Data Analysis : Methods and Case Studies*. Springer, New York, 2002.
- [RS05] J. O. Ramsay and B. W. Silverman. *Functional Data Analysis*. Springer, New York, 2005.

- [Sch59] H. Scheffé. *The analysis of variance*. Wiley Interscience, 1959.
- [She95] Y. Sheena. Unbiased estimator of risk for an orthogonally invariant estimator of a covariance matrix. *Journal of the Japan Statistical Society*, 25(1), 1995.
- [SHK⁺14] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout : A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15(1), 2014.
- [Ste99] M. L. Stein. *Interpolation of Spatial Data*. Springer-Verlag, New York, 1999.
- [Sto85] C. J. Stone. Additive regression and other nonparametric models. *Annals of Statistics*, 14(2), 1985.
- [Tsu16] H. Tsukuma. Estimation of a high-dimensional covariance matrix with the Stein loss. *Journal of Multivariate Analysis*, 148, 2016.
- [TW13] S. Taskinen and D. I. Warton. Robust tests for one or more allometric lines. *Journal of Theoretical Biology*, 333, 2013.
- [WBCL08] L. Wang, L. D. Brown, T. Cai, and M. Levine. Effect of mean on variance function estimation in nonparametric regression. *Annals of Statistics*, 36(2), 2008.
- [WSSB05] P. Wallman, M. Svensson, H. Sverdrup, and S. Belyazid. ForSAFE – an integrated process-oriented forest model for long-term sustainability assessments. *Forest Ecology and Management*, 207(1), 2005.
- [WWFW06] D. I. Warton, I. J. Wright, D. S. Falster, and M. Westoby. Bivariate line-fitting methods for allometry. *Biological Reviews*, 81(2), 2006.
- [Yin91] Z. Ying. Asymptotic properties of a maximum likelihood estimator with data from a Gaussian process. *Journal of Multivariate Analysis*, 36, 1991.