HDR

En vue de l'obtention de l'

HABILITATION À DIRIGER DES RECHERCHES

Délivré par : *l'Université Toulouse 3 Paul Sabatier (UT3 Paul Sabatier)*

Présentée et soutenue le *11/10/2019* **par :**

AGNÈS LAGNOUX

Statistique et excursions de processus, grandes déviations conditionnelles et analyse de sensibilité

JURY

BERCU BERNARD	Université Bordeaux 1
GAMBOA FABRICE	Université de Toulouse
GANTERT NINA	Technische Universität München
GARNIER JOSSELIN	École Polytechnique
GREMAUD PIERRE	NC State University
LAURENT BÉATRICE	INSA de Toulouse

École doctorale et spécialité :
MITT : Domaine Mathématiques : Mathématiques appliquées
Unité de Recherche :
Institut de Mathématiques de Toulouse (UMR 5219)
Parrain :
GAMBOA Fabrice
Rapporteurs :
GANTERT Nina, *GARNIER Josselin* et *GREMAUD Pierre*

Remerciements

Tout d'abord, j'adresse un grand merci à Fabrice qui m'a sollicitée pour participer à de nombreux projets de recherche et qui a bien voulu parrainer cette habilitation. Je souhaite à nouveau remercier Dominique Bakry et Pascal Lezaud pour m'avoir incitée à découvrir le monde de la recherche et accompagnée dans mes premiers pas.

Je suis très honorée que Nina Gantert, Josselin Garnier et Pierre Gremaud aient accepté d'être rapporteurs. Je les remercie pour leur lecture attentive du manuscrit. C'est également un grand plaisir que Bernard Bercu et Béatrice Laurent-Bonneau fassent partie de mon jury. Qu'ils soient tous remerciés d'avoir fait le déplacement (plus ou moins long) pour ma soutenance.

Une pensée chaleureuse se dirige vers mes co-auteurs avec lesquels j'ai pris plaisir à travailler. Ils m'ont beaucoup appris. Qu'ils en soient sincèrement remerciés. Une autre pensée concerne les collègues de l'Institut de Mathématiques et du département Math-Info de l'Université Toulouse Jean-Jaurès que j'ai plaisir à cotoyer.

Un mot enfin pour exprimer ma gratitude envers mes amis, ma famille et en particulier mes parents. Un immense merci à toi Éric et à mes trois poussins : Adèle, Élise et Victor.

Résumé

Nous donnons un aperçu de nos contributions principales obtenues depuis 2008 dans le cadre de nos recherches à l'Institut de Mathématiques de Toulouse (Maître de conférence à l'Université Toulouse 2 Jean-Jaurès). Nous discutons aussi en fin de manuscrit de nos activités de recherche en lien avec le milieu industriel et en particulier de nos activités de recherche appliquées avec EDF en Post-Doctorat au sein de l'Institut de Mathématiques de Toulouse (IMT).

Mots clefs par thématiques de recherche

Branchement avec duplication des trajectoires, estimation de la probabilité d'événements rares, Monte Carlo, processus de ramification, simulation, réduction de la variance, densité du premier temps d'atteinte, cartes conformes.
Processus gaussiens, krigeage, formules un à part (Leave-One-Out formulas), validation croisée, variations, processus gaussiens contraints, consistance forte, normalité asymptotique, échantillonnage spatial, asymptotique à domaine fixe.
Excursions Browniennes, mouvement Brownien réfléchi, processus de Lindley, théorème d'invariance de Donsker, score local, analyse de séquence biologique.
Processus autorégressif, statistique de Bickel-Rosenblatt, test d'ajustement, test d'hypothèses, estimation non paramétrique, estimateur de densité de Parzen-Rosenblatt, processus résiduel.
Principes de grandes déviations et de déviations modérées, inégalité de Berry-Esseen, queues lourdes, distribution conditionnelle, problèmes combinatoires, hachage avec essai linéaire.
Analyse de sensibilité, décomposition de Hoeffding, indices de Sobol, distance de Cramér-von Mises, plan d'échantillonnage Pick-Freeze, méthode (fonctionnelle) Delta, inégalités de concentration, consistance forte, normalité asymptotique, efficacité asymptotique, intervalles de confiance, métamodélisation.

Plan du manuscrit

Ce document s'articule en trois parties principales indépendantes les unes des autres. Après une rapide introduction dans le Chapitre 1, je présente les résultats concernant les processus dans le Chapitre 2. En particulier, mes travaux de thèse et leur suite sont résumés dans la Section 2.1 tandis que la Section 2.2 est dédiée à l'estimation de paramètres de fonction de covariance pour des processus gaussiens. S'ensuit la Section 2.3 dans laquelle je présente quelques résultats sur le maximum d'excursions browniennes pour des applications en biologie. Le Chapitre 2 se termine par la Section 2.4 résumant des résultats concernant les processus autorégessifs et la statistique de Bickel-Rosenblatt. Le Chapitre 3 est consacré aux principes de grandes déviations tandis que le Chapitre 4 concerne l'analyse de sensibilité. Enfin, j'évoque rapidement dans le Chapitre 5 mes travaux plus appliqués en lien avec l'industrie. Je conclue par le Chapitre 6 en présentant mes travaux de recherche en cours et quelques perspectives. Une liste des notations suit ce dernier chapitre et précède la bibliographie.

Liste des publications

Mes publications sont disponibles sur ma page web personnelle :

https://perso.math.univ-toulouse.fr/lagnoux/

Articles publiés

Revues à comité de lecture

[J19] F. Bachoc, A. Lagnoux, A. F. López-Lopera. Maximum likelihood estimation for Gaussian processes under inequality constraints. **Accepté EJS** (2019). hal-01772560.

[J18] G. Sarrazin, J. Morio, A. Lagnoux, M. Balesdent, L. Brevault. Sensitivity Analysis of Risk Assessment with Data-Driven Dependence Modeling. **Accepté ESREL** (2019).

[J17] T. Klein, A. Lagnoux, P. Petit. A conditional Berry-Esseen inequality. **Journal of Applied Probability** (2019). Vol. 56, no. 1, pp. 76-90. hal-01801795v1.

[J16] A. Lagnoux, S. Mercier, P. Vallois. Probability density of the local score position. **Stochastic Processes and their Applications** (2018). Available online. hal-01835781.

[J15] A. Lagnoux, T.M.N Nguyen, F. Proïa. On the Bickel-Rosenblatt test of goodness-of-fit for the residuals of autoregressive processes. **ESAIM Probability & Statistics** (2018). hal-01551093.

[J14] F. Gamboa, T. Klein, A. Lagnoux. Sensitivity analysis based on Cramér-von Mises distance. **SIAM JUQ** (2017). hal-01163393.

[J13] F. Bachoc, A. Lagnoux, T.M.N. Nguyen. Cross-validation estimation of covariance parameters under fixed-domain asymptotics. **Journal of Multivariate Analysis** (2017). Vol. 160, pp. 42-67. hal-01377854.

[J12] A. Lagnoux, P. Lezaud. Multilevel branching splitting algorithm for estimating rare event proba. **Simulation Modelling Practice and Theory** (2017). Vol. 72, pp. 150-167. hal-01269766.

[J11] A. Lagnoux, S. Mercier, P. Vallois. Statistical significance based on the length and position of the local score of i.i.d. sequences. **Bioinformatics** (2017). Vol. 33, no. 5, pp. 654-660. hal-01301246.

[J10] F. Gamboa, A. Janon, T. Klein, A. Lagnoux, C. Prieur. Statistical inference for Sobol pick freeze Monte Carlo method. **Statistics** (2015). hal-00804668.

[J9] A. Lagnoux, S. Mercier, P. Vallois. Probability that the maximum of the reflected Brownian motion over a finite interval [0 ; t] is achieved by its last zero before t. **Electronic Communications in Probability** (2015). Vol. 20, no. 62, pp. 1-9. hal-01214773.

[J8] C. Chabriac, A. Lagnoux, S. Mercier, P. Vallois. Elements related to the largest complete excursion of a reflected BM stopped at a fixed time. Application to local score. **Stochastic Processes and their Applications** (2014). Vol. 124, no. 12, pp. 4202-4223. hal-00857402.

[J7] F. Gamboa, A. Janon, T. Klein, A. Lagnoux. Sensitivity analysis for multidimensional and functional outputs. **Elect. J. Stat** (2014). Vol. 8, no. 1, pp. 575–603. hal-00881112.

[J6] J .C. Fort, T. Klein, A. Lagnoux, B. Laurent. Estimation of the Sobol indices in a linear functional multidimensional model. **Journal of Statistical Planning and Inference** (2013). Vol. 143, no. 9, pp. 1590-1605. hal-00685998.

[J5] A. Janon, T. Klein, A. Lagnoux, M. Nodet, C. Prieur. Asymptotic normality and efficiency of a Sobol index estimator. **ESAIM Probability & Statistics** (2013). Online publication. hal-00665048.

[J4] J.M. Azaïs, S. Bercu, J.C. Fort, A. Lagnoux-Renaudie, P. Lé. Simultaneous confidence bands in curve prediction. **JRSSC** (2010). Vol. 59, pp. 889-904. `hal-00644155`.

[J3] A. Lagnoux-Renaudie. A two-step branching splitting model under cost constraint. **Journal of Applied Probability** (2009). Vol. 46, no. 2, pp. 429-452. `hal-00644145`.

[J2] A. Lagnoux-Renaudie. Effective splitting model under cost constraint. **Stochastic Processes and their Appl.** (2008). Vol. 118, no.10, pp. 1820-1851. `hal-00644141`.

[J1] A. Lagnoux . Rare event simulation. **Probability in the Engineering and the Informational Sciences** (2006). Vol. 20, no. 1, pp. 45-66. `hal-00644139`.

Notes et proceedings

[N2] F. Gamboa, A. Janon, T. Klein, A. Lagnoux. Sensitivity indices for multivariate outputs. **C.R.A.S.** (2013). Vol. 351, no. 7-8, pp. 307-310. `hal-00800847`.

[N1] An adaptive branching splitting model under cost constraint for rare event analysis. A. Lagnoux. **Proceedings of 6th St. Petersburg Workshop on Simulation** (2009). pp. 721-724 (extended abstract).

Articles soumis

[S3] F. Bachoc, A. Lagnoux. Fixed-domain asymptotic properties of composite likelihood estimators for Gaussian processes. **En révision** (2019). `hal-02079975`.

[S2] B. Demory, M. Henner, A. Lagnoux, T.M.N. Nguyen. Expected Improvement applied to an industrial context - Prediction of new geometries increasing the efficiency of fans. **En révision** (2019). `hal-02044258`.

[S1] J.M. Azais, F. Bachoc, T. Klein, A. Lagnoux, T.M.N. Nguyen. Semi-parametric estimation of the variogram of a Gaussian process with stationary increments. **Soumis** (2019). `hal-01802830`.

Travaux en cours de rédaction

[P5] F. Brosset, F. Barthe, T. Klein, A. Lagnoux, P. Petit. Large deviations at logarithmic scale for sums of heavy-tailed random variables. **En cours de rédaction** (2019).

[P4] T. Klein, A. Lagnoux, P. Petit. Deviations results for hashing with linear probing. **En cours de rédaction** (2019).

[P3] M. Buisson, A. Lagnoux, T.M.N. Nguyen, M. Ribaud. Upper confidence bound and expected improvement in an industrial context. **En cours de rédaction** (2019).

[P2] N. Bousquet, E. Chassot, S. Da Veiga, B. Iooss, A. Lagnoux. Calibration, sensitivity analysis and classification in a biodynamical model. Application to the Indian Ocean Yellowfin DEB model. **En cours de rédaction** (2019).

[P1] F. Gamboa, T. Klein, A. Lagnoux, L. Moreno. Sensitivity analysis in general metric spaces. **En cours de rédaction** (2019). `hal-02044223`.

Rapports techniques

[R2] B. Iooss, T. Klein, A. Lagnoux. Stochastic numerical codes and sensitivity analysis. **Rapport technique - Projet NEEDS-ASINCRONE** (2014). 7 pages.

[R1] J.-M. Azaïs, S. Gadat, A. Lagnoux, C. Mercadier. Méthode d'événements rares pour l'intégrité et la continuité EGNOS - GALILEO. **Rapport technique - Projet THALES-CNES** (2011). 30 pages.

Manuscrit de thèse

[T1] A. Lagnoux-Renaudie. Analyse des modeles de branchement avec duplication des trajectoires pour l'étude des événements rares. **Manuscrit de thèse** (2006).`tel-00129752v1`.

Dans les listes précédentes, les références [J1], [J2], [J3] et [T1] correspondent à mes travaux de thèse et la référence [J4] correspond aux résultats obtenus pendant mon PostDoc. Toutes les autres références correspondent à des activités de recherche menées au sein de l'Institut de Mathématiques de Toulouse depuis mon recrutement à l'Université Toulouse 2 Jean Jaurès.

Table des matières

Chapitre 1

Introduction

Les activités de recherche dont je présente les contributions dans ce manuscrit ont été guidées par de nombreuses recontres. Dans cette section, j'en donne la chronologie (qui n'est pas celle du manuscrit) ainsi que les diverses thématiques abordées.

Mon initiation à la recherche a été rendue possible et encouragée par Dominique Bakry (IMT) que j'ai eu la chance et le plaisir d'avoir comme enseignant dans l'U.E. Modèles Stochastiques en Maîtrise (non sans difficultés !). Il m'a poussée à faire un DEA puis une thèse qu'il a accepté de co-encadrer avec Pascal Lezaud (IMT-ENAC). Le sujet portait sur l'estimation de la probabilté d'événements rares par des schémas de type Monte Carlo accélérés. Après ma thèse, j'ai eu le plaisir de continuer à travailler dans cette direction avec Pascal Lezaud et de clore en quelque sorte le sujet, du moins les questions auxquelles je m'étais intéressée. Les résultats de cette collaboration, publiés dans [J12], sont présentés dans la Section 2.1 du Chapitre 2, qui décrit également mes travaux de thèse.

Après ma thèse, j'ai travaillé sur des problématiques industrielles de statistique appliquée dans le cadre de mon Post-Doctorat à l'IMT sur un contrat EDF, co-encadré par Jean-Marc Azaïs (IMT) et Jean-Claude Fort (Université Paris V et à l'époque IMT). Le sujet portait sur la prédiction de courbes de charge. La Section 5.1 de ce manuscrit donne un aperçu du travail mené qui s'est concrétisé par l'article [J4].

Dès mon arrivée à l'Université Toulouse 2 Jean Jaurés en tant que Maître de COnférence, Sabine Mercier et Claudie Chabriac qui entamaient une collaboration sur l'analyse théorique de séquences biologiques avec Pierre Vallois (Université de Lorraine) m'ont proposé d'y participer. Nous avons établi quelques propriétés sur les excursions browniennes. Chacun a apporté ses connaissances et savoir-faire pour aboutir à une collaboration enrichissante et féconde donnant lieu à quatre articles, trois théoriques [J8], [J9] et [J16], le dernier plus appliqué et à destination des biologistes [J11]. Ce travail est présenté en Section 2.3 du Chapitre 2.

Parallèlement à ces activités de recherche en Probabilité, Fabrice Gamboa porteur du projet ANR Costa Brava, dont le thème était l'analyse de sensibilité, m'a proposé d'y participer. Cela a été l'occasion d'approfondir mes connaissances en statistique, de travailler avec plusieurs membres de l'IMT (Jean-Claude Fort, Fabrice Gamboa, Thierry Klein, Béatrice Laurent-Bonneau (IMT-INSA)), de rencontrer d'autres chercheurs académiques dont Clémentine Prieur (LJK) et Alexandre Janon (Université Paris Sud) avec lequel je continue d'interagir. De plus, cela m'a permis de découvrir un peu plus le monde de l'industrie avec des ingénieurs du CEA et de l'IFP dont Sébastien Da Veiga (Safran) et Amandine Marrel (CEA)... Nous nous sommes intéressés aux propriétés théoriques des indices de Sobol puis à leurs généralisations. Ce projet ANR m'a permis de rebondir dès le début de ma carrière d'enseigant-chercheur

sur un nouveau sujet de recherche. Le Chapitre 4 est consacré aux résultats obtenus. Ils ont fait l'objet des publications [N2], [J5], [J6], [J7], [J10] et [J14] et de la prépublication [P1].

Suite au recrutement de Pierre Petit à l'IMT en 2011, Thierry Klein (IMT-ENAC) nous a proposé de travailler sur un sujet qu'il connaissait bien : le hachage avec essais linéaires et plus généralement aux sommes de variables aléatoires conditionnées. Dans le contexte du hachage, les queues de distributions sont lourdes et seul un théorème central limite avait été établi. Il s'agissait alors de montrer un principe de grandes déviations. La première étape de ce travail de longue haleine a été de prouver un résultat de type Berry-Esseen pour les sommes de variables aléatoires conditionnées. Après de nombreuses séances de travail studieuses et laborieuses, nous avons finalement réussi à établir un catalogue complet du comportement asymptotique de la variable d'intérêt dans le cadre du hachage avec essais linéaires. Ces résultats sont présentés dans le Chapitre 3. Ils ont fait l'objet de l'article [J17] et des prépublications [P4] et [P5].

De nombreuses pistes de recherche ont été envisagées et explorées avec et grâce à Thi Mong Ngoc (Jade) Nguyen (Université de Sciences d'Ho Chi Minh, Vietnam) lors de sa venue en France à deux reprises : sur un support de professeur invité à l'Université Toulouse 2 Jean Jaurès en juin 2015 puis pour un Post-Doctorat de 18 mois à partir de février 2016 dans le cadre de l'ANR Pepito. Nous avons ainsi eu l'occasion d'étudier les processus déterminantaux, de découvrir le krigeage et l'amélioration attendue (Expected Improvement) que nous avons ensuite mis à profit pour proposer aux ingénieurs Valeo, membres de l'ANR, des géométries de ventilateurs prometteuses et inédites. Les techniques de borne de confiance supérieure (Upper Bound Confidence) ont aussi été exploitées et comparées à l'amélioration attendue. Les prépublications [S2] et [P3] ainsi que la Section 5.4 présentent les résultats obtenus. Parallèlement, Bernard Bercu, directeur de thèse de Jade, l'a encouragée à étudier dans le prolongement de sa thèse la statistique de Bickel-Rosenblatt. En collaboration avec Frédéric Proïa (Université d'Angers), autre doctorant de Bernard, nous avons ainsi écrit l'article [J15] dont les résultats sont présentés en Section 2.4. Enfin, nous avons aussi travaillé avec François Bachoc, tout juste recruté à l'IMT, sur les estimateurs par validation croisée des paramètres des fonctions de covariance de processus gaussiens. Voir la Section 2.2 du Chapitre 2 et la référence [J13].

Celle collaboration fructueuse avec François s'est poursuivie jusqu'à maintenant. En particulier, avec Andrés López-Lopera (Doctorant Mines de Saint Etienne), nous avons étudié le comportement asymptotique de l'estimateur par maximum de vraisemblance des paramètres des fonctions de covariance de processus gaussiens contraints par des conditions d'inégalités [S19]. Avec Jean-Marc Azaïs, Thierry Klein et Thi Mong Ngoc Nguyen, nous nous sommes intéressés à des estimateurs par variations pour des processus gaussiens à accroissements stationnaires [S1]. Nous avons enfin étudié le comportement asymptotique des estimateurs par vraisemblance composite en tant qu'alternative aux estimateurs par maximum de vraisemblance [S3]. L'ensemble de ces résultats sont résumés dans la Section 2.2 du Chapitre 2.

Actuellement, je continue à travailler avec François Bachoc sur les processus gaussiens, Pierre Petit et Thierry Klein sur la thématique des grandes déviations, Fabrice Gamboa et Thierry Klein en analyse de sensibilité... Ces activités de recherche se trouvent renforcées et enrichies par le co-encadrement de deux thèses : celle de Gabriel Sarazin commencée en octobre 2017, co-encadrée avec Jérôme Morio (ONERA), à l'interface entre analyse de sensibilté et estimation de la probabilité d'événements rares, et celle de Fabien Brosset démarrée en février 2017, co-encadrée avec Franck Barthe (IMT) et concernant des problématiques de grandes déviations. Dans le Chapitre 6, je présente mes travaux en cours et mes perspectives scientifiques.

Chapitre 2

Processus aléatoires : branchement avec duplication de trajectoires, estimation de covariance et excursions

Ce chapitre s'articule principalement en trois parties : l'une, algorithmique, concerne les travaux initiés pendant ma thèse, une autre, statistique, étudie l'estimation des paramètres de la fonction de covariance de processus gaussiens ; enfin la dernière, probabiliste, traite du maximum des excursions browniennes et autres quantités relatives.

2.1 Processus de branchement et estimation de la probabilité d'événements rares

Dans de nombreux domaines appliqués, l'estimation de la probabilité d'occurrence d'un événement rare est une question cruciale (souvent en raison du risque associé à cet événement) ; cet événement pouvant être par exemple une défaillance catastrophique en fiabilité des systèmes, un risque de collision entre avions, de tremblement de terre... Les événements rares sont caractérisés par des probabilités de l'ordre de 10^{-9} à 10^{-12}. Par exemple, en télécommunication pour certains protocoles ou liaisons, la probabilité de perte d'un paquet d'informations est inférieure à 10^{-9}.
Diverses voies se présentent pour l'étude de ces risques :

- l'analyse statistique des événements extrêmes qui s'appuie principalement sur les lois asymptotiques des extrêmes (Weibull, Fréchet, Gumbel) mais nécessite une longue période d'observation [4] ;
- la modélisation qui conduit à estimer la probabilité de l'événement rare, soit par une approche analytique [223], soit par la simulation.

Dans ma thèse, j'ai abordé l'aspect simulation basé sur la méthode de Monte-Carlo qui s'appuie sur la loi forte des grands nombres. Cependant, cette méthode de simulation s'avère inefficace pour estimer un événement de probabilité 10^{-9} à 10^{-12}, en raison du nombre trop important d'échantillons à générer. Pour résoudre ce problème, de nombreuses méthodes de simulation accélérée ont été proposées telles que l'échantillonnage préférentiel ("Importance Sampling" (IS)) et la méthode RESTART [253, 254]. Une

approche particulaire [54, 85] permet d'obtenir des résultats théoriques sur la convergence de ce type d'algorithme.
Nous nous intéressons à l'efficacité réelle de ces méthodes à multi-trajectoires préférentielles en terme de coût de simulation. En raison de leur complexité, l'analyse mathématique directe s'avère impraticable. Cependant, une étude analytique rapide montre que ces méthodes reposent tous sur un même "squelette", sorte de modèle simplifié pour lequel nous avons mené une étude précise et obtenu des résultats exacts et non asymptotiques.
Je présente succinctement dans la Section 2.1.1 les résultats obtenus dans ma thèse sous la direction de Dominique Bakry (IMT) et Pascal Lezaud (IMT-ENAC) et publiés dans [J1], [J2] et [J3]. La Section 2.1.2 est consacrée à une analyse menée ultérieurement avec P. Lezaud en dimension supérieure à un conduisant à l'article [J12].

2.1.1 La méthode de branchement avec duplication des trajectoires en dimension 1

Le but est d'estimer la probabilité $\mathbb{P}(A)$ d'un événement rare A correspondant par exemple au dépassement d'un certain niveau L par un processus $Y(t)$. Contrairement aux algorithmes d'échantillonnage préférentiel [155], dans les algorithmes multi-trajectoires préférentiels, le système évolue selon la mesure de probabilité initiale. Cette technique s'appuie sur l'hypothèse qu'il existe des états intermédiaires identifiables visités par le système plus souvent que l'événement rare lui-même : nous définissons donc une suite de $M+1$ ensembles B_i emboîtés

$$A = B_{M+1} \subset B_M \subset \ldots \subset B_2 \subset B_1$$

où B_{M+1} correspond à A. On obtient de la sorte une partition de l'espace d'états en régions $B_i \backslash B_{i+1}$ appelées *régions d'importance*.
Dans ce cas, nous avons la formule produit

$$\mathbb{P}(A) = \mathbb{P}(A|B_M)\mathbb{P}(B_M|B_{M-1}) \ldots \mathbb{P}(B_2|B_1)\mathbb{P}(B_1). \tag{2.1}$$

La probabilité d'intérêt $\mathbb{P}(A)$ s'écrit ainsi comme le produit de $M+1$ quantités (probabilités conditionnelles) "moins rares" et donc plus faciles à estimer et avec plus de précision que $\mathbb{P}(A)$ elle-même, pour un coût de simulation donné.
Dans cette méthode, appelé *méthode de branchement avec duplication des trajectoires*, une apparition plus fréquente de A est réalisée en dupliquant le processus en R_i sous-processus dès qu'il entre dans une région B_i où sa chance d'atteindre l'événement rare est plus grande. On privilégie ainsi les trajectoires favorables. Un principe implicite nécessaire est que le niveau B_{i+1} ne peut être atteint depuis B_{i-1} sans que ne le soit aussi B_i. Nous supposons naturellement que Y évolue continuement et que tous les seuils intermédiaires sont atteints si le dernier seuil l'est. En fait, la dynamique qui nous intéresse n'est pas directement celle des particules, mais plutôt celle de la chaîne de Markov sous-jacente observée à chaque fois que la particule atteint un niveau intermédiaire B_k. Cette chaîne de Markov sous-jacente sera notée $(X_k)_{0 \leqslant k \leqslant M+1}$. Plus précisément, dans ce modèle et contrairement aux algorithmes particulaires, nous ne tenons pas compte de l'évolution du processus entre les niveaux mais seulement du franchissement des niveaux par ce dernier.
Précisons que nous ne considèrerons, dans cette section, que des modèles uni-dimensionnels comme ceux introduits par Garvels [103] et dans la méthode RESTART [254]. L'idée de base consiste alors à faire un tirage de Bernoulli et vérifier si B_1 est atteint ou non. Si c'est le cas, on scinde le tirage en R_1 tirages de Bernoulli $\{0,1\}$ et on vérifie si B_2 est atteint ou non, et ainsi de suite. Plus précisément, chaque fois

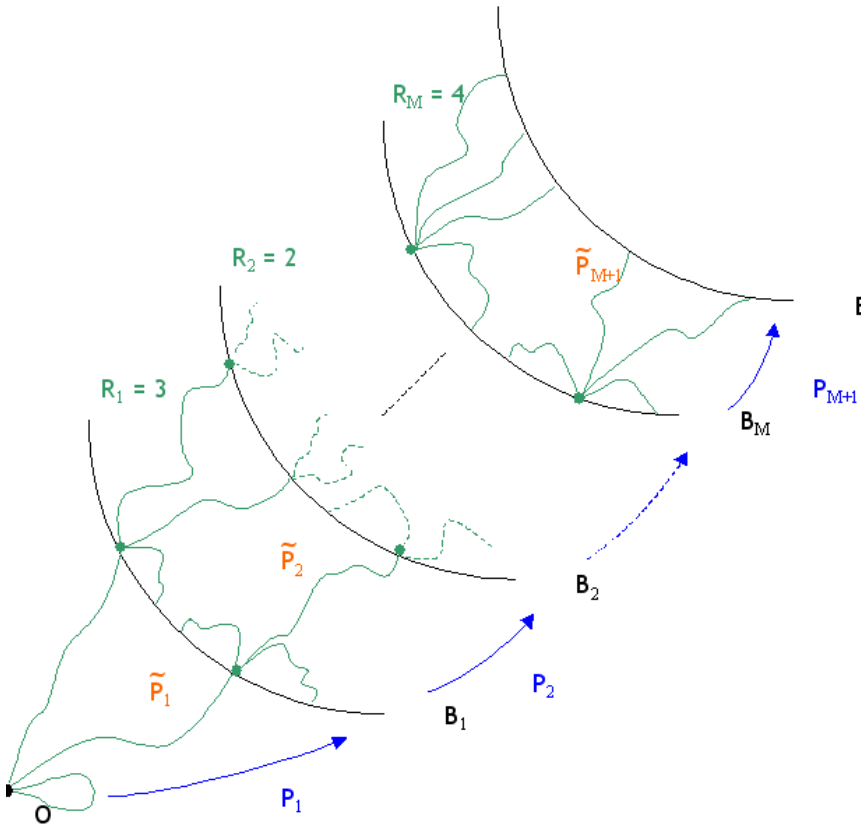


FIGURE 2.1 – Branchement avec duplication des trajectoires.

que l'événement B_i se produit, on simule R_i tirages et on renouvelle cette technique chaque fois que l'événement B_{i+1} se produit. Si aucun seuil supérieur, ni A, n'est atteint, on arrête le tirage concerné. En itérant de façon indépendante N fois cette procédure, on aura considéré $NR_1 \dots R_M$ tirages, en tenant compte du fait que par exemple, si un niveau B_i à la i-ième étape n'a pas été atteint, les $R_i \dots R_M$ tirages possibles ont échoué.

Les paramètres inconnus de l'algorithme sont donc le nombre N de particules envoyées au départ, les nombres $R_1, \dots, R_M$ de retirage, les probabilités $P_1, \dots, P_{M+1}$ de transition et le nombre M de niveaux.

Un estimateur sans biais de $\mathbb{P}(A)$ est naturellement donné par

$$\widehat{P} = \frac{Z_1}{N} \times \left(\prod_{n=1}^{M-1} \frac{Z_{n+1}}{R_n Z_n} \right) \times \frac{Z_{M+1}}{R_M Z_M} = \frac{Z_{M+1}}{N R_1 \dots R_M}. \tag{2.2}$$

où Z_n représente le nombre de particules ayant atteint le niveau B_n. Notons que cet algorithme peut être représenté par N processus de branchement indépendants. Introduisons les quantités suivantes

$$m_0 = P_1, \qquad m_n = R_n P_{n+1}, \quad n = 1, \dots, M+1,$$

nombres moyens de particules réussissant à atteindre le niveau supérieur, ainsi que

$$\begin{aligned} r_0 &= 1, \qquad r_n = R_1 \dots R_n, \quad n = 1, \dots, M, \\ p_0 &= 1, \qquad p_n = P_1 \dots P_n, \quad n = 1, \dots, M+1. \end{aligned}$$

Comme fait dans [254], on déduit l'expression de la variance de l'estimateur $\widehat{P}$ par récurrence.

Proposition 2.1 (Variance de l'estimateur en dimension 1)**.** *En utilisant les notations précédemment introduites, il vient*

$$\mathrm{Var}(\widehat{P}) = \frac{\mathbb{P}(A)^2}{N}\left[\sum_{i=0}^{M}\frac{1}{r_i}\left(\frac{1}{p_{i+1}}-\frac{1}{p_i}\right)\right]. \tag{2.3}$$

Il s'agit maintenant de décrire le coût d'une simulation : chaque fois qu'une particule est générée, cela entraîne un coût moyen unitaire décrit par une fonction c positive. On suppose que

— le coût c d'une particule pour atteindre B_i partant de B_{i-1} dépend seulement de P_i et non du niveau de départ ni du point de départ (nous travaillons en dimension 1) ;

— c est décroissante en x (ce qui signifie que plus la probabilité de transition est petite, plus difficile est la transition et par conséquent plus le coût est élevé) ;

— c converge vers une constante (en général petite) quand x tend vers 1.

Proposition 2.2 (Coût de l'algorithme en dimension 1)**.** *En utilisant les notations précédentes, il vient que le coût moyen de l'algorithme est donné par*

$$C = N\sum_{i=0}^{M} c(P_{i+1})r_i p_i \tag{2.4}$$

Pour minimiser la variance de $\widehat{P}$ à coût de simulation fixé, nous procédons en trois étapes.

1) On optimise la variance en $N, R_1, \ldots, R_M$ à $P_1, \ldots, P_{M+1}$ fixés (*i.e.* aux niveaux B_i fixés) et on obtient

$$R_i = \frac{r_i}{r_{i-1}} = \sqrt{\frac{c(P_i)}{c(P_{i+1})}}\sqrt{\frac{1}{P_i P_{i+1}}}\sqrt{\frac{1-P_{i+1}}{1-P_i}} \quad i = 1, \ldots, M, \tag{2.5}$$

$$N = \frac{1}{\sqrt{c(P_1)}}\frac{C\sqrt{1/P_1 - 1}}{\sum_{i=1}^{M+1}\sqrt{c(P_i)}\sqrt{\frac{1}{P_i}-1}}. \tag{2.6}$$

2) On injecte ces valeurs optimales dans la variance et on déduit les valeurs optimales des P_i pour $i = 1 \ldots M+1$ sous la contrainte $\mathbb{P}(A) = P_1 \ldots P_{M+1}$:

$$P_i = \mathbb{P}(A)^{\frac{1}{M+1}} \quad i = 1, \ldots, M+1. \tag{2.7}$$

3) On injecte ces valeurs optimales dans la variance et on déduit la valeur optimale du nombre de seuils M : $M = [\frac{\ln \mathbb{P}(A)}{y_0}] - 1$ ou $M = [\frac{\ln \mathbb{P}(A)}{y_0}]$ avec y_0 solution de $F(y) = 0$ où

$$F(y) := (2(1-e^y)+y)c(e^y) - y(1-e^y)e^y c'(e^y) = 0, \text{ avec } y = \frac{\ln \mathbb{P}(A)}{M+1}. \tag{2.8}$$

Notons que M croît lorsque $\mathbb{P}(A)$ décroît et avec cette valeur de M, nous obtenons

$$R_i \approx 5 \text{ et } P_i \approx \frac{1}{5}. \tag{2.9}$$

Ainsi les nombres optimaux de retirage et les probabilités de transition optimales sont "indépendants" de la probabilité de l'événement rare.

Remarques sur l'algorithme optimal

Tout d'abord, notons que les valeurs optimales pour les $(R_i)_i : R_i = \frac{1}{P_0} := R$ et $(P_i)_i : P_i = P_0$ entraînent que

$$R_i P_{i+1} = 1 \tag{2.10}$$

Ce résultat n'est pas surprenant puisqu'il signifie que le processus de branchement sous-jacent est un processus de Galton-Watson critique. L'équation (2.10) traduit un certain équilibre entre un nombre de répliques trop important qui entraînerait une explosion du coût et un nombre de répliques insuffisant conduisant à un arrêt prématuré de l'algorithme.

Notons qu'il est aussi possible d'établir un lien entre l'algorithme par duplication des trajectoires et l'échantillonnage préférentiel. En effet, en un certain sens la méthode étudiée ici peut être vue comme une méthode d'échantillonnage préférentiel adaptatif. Les détails se trouvent dans mon manuscrit de thèse [T1].

Choix pratique des nombres de retirages

Bien que nous aimerions prendre R_i tel que $R_i P_{i+1} = 1$ (cf. équation (2.10)), nous sommes contraints de choisir R_i entier et positif. Lorsque la valeur optimale du nombre de retirage n'est pas un entier, un choix naturel est de prendre l'entier le plus proche. Mais dans ce cas, la criticalité du processus de Galton-Watson sous-jacent est perdue. Deux autres stratégies visant à surmonter ce problème ont été proposées dans [J2]. Plus précisément, nous avons considéré les stratégies suivantes et nous avons testé leur efficacité en terme d'intervalles de confiance en utilisant des techniques d'encadrements fins basés sur des fonctions choisies dans les groupes de Lie de faible dimension et des résultats concernant les processus de branchement en environnement aléatoire.

— Stratégie déterministe. Nous prenons pour R l'entier le plus proche de la valeur optimale $1/P_0$. Alors pour α assez petit,

$$\mathbb{P}\left(\frac{|\widehat{P} - \mathbb{P}(A)|}{\mathbb{P}(A)} \geqslant \alpha\right) \leqslant 2\exp\left\{-\frac{\alpha^2}{4}\frac{N\mathbb{E}[Z_{M+1}]^2}{\mathrm{Var}(Z_{M+1})}\left(1-\frac{\alpha}{2}\right)\right\}.$$

— Seconde stratégie : générer à chaque succès. Soient $k \in \mathbb{N}$ et $\delta \in [0,1[$ tels que $1/P_0 = k + \delta$. Pour s'approcher au plus de l'algorithme optimal, on autorise R à varier : chaque fois qu'une particule atteint un niveau supérieur, on génère une réalisation R d'une loi de Bernoulli de paramètre $p = 1-q$. Pour α assez petit et c_2 une constante explicite, nous obtenons dans ce cas

$$\mathbb{P}\left(\frac{|\widehat{P} - \mathbb{P}(A)|}{\mathbb{P}(A)} \geqslant \alpha\right) \leqslant 2\exp\left\{-\frac{c_2}{c(P_0)}\frac{\alpha^2}{(M+1)^2}\right\}.$$

Ce résultat est meilleur que celui obtenu dans la stratégie déterministe puisque ici la borne supérieure est en $\exp\left\{-\frac{(Const)}{(M+1)^2}\right\}$ comme dans le cas optimal.

— Troisième stratégie : générer un environnement aléatoire. On génère un environnement aléatoire décrit par la suite $(R_1, R_2, \ldots, R_M)$ au début de la simulation puis on optimise l'algorithme en $\mathbb{E}[R]$. Nous travaillons ainsi en environnement aléatoire et obtenons des bornes du type de celles établies dans la stratégie où l'on génère à chaque succès.

Nous avons vérifié numériquement que le premier algorithme aléatoire est le plus efficace tandis que la stratégie déterministe fournit les moins bons résultats.

Choix pratique des probabilités de transition

En pratique, les probabilités de transition sont généralement inconnues ; à défaut, nous savons faire évoluer les particules du niveau B_i vers le niveau suivant B_{i+1} (par exemple, comportement markovien). Il devient donc impossible de déduire les nombres de retirage optimaux. Pour pallier ce problème, nous avons proposé une méthode adaptative en deux phases.

— Phase 1. Nous générons ρ_N particules. Les nombres de retirage sont fixés arbitrairement $(R_i^0)_{i=1\dots M}$. De l'étape i $(i = 1, \dots, M+1)$, nous obtenons un estimateur $\widetilde{P}_i^{(1)}$ de P_i égal à la fraction des particules ayant réussi à atteindre B_i depuis B_{i-1}. Maintenant, pour tout $i = 1, \dots, M$, posons

$$\widetilde{R}_i = \frac{1}{\sqrt{\widetilde{P}_i \widetilde{P}_{i+1}}} \sqrt{\frac{1-\widetilde{P}_{i+1}}{1-\widetilde{P}_i}},$$

qui seront les nombres de retirage de la deuxième phase.

— Phase 2. Nous générons $N - \rho_N$ particules. Les nombres de retirage sont $(\widetilde{R}_i)_{i=1\dots M}$ précédemment obtenus. De l'étape i $(i = 1, \dots, M+1)$, nous obtenons un estimateur $\widetilde{P}_i^{(2)}$ de P_i pendant la simulation.

L'optimisation de l'algorithme par minimisation de la variance à coût de simulation fixé a montré que le budget à consacrer à la première phase d'apprentissage doit être de l'ordre de $N^{1/3}$, ce qui est loin d'être intuitif.

Nous avons montré son efficacité en considérant le processus d'Ornstein-Uhlenbeck et en le comparant numériquement à d'autres algorithmes adaptatifs dont celui introduit par Aldous et Vazirani [5], une méthode qui garde le système particulaire en vie [163], et une méthode où les niveaux sont placés au cours de la simulation [55]. Voir [J3] pour plus de détails et pour les résultats numériques.

2.1.2 Le modèle en dimension supérieure

En dimension supérieure, des algorithmes pour estimer les probabilités d'événements rares ont été proposés tels que la "subset simulation" [13] aussi basée sur un partionnement de l'espace en sous-ensembles imbriqués. Des techniques d'échantillonnage préférentiel ont également été élaborées dans ce cadre, basés sur des "design points" [12, 145] ou des pré-échantillons adaptatifs [11]. Lorsque la complexité du modèle générant l'événement rare augmente, il semble difficile d'utiliser cette technique efficacement [229].

En dimension supérieure, en plus du choix optimal des paramètres de l'algorithme, se pose le problème du choix de la fonction d'importance qui définit la forme des niveaux. Nous illustrons l'importance de ce choix dans l'exemple suivant représenté par la Figure 2.2.

Soit $M = 1$ et supposons que ∂B_1 est partitionné en deux sous-ensembles tels que :

$$\begin{cases} \gamma_1(1) = 10^{-2},\ \gamma_1(2) = 0.5, \\ f_1(\mathbf{1}) = 10^{-1},\ f_1(2) = 10^{-3}. \end{cases}$$

Nous obtenons alors $p = 1.5 \cdot 10^{-3}$ et $\gamma_1(\mathbf{1}) = 0.51$.

En simulant des particules issues de O, nous nous attendons à ce que 51% des particules atteignent ∂B_1 dont 50% $\partial B_1^{(2)}$ et seulement 1% en $\partial B_1^{(1)}$. Néanmoins, les particules de $\partial B_1^{(1)}$ sont 100 fois plus susceptibles d'atteindre A que celles de $\partial B_1^{(2)}$. Ainsi, avec cette forme de niveau pour $partiel B_1$, près de 50% des particules simulées le sont inutilement.

La construction de la fonction d'importance lorsque la probabilité cible a une caractérisation de type grande déviation est discutée dans [83]. Ce contexte est également pris en compte dans [223]. Néanmoins, il semble difficile de traduire les résultats obtenus dans le cadre de cette section.

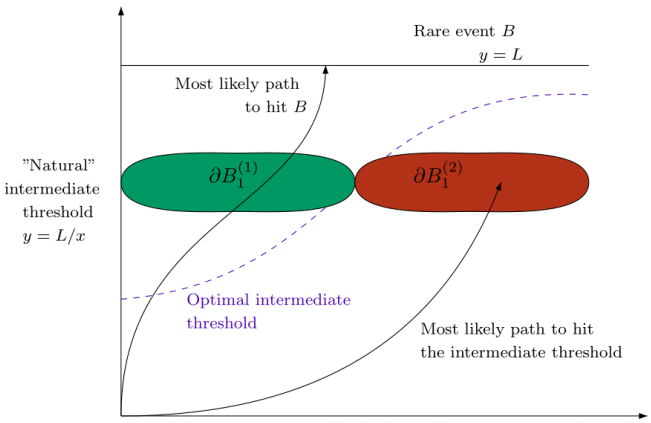


FIGURE 2.2 – Le choix choice de la fonction d'importance sur un exemple.

Dans cette section, nous étudions théoriquement l'algorithme proposé dans [106] et partiellement étudié dans [103]. Nous illustrons sur un exemple concret comment déformer les niveaux intermédiaires de façon à se rapprocher de l'algorithme optimal. Tous les résultats et preuves de cette section se trouvent dans [J12]. S'y trouvent aussi, d'une part une étude quantifiant la sensibilité de la variance lorsqu'un niveau intermédiaire est supprimé permettant ensuite de décider si un niveau doit être conservé ou supprimé ; d'autre part, une analyse de sensibilité de la variance par rapport à une déformation des niveaux. Je ne présenterai pas dans cette section ces deux dernières pistes de recherche.

Ici, nous supposons que chaque frontière ∂B_k de B_k est partitionnée en s sous-ensembles disjoints, notés $\partial B_k^{(i)}$, tels que :

$$\partial B_k = \bigcup_{i=1}^{s} \partial B_k^{(i)}, \quad k = 1, \dots, M.$$

Sans perte de généralité, nous supposons que tous les ∂B_k ont le même nombre s de sous-ensembles. Définissons maintenant τ_k comme le temps d'atteinte de ∂B_k. Ainsi, la probabilité cible s'écrit $\mathbb{P}(A) = \mathbb{P}(\tau_{M+1} < \infty)$. Rappelons que nous notons $(X_k)_{0 \leqslant k \leqslant M+1}$ la chaîne de Markov sous-jacente du processus continu $Y = (Y(t),\ t \geqslant 0)$. Ici nous aurons donc $X_k = i$, et de manière équivalente $Y_{\tau_k} \in \partial B_k^{(i)}$ si la particule au temps τ_k appartient à $\partial B_k^{(i)}$.

L'algorithme en dimension supérieure

Pour estimer la probabilité $\mathbb{P}(A)$, nous procédons comme dans [106] et [103].

Initialisation. Nous générons indépendamment N particules du même point de départ O. Un nombre aléatoire Z_1 de particules atteint le seuil B_1, où Z_1 a une distribution binomiale de paramètres N et $\gamma_1(\mathbf{1})$. Ces particules Z_1 sont réparties sur les sous-ensembles $\partial B_1^{(i)}$ selon une variable aléatoire multinomiale $\text{Mult}(Z_1, \mu_1)$. Soit Z_1 le vecteur aléatoire correspondant $(Z_{11}, \dots, Z_{1r})$.

Etape n $(2 \leqslant n \leqslant M)$. Chacune des Z_{n-1} particules dans ∂B_{n-1} est dupliquée R_{n-1} fois. Ces nouvelles particules évoluent selon la dynamique du processus initial Y et le nombre Z_{nj} de particules ayant atteint $\partial B_n^{(j)}$ est encore aléatoire et se décompose de la façon suivante :

$$Z_{nj} = \sum_{i=1}^{s} Y_{nj}^i, \tag{2.11}$$

où Y_{nj}^i est le nombre de particules issues de $\partial B_{n-1}^{(i)}$ ayant atteint $\partial B_n^{(j)}$ dont le nombre total $Y_n^i = \sum_{j=1}^{s} Y_{nj}^i$ est une variable binomiale de paramètres $R_{n-1} Z_{(n-1)i}$ et $g_{n-1}(i)$.

Chaque $\mathbf{Y}_n^i = (Y_{n1}^i, \ldots, Y_{ns}^i)$ conditionné à Y_n^i suit la loi multinomiale de paramètres Y_n^i et $\mathbb{P}(X_n = \cdot \mid X_{n-1} = i \ ; \ \tau_n < \infty)$.

Dernière étape. Chacune des Z_M particules dans ∂B_M est dupliquée R_M fois. Ces nouvelles particules évoluent selon la dynamique du processus initial Y et le nombre Z_M de particules ayant atteint ∂B_{M+1} se décompose de la façon suivante :

$$Z_{M+1} = \sum_{i=1}^{s} Y_{M+1}^i, \tag{2.12}$$

où Y_{M+1}^i représente le nombre de particules issues de $\partial B_M^{(i)}$ ayant atteint ∂B_{M+1}. Conditionné à $(Z_{M1}, \ldots, Z_{Ms})$, les v.a. Y_{M+1}^i, $i = 1, \ldots, r$ sont indépendantes et de loi binomiale de paramètres $R_M Z_{Mi}$ et $f_M(i)$.

Optimisation de l'algorithme en dimension supérieure

L'estimateur sans biais de $\mathbb{P}(A)$ est encore naturellement donné par (2.2) et sera encore noté $\widehat{P}$. Avant de déterminer sa variance, nous introduisons quelques notations supplémentaires propres à la dimension supérieure. Pour tout $k = 1, \ldots, M$, la mesure γ_k sur la frontière ∂B_k est définie par :

$$\gamma_k(i) = \mathbb{P}(X_k = i \ ; \ \tau_k < \infty).$$

Cette mesure agit sur les fonctions f définies sur ∂B_k selon $\gamma_k(f) = \mathbb{E}\left[f(X_k)\mathbb{1}_{\tau_k < \infty}\right]$ de telle sorte que $\gamma_k(\mathbf{1}) = \mathbb{P}(\tau_k < \infty)$ est la probabilité que la particule atteigne B_k ($\mathbf{1}$ est la fonction unité). Les mesures γ_k n'étant pas des mesures de probabilités, nous introduisons leurs versions normalisées μ_k. Les fonctions f_k définies par :

$$f_k(i) = \mathbb{P}(\tau_{M+1} < \infty \mid X_k = i \ , \ \tau_k < \infty), \quad k = 1, \ldots, M,$$

quantifient la difficulté d'atteindre A en partant de $\partial B_k^{(i)}$. Enfin, nous introduisons les fonctions g_{k-1} définies par :

$$g_{k-1}(i) = \mathbb{P}(\tau_k < \infty \mid X_{k-1} = i \ , \ \tau_{k-1} < \infty),$$

pour tout $k = 2, \ldots, M$. Les notations sont présentées dans la Figure 2.3 avec quelques relations de transport évidentes.

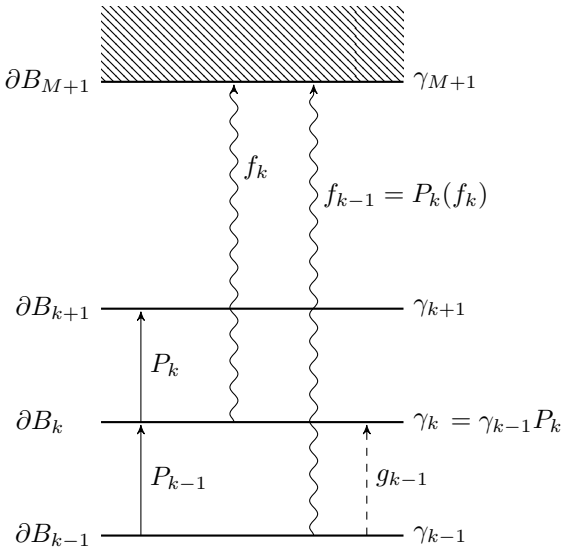


FIGURE 2.3 – Notations introduites pour l'algorithme en dimension supérieure à un.

Proposition 2.3 (Variance de l'estimateur en dimension supérieure)**.** *Le coefficient de variation est donné par :*

$$\frac{\mathrm{Var}(\widehat{P})}{\mathbb{P}(A)^2} = \sum_{k=1}^{M} \frac{1}{\gamma_k(\mathbf{1})} \left(\frac{1}{r_{k-1}} - \frac{1}{r_k} \right) \frac{\mathrm{Var}_{\mu_k}(f_k)}{\mathbb{E}^2_{\mu_k}(f_k)} + \sum_{k=0}^{M} \frac{1}{r_k \gamma_k(\mathbf{1})} \frac{1-\mu_k(g_k)}{\mu_k(g_k)}. \tag{2.13}$$

La variance est donc divisée en deux parties. La première somme décrit la variabilité due à la forme des seuils ∂B_k (définis par les f_k), tandis que la seconde décrit la variabilité due au nombre de seuils M, aux nombres de réplication R_k et à la position des seuils (traduit en termes de P_k et g_k). La formule de la variance (2.13) peut être comparée aux expressions obtenues dans [103, Equation (2.21)], [160], [53] et [107].

Nous considérons maintenant un coût de simulation (réaliste) qui tient compte de la probabilité $P_k(i,j)$ d'atteindre $\partial B_k^{(j)}$ partant de $\partial B_k^{(i)}$. En fait, même si l'algorithme présenté ici est basé sur la simulation de variables aléatoires multinomiales, l'introduction de ce nouveau coût permet de considérer la dynamique d'une particule entre deux seuils successifs à travers les fonctions g_k. Ainsi, nous associons à chaque particule de $\partial B_k^{(i)}$ un coût unitaire $c_k(i)$ qui dépend du seuil de départ et de la difficulté $g_k(i)$ à atteindre le seuil suivant. Plus précisément, nous supposons que :

$$c_0 = c(\gamma_1(\mathbf{1})), \quad c_k(i) = c(g_k(i)), \quad k = 1, \dots, M,$$

où c vérifie les mêmes hypothèses que dans la Section 2.1.1.

Proposition 2.4 (Coût de la simulation)**.** *Le coût (moyen) est donné par :*

$$C_{M+1} = Nc_0 + \sum_{n=1}^{M} r_n \sum_{i=1}^{s} \gamma_n(i) c_n(i) = \sum_{n=0}^{M} r_n \gamma_n(c_n). \tag{2.14}$$

Au vu de la décomposition (2.13) de la variance en deux parties, nous obtenons naturellement que l'optimisation se fait selon deux étapes successives dont le résultat est donné dans la proposition qui suit.

Proposition 2.5 (Optimisation de l'algorithme)**.** *Les paramètres optimaux obtenus en minimisant la variance de l'estimateur à coût de simulation fixé sont les suivants :*

(i) les fonctions f_k ne doivent pas dépendre du point intial dans ∂B_k (condition d'isoprobabilité) ;

(ii) les valeurs optimales des paramètres N, M, $(R_k)_{k=1,\dots,M}$ et $(P_k)_{k=1,\dots,M+1}$ sont celles du cas unidimensionnel (i.e. $s = 1$). Plus précisément, N se déduit de C_{M+1} et tous les R_k sont égaux (à un certain R_0). En outre, pour arriver à un compromis entre un arrêt prématué de l'algorithme ($R_k P_{k+1} \ll 1$) et un coût prohibitif ($R_k P_{k+1} \gg 1$), la condition $R_k P_{k+1} = 1$ doit être satisfaite. Enfin, M est donné par la relation $R_0 \mathbb{P}(A)^{1/(M+1)} = 1$.

Application au processus d'Ornstein-Uhlenbeck en dimension 2

Nous présentons maintenant une procédure adaptative ad-hoc afin de définir des niveaux intermédiaires proches des optimaux définis par la condition d'isoprobabilité. Pour ce faire, nous considérons le processus d'Ornstein-Uhlenbeck en dimension 2 défini par :

$$\begin{cases} \mathrm{d}Y(t) = -\Lambda Y(t)\,\mathrm{d}t + \sigma\,\mathrm{d}W(t),\ t > 0 \\ X_0 = x \in \mathbb{R}^2 \end{cases}$$

où $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2)$ avec $\lambda_1 = 1$, $\lambda_2 = 0.2$, $\sigma = 0.3$ et W un mouvement Brownien (MB) standard en dimension 2.

Nous souhaitons ici estimer la probabilité que le processus d'Ornstein-Uhlenbeck issu de $x = (0.05, 0)$ atteigne le cercle centré en 0 et de rayon 1.5 avant de revenir en 0. Dans ce qui suit, nous prendrons $M = 2$, $A = B_{M+1} = B_3 = D(0, 1.5)$ et dans un premier temps $B_1 = D(0, 0.5)$, $B_2 = D(0, 1)$. La simulation du processus se fait en utilisant Mathematica [260] selon un schéma d'Euler de pas 0.01. Nous commençons par générer $N = 300$ particules indépendantes issues de $x = (0.05, 0)$ et nous considérons comme premier niveau ∂B_1 le cercle de rayon 0.5 et centré en 0. Tout d'abord, nous estimons la densité de la mesure d'occupation du processus dans ∂B_1 [1] par rapport à la mesure de Lebesgue. Cette estimation est basée sur le noyau de von Mises et comme attendu (puisque $\lambda_1 > \lambda_2$), cette densité (représentée dans la Figure 2.4 (à gauche)) est loin d'être uniforme.

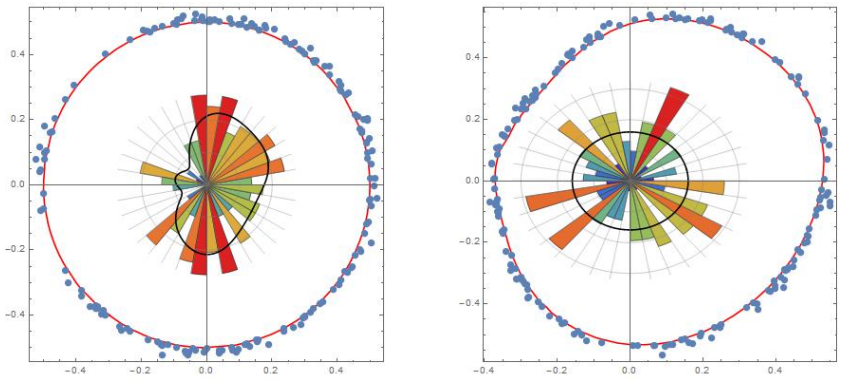


FIGURE 2.4 – La densité de la mesure d'occupation du premier seuil intermédiaire et son estimation (trait noir) basée sur le noyau de von Mises. A gauche, le seuil est le cercle centré de rayon $0,5$ tandis qu'à droite, le seuil est l'image conforme du cercle.

Nous avons vu dans la Proposition 2.5 que l'efficacité de l'algorithme sera optimale en prenant des mesures d'occupation du processus uniformes sur les niveaux intermédiaires. Dans notre cas, puisque $\lambda_1 > \lambda_2$, l'intuition invite à prendre des ellipses pour les nivaux intermédiares, ce qui est confirmé numériquement par la Figure 2.4 (à droite). D'un point de vue théorique, ceci est cohérant avec [8, Theorem 1.3] qui établit que, pour tout $x \in \mathbb{R}^2$, $(Z(t),\ t \geqslant 0) = \left(\frac{\sqrt{2}}{\sigma\sqrt{\log t}} Y(t),\ t \geqslant 0\right)$ admet l'ellipse $\mathcal{E} = \{y = (y_1, y_2) \in \mathbb{R}^2;\ \lambda_1 y_1^2 + \lambda_2 y_2^2 \leqslant 2\}$ comme ensemble de valeurs d'adhérance lorsque t tend vers l'infini.

Dans un second temps, notre objectif est de déformer le premier seuil de façon à se rapprocher du seuil optimal pour lequel la mesure d'occupation est uniforme. Comme le processus vit dans le plan, nous pouvons utiliser une carte conforme, $\varphi_1 : B_1 \to \Omega_1$, afin d'obtenir une mesure d'occupation uniforme. Notons que les cartes conformes sont très pratiques en tant que transformations planaires car elles ne permettent que des rotations locales et des changements d'échelles évitant ainsi des distorsions perturbatrices, en particulier pour les domaines de la forme $\partial\Omega_1 = \varphi_1(\partial B_1)$. Nous suivons la procédure décrite dans [258] pour construire la carte conforme φ_1. Une fois la carte déterminée (voir Figure2.4 (à droite)), nous relançons l'algorithme en utilisant $\partial\Omega_1$ au lieu de ∂B_1.

Insistons sur le fait que notre intention ici n'est pas de proposer un nouvel algorithme utilisant des cartes conformes. Travailler en dimension 2 est déjà difficile et envisager une plus grande dimension devient encore plus complexe. Néanmoins, les fonctions harmoniques ou les cartes quasi-conformes [3, 115] sont la généralisation naturelle des transformations conformes en dimension supérieure. Dans notre contexte,

1. Puisque nous travaillons avec un processus continu, les particules évoluent jusqu'à avoir atteint ∂B_1 ou le disque $D(0, 0.01)$ au lieu de 0.

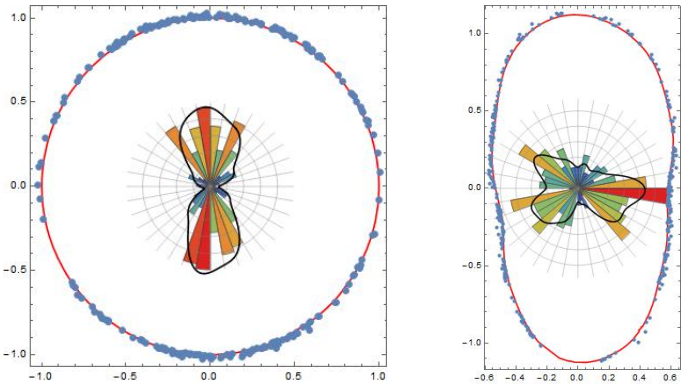


FIGURE 2.5 – En utilisant $R_1 = 2$ pour les particules ayant atteint le premier seuil déformé, nous faisons évoluer ces particules jusqu'à ce qu'elles atteignent le seuil suivant ou le cercle intérieur de rayon 0.01. Les densités empiriques de la mesure d'occupation du cercle unitaire (à gauche) et du seuil déformé (à droite) et leurs estimations respectives (trait noir) reposant sur le noyau de von Mises sont représentées.

nous pourrions commencer par estimer la densité de la mesure d'occupation sur une sphère, considérée comme une forme volume. Ensuite, nous pourrions tenter de déterminer une métrique riemannienne g telle que la forme volume riemannienne associée est égale à la précédente. Enfin, nous pourrions déformer la métrique g dans la métrique uniforme à travers un flot de Ricci, par ex. Nous obtiendrions alors une nouvelle variété riemannienne homéomorphe à la sphère. Pour plus de détails, voir [197].

2.2 Processus gaussiens et estimation de paramètres de covariance

Dans cette section, je présente quelques résultats sur l'estimation de paramètres de fonction de covariance de processus gaussiens. Ces travaux ont été menés avec Jean-Marc Azaïs (IMT), François Bachoc (IMT), Thierry Klein (IMT-ENAC), Andrés López-Lopera (Doctorant Ecole des Mines de Saint Etienne) et Thi Mong Ngoc Nguyen (Université des Sciences d'Ho Chi Minh, Vietnam). Ils ont fait l'objet d'une publication [J13] et de trois prépublications [S19], [S1] et [S3].

Les processus gaussiens

Un processus gaussien Y sur $\mathbb{R}^d$ est un processus stochastique de $\mathbb{R}^d$ dans $\mathbb{R}$ tel que pour tout $p \in \mathbb{N}^*$ et pour tout q-uplet $(x_1, \ldots, x_q)$ d'éléments de $\mathbb{R}^d$, le vecteur aléatoire $(Y(x_1), \ldots, Y(x_q))$ est gaussien [210]. Un processus gaussien est caractérisé par sa fonction moyenne $m : \mathbb{R}^d \to \mathbb{R}$ et sa fonction de covariance $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$. Les fonctions de covariance les plus classiques sont les fonctions exponentielle, gaussienne et de Matérn. D'autres exemples sont donnés dans [241, 210]. La fonction de covariance k est symétrique et semi-définie positive, *i.e.*, pour tout $q \in \mathbb{N}^*$ et pour tous $x_1, \ldots, x_p \in \mathbb{R}^d$, la matrice $(k(x_i, x_j))_{1 \leqslant i,j \leqslant q}$ est symétrique semi-définie positive.

Les processus gaussiens sont devenus très populaires en raison de leur simplicité et de leur flexibilité permettant ainsi de modéliser une large classe de modèles. Leur attractivité réside aussi dans le fait que comme pour la distribution gaussienne, ils se prêtent bien à une analyse théorique fine (voir les résultats énoncés dans la suite de cette section). Pour ces raisons, les processus gaussiens sont depuis quelques années largement utilisés en statistique spatiale afin d'interpoler les observations et proposer un métamodèle, de krigeage par exemple. Les domaines d'application sont nombreux. Citons par exemple la géostatistique [178], l'approximation de code de calcul [222, 224, 20], la calibration [198, 22], l'optimisation globale [137], l'apprentissage [210]...

Cadre de l'étude

Dans ce qui suit, nous considérons un processus gaussien Y centré, défini sur $[0,1]^d$ et à valeurs réelles. Nous supposons ensuite que la fonction de covariance k est stationnaire, *i.e.*, $k(x_1,x_2) = k(x_3,x_4)$ dès que $\|x_2 - x_1\| = \|x_4 - x_3\|$ pour tous $x_1,\dots,x_4 \in \mathbb{R}^d$. Dans ce cas, nous posons $k(x_1,x_2) = k(x_2 - x_1)$ et identifions k à une fonction de $\mathbb{R}^d$ dans $\mathbb{R}$.

Le plan d'expérience consiste en n points d'observation $0 \leqslant x_1 \leqslant \ldots \leqslant x_n \leqslant 1$ et les valeurs observées correspondantes sont $y_1 = Y(x_1),\dots,y_n = Y(x_n)$. Le théorème du conditionnement gaussien [210] affirme alors que Y conditionné à $y = (y_1,\dots,y_n)^\top$ est un processus gaussien de fonction moyenne m_n et de fonction de covariance k_n données par

$$m_n(u) = r_n(u)R_n^{-1}y \tag{2.15}$$

et

$$k_n(u,v) = k(u,v) - r_n(u)R_n^{-1}r_n(v) \tag{2.16}$$

où $x = (x_1,\dots,x_n)^\top$ est le vecteur des points d'observation et $r_n(u)$ est le vecteur de coordonnées $k(u,x_i)$ pour $i = 1,\dots,n$ tandis que R_n est la matrice de taille $n \times n$ de coordonnées $k(x_i,x_j)$ pour $i,j = 1,\dots,n$. Le caractère gaussien de Y conditionné aux observations est l'une des principales raisons pour lesquelles les processus gaussiens sont très populaires en pratique. La fonction moyenne m_n fournit une approximation du processus Y en considérant les observations tandis que la fonction de covariance $k_n(u,u)$ donne un indicateur de l'incertitude sur la valeur de $Y(u)$.

Estimation de la fonction de covariance

Dans ce qui suit, nous supposerons généralement que le processus Y a une fonction moyenne m nulle (sauf mention contraire explicite). Lorsqu'on considère un processus gaussien, on doit estimer sa fonction de covariance. Habituellement, on suppose que la fonction de covariance appartient à une famille paramétrique donnée (voir [1] pour une revue des familles classiques). Dans ce cas, l'estimation se résume à estimer les paramètres de covariance correspondants. Les principales techniques d'estimation qui ont été examinées reposent sur le maximum de vraisemblance [241], la vraisemblance composite [265, 16, 17] et les estimateurs par variations [128, 7].

Dans notre travail, nous supposons que la fonction de covariance appartient à une famille paramétrique de la forme $\{\sigma^2 k_\alpha ; \sigma^2 > 0, \alpha \in A\}$, où $A \subset \mathbb{R}^p$ et k_α est une fonction de corrélation pour tout $\alpha \in \mathbb{R}^p$. La (vraie) fonction de covariance de Y est donnée par $\sigma_0^2 k_{\alpha_0}$ où $\sigma_0^2 > 0$ et $\alpha_0 \in \mathbb{R}^p$. Classiquement, les paramètres de covariance σ_0^2 et α_0 sont estimés par maximum de vraisemblance [210, 241]. L'estimateur par maximum de vraisemblance (MLE) consiste à maximiser la log vraisemblance $\mathcal{L}_n(\sigma^2,\alpha)$ donnée par

$$\mathcal{L}_n(\sigma^2,\alpha) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2}\ln(\det(R_{n,\alpha})) - \frac{1}{2\sigma^2}y^\top R_{n,\alpha}^{-1}y, \tag{2.17}$$

où $R_{n,\alpha}$ est la matrice $n \times n$ donnée par $(k_\alpha(x_i,x_j))_{1\leqslant i,j\leqslant n}$, $y = (y_1,\dots,y_n)^\top$ est le vecteur des observations et $\det(M)$ représente le déterminant de la matrice M. Ainsi, nous avons naturellement

$$(\hat{\sigma}^2_{ML},\hat{\alpha}_{ML}) \in \underset{\sigma^2>0,\alpha\in A}{\operatorname{argmin}} \left(n\ln(\sigma^2) + \ln(\det(R_{n,\alpha})) + \frac{1}{\sigma^2}y^\top R_{n,\alpha}^{-1}y \right). \tag{2.18}$$

Cadre asymptotique et état de l'art sur le MLE

Pour ces analyses, deux contextes ont été envisagés. Le cadre asymptotique de remplissage (fixed-domain asymptotic ou parfois infill asymptotic) [72, 241], correspond au cas où de plus en plus de données sont

observées dans un domaine d'échantillonnage borné fixe de $\mathbb{R}^d$. Au contraire, le cadre asymptotique par expansion (increasing-domain asymptotic) correspond au cas où le domaine d'échantillonnage dans $\mathbb{R}^d$ augmente avec le nombre de points observés et la distance entre deux points d'échantillonnage ne tend pas vers 0. Le comportement asymptotique de l'estimateur par maximum de vraisemblance (MLE) des paramètres de covariance peut être tout à fait différent selon le cas [266]. En effet, dans le cadre asymptotique par expansion, en général, pour tous les paramètres de covariance (identifiables), le MLE est consistant et asymptotiquement normal sous certaines conditions de régularité. La matrice de covariance asymptotique est égale à l'inverse de la matrice d'information de Fisher (asymptotique) [17, 73, 175, 231]. La situation est significativement différente lorsqu'on travaille dans le cadre asymptotique de remplissage. En effet, on distingue deux types de paramètres de covariance : les paramètres microergodiques et non microergodiques. Un paramètre de covariance est *microergodique* si, pour deux valeurs différentes de celui-ci, les deux mesures gaussiennes correspondantes sont orthogonales, voir [126, 241]. Il est *non microergodique* si, pour deux valeurs différentes, les deux mesures gaussiennes correspondantes peuvent être équivalentes. Les paramètres non microergodiques ne peuvent pas être estimés de manière consistante, mais une mauvaise spécification peut conduire asymptotiquement à la même inférence statistique qu'une spécification correcte [237, 238, 240, 264]. Dans le cas des fonctions de covariance isotropes de Matérn avec $d \leqslant 3$, [264] montre que seule une quantité reparamétrisée obtenue à partir de la variance et des paramètres d'échelle est microergodique. La normalité asymptotique du MLE de ce paramètre microergodique est alors prouvée [141]. Des résultats similaires pour le cas particulier de la fonction de covariance exponentielle ont été obtenus précédemment dans [262].

Le MLE est généralement considéré comme la meilleure option pour estimer les paramètres de covariance d'un processus gaussien (au moins dans notre cadre, où la véritable fonction de covariance appartient au modèle paramétrique, voir [16, 19]). Néanmoins, l'évaluation de la fonction de vraisemblance nécessite de résoudre un système d'équations linéaires et de calculer un déterminant. Pour un ensemble de n observations, le fardeau computationnel est en $O(n^3)$, ce qui rend cette méthode impraticable lorsque la taille de l'échantillon devient grande. Ce fait motive la recherche de méthodes d'estimation proposant un compromis entre complexité computationnelle et efficacité statistique. Parmi ces méthodes, on peut citer l'approximation de faible rang (voir [242] et les références qu'il contient), l'approximation creuse [117], l'atténuation de covariance (covariance tapering) [99, 142], l'approximation par champs aléatoires gaussienns [80, 220], l'agrégation de sous-mdèles [49, 84, 118, 221, 246, 250], les vraisemblances composites...

Une alternative au MLE

La vraisemblance composite fait référence à une classe générale de fonctions basées sur la vraisemblance de marginales ou d'événements conditionnels [251]. Ces méthodes d'estimation a deux avantages notables : elles sont généralement attrayantes lorsque la taille de l'échantillon est grande et elles peuvent être utiles lorsqu'il est difficile de déterminer la vraisemblance complète. Les estimateurs par vraisemblance composite (CLE) maximisent la somme sur $i = 1, \ldots, n$ de la log vraisemblance conditionnelle de y_i connaissant un sous-ensemble $\{y_1, \ldots, y_n\} \backslash \{y_i\}$ correspondant à des points d'observation proches de x_i. Ils ont été considérés dans [177, 195, 243, 252]. Nous expliquerons en détails cette méthode dans la Section 2.2.1. Plus généralement, le principe de conditionnement basé sur des points d'observation voisins plutôt que sur l'ensemble des points d'observation a largement été appliqué pour les processus gaussiens [109, 110].

Malgré leur popularité dans la pratique, aucun résultat asymptotique dans le cadre asymptotique de remplissage n'existe pour les estimateurs CLE. Les résultats existants traitent le cas de la fonction de covariance exponentielle en dimension une. Dans le cas exponentiel, en raison de la propriété de Markov, le CLE conditionné à $\{y_{i-K}, \ldots, y_{i-1}\}$ quelle que soit la valeur de K est simplement le CLE conditionné

à y_{i-1} et coïncide donc aussi avec le MLE (voir [262]). Il est alors prouvé que le CLE du paramètre microergodique est asymptotiquement gaussien. Lorsque chaque observation est conditionnée à ses deux observations voisines les plus proches, alors le CLE du paramètre microergodique est également asymptotiquement gaussien [J13] (voir Section 2.2.2 pour plus de détails). Enfin, notons que les estimateurs de vraisemblance par paires ont été analysés récemment, dans le cas de la fonction de covariance exponentielle en dimension une [21]. Les auteurs somment sur des paires d'observations les log vraisemblances en dimension deux correspondantes.

Plan de la section

Dans cette section, nous considérons le MLE, le CLE et les estimateurs par variations dans des contextes et des problématiques différents. Dans tous les cas, nous travaillons dans le cadre asymptotique de remplissage. Dans la Section 2.2.1, nous étudions l'estimateur par vraisemblance composite et ses propriétés asymptotiques (cf. [S3]). Nous le comparons à l'estimateur par variations étudié en Section 2.2.3 (cf. [S1]). Enfin, nous considérons dans la Section 2.2.4 des processus gaussiens contraints par des inégalités et étudierons le MLE et ses propriétés asymptotiques dans ce cadre (cf. [S19]).

2.2.1 Vraisemblance composite

Définissons maintenant rigoureusement l'estimateur par vraisemblance composite. Soient $K \in \mathbb{N}$ et $L \in \mathbb{N}$ fixés. Le CLE maximise la somme sur $i = K+1, \dots, n-L$ de la log vraissemblance de y_i conditionnée aux observations $\{y_{i-K}, \dots, y_{n-L}\} \backslash \{y_i\}$, voisines de y_i. Il est donc donné par

$$(\hat{\sigma}^2_{CL}, \hat{\alpha}_{CL}) \in \underset{\sigma^2>0, \alpha \in A}{\operatorname{argmax}} \sum_{i=K+1}^{n-L} \mathcal{L}_{\sigma^2,\alpha}\left(y_i | y_{i-K}, \dots, y_{i-1}, y_{i+1}, \dots, y_{i+L}\right), \tag{2.19}$$

où $\mathcal{L}_{\sigma^2,\alpha}(y_i | y_{i-K}, \dots, y_{i-1}, y_{i+1}, \dots, y_{i+L})$ est défini comme le logarithme de la densité de y_i sachant $y_{i-K}, \dots, y_{i-1}, y_{i+1}, \dots, y_{i+L}$ sous les paramètres de covariance σ^2 et α. Le coût computationnel du critère défini par (2.19) est en $O(n)$ si K et L sont fixés, à comparer à $O(n^3)$ pour le MLE défini en (2.18).

Pour tout $i \in \{K+1, \dots, n-L\}$, définissons le vecteur $r_{K,L,\alpha;i}$ par

$$r_{K,L,\alpha;i} = \left(k_\alpha(x_{i-K}, x_i), \dots, k_\alpha(x_{i-1}, x_i), k_\alpha(x_{i+1}, x_i), \dots, k_\alpha(x_{i+L}, x_i)\right)^\top,$$

le vecteur des observations locales $y_{K,L;i}$ par

$$y_{K,L;i} = \left(y_{i-K}, \dots, y_{i-1}, y_{i+1}, \dots, y_{i+L}\right)^\top$$

et la matrice $R_{K,L,\alpha;i}$ par

$$R_{K,L,\alpha;i} = \begin{pmatrix} k_\alpha(x_{i-K}, x_{i-K}) & \dots & k_\alpha(x_{i-K}, x_{i-1}) & k_\alpha(x_{i-K}, x_{i+1}) & \dots & k_\alpha(x_{i-K}, x_{i+L}) \\ \vdots & & & & & \vdots \\ k_\alpha(x_{i-1}, x_{i-K}) & \dots & k_\alpha(x_{i-1}, x_{i-1}) & k_\alpha(x_{i-1}, x_{i+1}) & \dots & k_\alpha(x_{i-1}, x_{i+L}) \\ k_\alpha(x_{i+1}, x_{i-K}) & \dots & k_\alpha(x_{i+1}, x_{i-1}) & k_\alpha(x_{i+1}, x_{i+1}) & \dots & k_\alpha(x_{i+1}, x_{i+L}) \\ \vdots & & & & & \vdots \\ k_\alpha(x_{i+L}, x_{i-K}) & \dots & k_\alpha(x_{i+L}, x_{i-1}) & k_\alpha(x_{i+L}, x_{i+1}) & \dots & k_\alpha(x_{i+L}, x_{i+L}) \end{pmatrix}.$$

D'après le théorème du conditionnement gaussien énoncé dans l'introduction de cette section, pour $i =$

$K+1,\dots,n-L$, nous avons

$$\mathcal{L}_{\sigma^2,\alpha}\left(y_i|y_{i-K},\dots,y_{i-1},y_{i+1},\dots,y_{i+L}\right) = -\frac{1}{2}\ln\sigma^2 - \frac{1}{2}\ln\left(1-r_{K,L,\alpha;i}^\top R_{K,L,\alpha;i}^{-1} r_{K,L,\alpha;i}\right) - \frac{\left(y_i - r_{K,L,\alpha;i}^\top R_{K,L,\alpha;i}^{-1} y_{K,L;i}\right)^2}{2\sigma^2\left(1-r_{K,L,\alpha;i}^\top R_{K,L,\alpha;i}^{-1} r_{K,L,\alpha;i}\right)}$$

(à une constante additive près).

Dans ce qui suit, nous considérons sans perte de généralité une grille régulière donnée par les points d'observation suivants $\{x_1 = 1/n,\dots,x_n = 1\}$. Soit A un sous-ensemble compact de $(0,\infty)$ et posons $k_\alpha(t) = k(\alpha t)$, où k est une fonction de corrélation stationnaire donnée. Les paramètres à estimer sont donc la variance σ^2 et le paramètre d'échelle spatiale α.

Estimation du paramètre de variance uniquement

Tout d'abord, nous supposons que A est réduit au singleton $\{1\}$, que $\alpha_0 = 1$ (*i.e.* la fonction de corrélation est connue) et que nous souhaitons déterminer les propriétés asymptotiques du CLE de la variance inconnue σ_0^2.

Introduisons maintenant le processus réduit Z tel que $Y = \sigma_0 Z$. Le processus Z est centré, gaussien de fonction de covariance k et d'observations $z_i = y_i/\sigma_0$. Par analogie à $y_{K,L;i}$, nous définissons aussi $z_{K,L;i}$. Pour tout $i = K+1,\dots,n-L$, la prédiction est donnée par

$$\hat{z}_i = \mathbb{E}[z_i|z_{K,L;i}] = r_{K,L,\alpha;i}^\top R_{K,L,\alpha;i}^{-1} z_{K,L;i}$$

tandis que la variance prédite est donnée par

$$\hat{\sigma}_i^2 = \mathrm{Var}(z_i|z_{K,L;i}) = 1 - r_{K,L,\alpha;i}^\top R_{K,L,\alpha;i}^{-1} r_{K,L,\alpha;i}. \tag{2.20}$$

Après quelques calculs, nous déduisons la valeur de $\hat{\sigma}_{CL}^2$ défini par (2.19) :

$$\frac{\hat{\sigma}_{CL}^2}{\sigma_0^2} = \frac{1}{n-L-K}\sum_{i=K+1}^{n-L}\frac{(z_i-\hat{z}_i)^2}{\hat{\sigma}_i^2}. \tag{2.21}$$

Clairement, c'est un estimateur sans biais dont la variance est donnée par

$$\mathrm{Var}\left(\frac{\hat{\sigma}_{CL}^2}{\sigma_0^2}\right) = \frac{2}{(n-L-K)^2}\sum_{i,j=K+1}^{n-L}\frac{\mathrm{Cov}\left(z_i-\hat{z}_i, z_j-\hat{z}_j\right)^2}{\hat{\sigma}_i^2\hat{\sigma}_j^2} \tag{2.22}$$

après application de la formule de Mehler. Dans la suite, nous supposons que k satisfait la condition suivante.

Condition 2.6. *La fonction de corrélation k a le développement suivant lorsque t tend vers 0*

$$k(t) = 1 - |t|^s + r(t), \tag{2.23}$$

où r est deux fois différentiable, $r(0) = 0$ et $0 < s < 3/2$. De plus, pour $1/2 \leqslant s < 3/2$, r'' est borné.

La Condition 2.6 signifie que le processus gaussien Y est continu mais pas dérivable. La quantité s supposée connue s'interprète comme un paramètre de régularité. Par exemple, la fonction de covariance exponentielle $\exp\{-|t|\}$ correspond à $s = 1$ [21, 262] et [J13]. Elle implique que $r(t) = O(t)$ pour tout

$0 < s < 3/2$ et $r(t) = O(t^2)$ pour $1/2 \leqslant s < 3/2$. La Condition 2.6 est satisfaite pour les fonctions de covariance Matérn $\tilde{k}_\nu$ définies par $\tilde{k}_\nu(t) = k_\nu(A_\nu^{-1/2\nu}t)$ avec $s = 2\nu$, A_ν une constante dépendant de ν uniquement et $r(t) = O(t^2)$.

Soit maintenant b la matrice inverse de la matrice B de taille $(K+L)\times(K+L)$ définie par $B_{i,j} = i^s + j^s - |i-j|^s$ (voir [S1] pour la preuve de l'inversibilité de B). Définissons de plus la matrice C de taille $(K+L)\times(K+L)$ donnée par $C_{i,j} = i^s j^s$. Lorsque $0 < s < 1/2$, la condition supplémentaire suivante sera nécessaire.

Condition 2.7. *Supposons que $K \in \mathbb{N}$, $L \in \mathbb{N}$ et $0 < s < 1/2$ sont tels que*

$$\sum_{k=1}^{K+L} b_{K,k} k^s \neq 0.$$

Nous avons prouvé que la Condition 2.7 est satisfaite pour $K = 0$ ou $L = 0$. De façon générale, nous n'avons pas montré que la Condition 2.7 est satisfaite. Cependant, numériquement cela semble être le cas pour toutes les valeurs de K et L que nous avons testées.
En utilisant les formules virtuelles un à part (Leave-One-Out virtual formulas - LOO) [16, Proposition 3.1], la variance prédite définie en (2.20) est donnée par

$$\hat{\sigma}_i^2 = \frac{1}{\left(\mathrm{Cov}_{|1}(z_2,\ldots,z_K,z_{K+1},z_{K+2},\ldots,z_{K+L+1})^{-1}\right)_{K,K}}$$

et l'erreur de prédiction est donnée par

$$z_i - \hat{z}_i = \frac{1}{b_{K,K}^{(n)}} \sum_{k=1}^{K+L} b_{K,k}^{(n)} \left(z_{i-K+k} - \left(1 - \frac{k^s}{n^s} + r\left(\frac{|k|}{n}\right)\right) z_{i-K} \right),$$

où $\mathrm{Cov}_{|l}$ représente la covariance conditionnelle sachant z_l.

Théorème 2.8 (Comportement asymptotique du CLE de la variance)**.** *Supposons que la Condition 2.6 est satisfaite et soient $K \geqslant 0$ et $L \geqslant 0$ tels que $K + L \geqslant 2$.*

(i) Si $0 < s < 1/2$, alors

$$\mathrm{Var}\left(\frac{\hat{\sigma}_{CL}^2}{\sigma_0^2}\right) = O\left(\frac{1}{n^{2s}}\right). \tag{2.24}$$

En outre, si la Condition 2.7 est vérifiée,

$$\mathrm{Var}\left(\frac{\hat{\sigma}_{CL}^2}{\sigma_0^2}\right) \sim \frac{4}{n^{2s}} \frac{(bCb)_{K,K}^2}{b_{K,K}^2} \int_0^1 (1-t)(1-t^s+r(t))^2 \mathrm{d}t. \tag{2.25}$$

(ii) Si $1/2 \leqslant s < 3/2$, alors

$$\mathrm{Var}\left(\frac{\hat{\sigma}_{CL}^2}{\sigma_0^2}\right) = O\left(\frac{1}{n}\right). \tag{2.26}$$

Remarquons que les valeurs du nombre de voisins n'ont aucun impact sur la vitesse de convergence, mais interviennent seulement dans la variance asymptotique dans le cas $0 < s < 1/2$. Clairement, on en déduit la consistance de $\hat{\sigma}^2$ dès que $0 < s < 3/2$ puisque la variance tend vers zéro. Lorsque $K + L = 1$, la preuve du Théorème 2.8 ne peut plus être appliquée mais on peut facilement prouver que (2.25) et (2.26) sont encore vérifiés.

Comparons le résultat précédent à celui obtenu pour l'estimateur par variations défini ultérieurement dans la Section 2.2.3. Remarquons que dans le cas où a est la séquence élémentaire d'ordre 1 : $a_0 = 1$ et $a_1 = -1$, nous avons montré que

$$\mathbb{E}[(C_{a,n} - \sigma_0^2)^2] = O\left(\frac{1}{n}\right),$$

ce qui signifie que l'estimateur par variations $C_{a,n}$ basé sur les différences $y_i - y_{i-1}$ conduit à la vitesse optimale $n^{1/2}$. En revanche, utiliser les différences $y_i - \mathbb{E}[y_i|y_{i-1}]$ conduit à la vitesse sous-optimale n^s lorsque $0 < s < 1/2$ d'après le Théorème 2.8. Ainsi, contrairement à l'intuition, utiliser $y_i - \mathbb{E}[y_i|y_{i-1}]$ semble être plus efficace que $y_i - y_{i-1}$. De façon analogue, les résultats précédents montrent qu'utiliser $y_i - \mathbb{E}[y_i|y_{i-K}, \ldots, y_{i-1}, y_{i+1}, \ldots, y_{i+L}]$ est moins efficace que $\sum_{j=0}^{L(a)-1} a_j y_{i+j}$ lorsque $0 < s < 1/2$, ce qui est loin d'être intuitif.
Le Théorème 2.8 montre donc que pour $0 < s < 1/2$, le CLE converge à la vitesse sous-optimale $n^s < n^{1/2}$, alors qu'il existe des estimateurs de σ_0^2 convergeant à la vitesse optimale $n^{1/2}$, par exemple le MLE défini dans (2.18) ou encore les estimateurs par variations quadratiques (voir la Section 2.2.3 ci-dessous). Cela peut être considéré comme un inconvénient du CLE, puisque contrairement au MLE, les estimateurs par variations requièrent un faible coût de calcul (en $O(n)$) pour une vitesse optimale de $n^{1/2}$ quelle que soit la valeur de $0 < s < 3/2$. Dans la proposition suivante, nous montrons que le CLE $\hat{\sigma}^2_{CL}$ converge vers une variable qui n'est pas gausienne lorsque $0 < s < 1/2$.

Proposition 2.9 (Non gaussianité de la distribution limite)**.** *Supposons que les Conditions 2.6 et 2.7 sont vérifiées. Alors pour $0 < s < 1/2$, $K \geqslant 0$ et $L \geqslant 0$ tels que $K + L \geqslant 1$, la v.a. $n^s\left(\hat{\sigma}^2_{CL}/\sigma_0^2 - 1\right)$ ne converge pas en loi vers une v.a. gaussienne. Plus précisément,*

$$n^s\left(\frac{\hat{\sigma}^2_{CL}}{\sigma_0^2} - 1\right) \xrightarrow[n\to\infty]{\mathcal{L}} \frac{(bCb)_{K,K}}{b_{K,K}}\left(\int_0^1 Z(t)^2 \mathrm{d}t - 1\right)$$

où Z représente le processus réduit associé au processus gaussien Y.

Estimation jointe des paramètres de variance et d'échelle spatiale
Supposons maintenant que A est un compact de $(0, \infty)$. Ainsi il s'agit d'estimer les paramètres de variance et d'échelle spatiale. Nous estimons alors le paramètre microergodique $\sigma_0^2\alpha_0^s$ (voir [264]). Supposons que les vrais paramètres du modèle σ_0^2 et α_0 sont fixés dans $(0, \infty)$, sans supposer nécessairement que α_0 appartient à A (simplement parce que σ^2 n'est pas restreint dans (2.19) et donc il existe $(\sigma^2, \alpha) \in (0, \infty) \times A$ tel que $\sigma^2\alpha^s = \sigma_0^2\alpha_0^s$).

Théorème 2.10 (Comportement asymptotique du CLE du paramètre microergodique)**.** *Supposons que k vérifie la Condition 2.6 et soient $K \geqslant 0$ et $L \geqslant 0$ tels que $K + L \geqslant 2$.*

(i) Si $0 < s < 1/2$, alors

$$\frac{\hat{\sigma}^2_{CL}\hat{\alpha}^s_{CL}}{\sigma_0^2\alpha_0^s} - 1 = \frac{\hat{\alpha}^s_{CL}}{n^s}\frac{(bCb)_{K,K}}{b_{K,K}}\left(\frac{\hat{\alpha}^s}{\alpha_0^s}\int_0^1 Z(t)^2\mathrm{d}t - 1\right) + o_{\mathbb{P}}\left(\frac{1}{n^s}\right). \tag{2.27}$$

(ii) Si $1/2 \leqslant s < 3/2$, alors

$$\frac{\hat{\sigma}^2_{CL}\hat{\alpha}^s_{CL}}{\sigma_0^2\alpha_0^s} - 1 = O\left(\frac{1}{\sqrt{n}}\right). \tag{2.28}$$

L'interprétation du Théorème 2.10 est la même que celle du Théorème 2.8. Ici, nous ne pouvons pas analyser la variance de $\hat{\sigma}^2_{CL}\hat{\alpha}^s_{CL}$, parce que $\hat{\alpha}^s_{CL}$ n'a pas d'expression explicite. C'est pourquoi le Théorèmme

2.10 est énoncé en termes de convergence en distribution et en probabilité plutôt qu'en termes de variance ou d'erreur quadratique moyenne.

Notons que lorsque $0 < s < 1/2$, l'approximation asymptotique de $\hat{\sigma}^2_{CL}\hat{\alpha}^s_{CL}/\hat{\sigma_0}^2\hat{\alpha_0}^s - 1$ dépend de la distribution de $\hat{\alpha}^s_{CL}$, pour laquelle peu de résultats sont connus ([266] considère la fonction de covariance exponentielle pour laquelle $s = 1$). Néanmoins, la variable aléatoire

$$\hat{\alpha}^s_{CL}\frac{(bCb)_{K,K}}{b_{K,K}}\left(\frac{\hat{\alpha}^s}{\alpha^s_0}\int_0^1 Z(t)^2\mathrm{d}t - 1\right) \tag{2.29}$$

n'est pas gaussienne. Par exemple, son minimum est $-\alpha^s_u(bCb)_{K,K}/b_{K,K} > -\infty$ où α_u est le supremum du compact A. De plus, cette v.a. n'est pas non plus constante car $\hat{\alpha}^{2s}_{CL}\int_0^1 Z(t)^2\mathrm{d}t$ ne l'est pas. En effet, $\int_0^1 Z(t)^2\mathrm{d}t$ a une variance non nulle et une probabilité non nulle d'appartenir à tout intervalle compact $[0, \epsilon]$ pour $\epsilon > 0$ relativement petit (voir [165, 171]) et $\hat{\alpha}^{2s}_{CL}$ est borné par α^{2s}_u. Enfin, si $A = \{\alpha_1\}$ pour un $\alpha_1 \in (0, \infty)$ donné, alors $n^s(\hat{\sigma}^2_{CL}\hat{\alpha}^s_{CL}/\sigma^2_0\alpha^s_0 - 1)$ converge vers une v.a. non gaussienne, dont la variance est proportionnelle à α^{4s}_1.

De la même façon que pour le Théorème 2.8, le Théorème 2.10 ne s'applique pas lorsque $K + L = 1$, mais nous avons montré que la conclusion reste valable malgré tout.

2.2.2 Validation croisée

Antérieurement aux travaux de la section précédente, nous avons montré dans [J13] que l'estimateur par validation croisée dans le modèle exponentiel est consistant et converge en distribution vers une loi gaussienne. Dans ce cas, le paramètre de régularité s est égal à 1. Ce résultat précise donc le cas (ii) du Théorème 2.10.

Plus précisément, dans cette section, le processus gaussien Y a pour fonction de covariance $\sigma^2_0 k_{\alpha_0}$ où $k_{\alpha_0}(t) = e^{-\alpha_0|t|}$. Ce processus est connu sous le nom de processus d'Ornstein-Uhlenbeck. Il est gouverné par l'équation différentielle stochastique suivante, appelée équation de Langevin,

$$\mathrm{d}Y(t) = -\alpha_0 Y(t)\mathrm{d}t + \sqrt{2\alpha_0}\sigma_0\mathrm{d}B(t),$$

où $(B(t),\ t \geqslant 0)$ est le MB standard. Le processus d'Ornstein-Uhlenbeck est largement utilisé pour modéliser des phénomènes physiques, biologiques, sociaux... Il possède de nombreuses propriétés mathématiques qui permettent de simplifier son analyse.

Le processus est observé sur un tableau triangulaire de points $(x^{(n)}_i)_{n\in\mathbb{N},i=1,\dots,n}$. Par souci de concision, notons simplement $(x^{(n)}_1, \dots, x^{(n)}_n) = (x_1, \dots, x_n)$. Les observations correspondantes sont données par $(y_1, \dots, y_n) = (Y(x_1), \dots, Y(x_n))$. Supposons que le modèle paramétrique est de la forme $\{\sigma^2 k_\alpha;\ (\sigma^2, \alpha) \in [\sigma^2_l, \sigma^2_u] \times [\alpha_l, \alpha_u]\}$ pour des valeurs fixées telles que $0 < \sigma^2_l \leqslant \sigma^2_u < +\infty$ et $0 < \alpha_l \leqslant \alpha_u < +\infty$ et où $k_\alpha(t) = e^{-\alpha|t|}$.

Nous étudions maintenant l'estimateur par validation croisée de (σ^2_0, α_0) reposant sur le score logarithmique considéré dans [210, 265]. Posons

$$\hat{Y}_{\alpha,-i}(x_i) = \mathbb{E}_{\sigma^2,\alpha}[y_i|y_1, \dots, y_{i-1}, y_{i+1}, \dots, y_n],$$

où l'espérance conditionnelle $\mathbb{E}_{\sigma^2,\alpha}$ est calculée en supposant que Y est centré de fonction de covariance $\sigma^2 k_\alpha$. Remarquons que $\hat{Y}_{\alpha,-i}(s_i)$ ne dépend pas de σ^2. Similairement, nous définissons

$$\hat{\sigma}^2_{\sigma^2,\alpha,-i}(x_i) = \mathrm{Var}_{\sigma^2,\alpha}(y_i|y_1, \dots, y_{i-1}, y_{i+1}, \dots, y_n).$$

Alors, l'estimateur par validation croisée (CVE) est donné par

$$(\hat{\sigma}^2_{CV}, \hat{\alpha}_{CV}) \in \underset{\sigma_l^2 \leqslant \sigma^2 \leqslant \sigma_u^2, \alpha_l \leqslant \alpha \leqslant \alpha_u}{\operatorname{argmax}} S_n(\sigma^2, \alpha),$$

où

$$S_n(\sigma^2, \alpha) = -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\sum_{i=1}^{n} \ln(\hat{\sigma}^2_{\sigma^2,\alpha,-i}(s_i)) - \sum_{i=1}^{n} \frac{(y_i - \hat{Y}_{\alpha,-i}(s_i))^2}{\hat{\sigma}^2_{2\sigma^2,\alpha,-i}(s_i)} \tag{2.30}$$

est égal à la log vraisemblance conditionnelle de y_i sachant $(y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n)^\top$ sous les paramètres de covariance (σ^2, α) (à une constante additive près). Le terme *validation croisée* met en lumière le fait que nous considérons des quantités un à part (Leave-One-Out quantities).
Une fois encore, nous considérons le paramètre microergodique $\sigma_0^2\alpha_0$ pour lequel [262] a prouvé la consistance et la normalité asymptotique du MLE défini en (2.18). De même que dans [262], nous avons considéré trois cas.

(i) Soit $\sigma^2 = \sigma_1^2$ défini par (2.30) avec $\sigma_1^2 > 0$ une constante fixée. Nous considérons le CVE $\hat{\alpha}_{CV,1}$ de $\alpha_1 = \alpha_0\sigma_0^2/\sigma_1^2$ qui minimise (2.30) avec $\sigma^2 = \sigma_1^2$.

(ii) Soit $\alpha = \alpha_2$ défini par (2.30) avec $\alpha_2 > 0$ une constante fixée. Nous considérons le CVE $\hat{\sigma}^2_{CV,2}$ de $\sigma_2^2 = \theta_0\sigma_0^2/\alpha_2$ qui minimise (2.30) avec $\alpha = \alpha_2$.

(iii) Nous considérons le CVE $\hat{\sigma}^2_{CV}\hat{\alpha}_{CV}$ de $\sigma_0^2\alpha_0$, où $\hat{\alpha}_{CV}$ et $\hat{\sigma}^2_{CV}$ sont les estimateurs de α_0 et σ_0^2.

Inspirés de [262], nous basons notre analyse sur la propriété de Markov du processus d'Ornstein-Uhlenbeck afin de gérer le fait que, lorsque n augmente, les observations $(y_1, \ldots, y_n)^\top$ deviennent de plus en plus corrélées. Dans notre cadre, nous avons

$$\hat{Y}_{\alpha,-i}(s_i) = -\sum_{\substack{j=1,\ldots,n;\\ j\neq i}} \frac{(R_\alpha^{-1})_{ij}}{(R_\alpha^{-1})_{ii}} y_j \quad \text{et} \quad \hat{\sigma}^2_{\alpha,\sigma^2,-i}(s_i) = \frac{\sigma^2}{(R_\alpha^{-1})_{ii}}, \tag{2.31}$$

d'après [16, 90, 265]. Nous n'énonçons dans ce manuscrit que les résultats concernant $\hat{\sigma}^2_{CV}\hat{\alpha}_{CV}$.

Théorème 2.11 (Consistance et normalité asymptotique du CVE $\hat{\sigma}^2_{CV}\hat{\alpha}_{CV}$)**.** *Supposons que*

$$\limsup_{n\to+\infty} \max_{i=2,\ldots,n} \Delta_i = 0, \tag{2.32}$$

où $\Delta_i = x_i - x_{i-1}$. *Soit* $\Theta = [\sigma_l^2, \sigma_u^2] \times [\alpha_l, \alpha_u]$, *où* σ_l^2, σ_u^2, α_l *et* α_u *sont fixés. Supposons qu'il existe* $(\tilde{\sigma}^2, \tilde{\alpha})$ *dans* Θ *tel que* $\tilde{\sigma}^2\tilde{\alpha} = \sigma_0^2\alpha_0$. *Soit* $(\hat{\sigma}^2_{CV}, \hat{\alpha}_{CV}) \in \Theta$ *une solution de*

$$S_n(\hat{\sigma}^2_{CV}, \hat{\alpha}_{CV}) = \max_{(\sigma^2,\alpha)\in\Theta} S_n(\sigma^2, \alpha). \tag{2.33}$$

Alors $(\hat{\sigma}^2_{CV}, \hat{\alpha}_{CV})$ *existe et est consistant :*

$$\hat{\sigma}^2_{CV}\hat{\alpha}_{CV} \xrightarrow[n\to+\infty]{p.s.} \sigma_0^2\alpha_0.$$

Supposons de plus que $\sigma_l^2\alpha_u < \sigma_0^2\alpha_0 < \sigma_u^2\alpha_l$ *ou* $\sigma_l^2\alpha_u > \sigma_0^2\alpha_0 > \sigma_u^2\alpha_l$. *Alors la normalité asymptotique est vérifiée*

$$\frac{\sqrt{n}}{\sigma_0^2\alpha_0\tau_n}(\hat{\sigma}^2_{CV}\hat{\alpha}_{CV} - \sigma_0^2\alpha_0) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0,1), \tag{2.34}$$

où τ_n *dépend du choix des points d'observation (voir équation (11) dans [J13]).*

Les conditions sur σ_l^2, σ_u^2, α_l et α_u assurent que $(\partial/\partial\alpha)S_n(\hat{\alpha}, \hat{\sigma}^2)$ et/ou $(\partial/\partial\sigma^2)S_n(\hat{\alpha}, \hat{\sigma}^2)$ sera nul p.s.

pour n suffisamment grand d'après la consistance des estimateurs. Dans [262], l'auteur suppose plutôt que l'espace des paramètres est $(0,\infty)\times[\alpha_l,\alpha_u]$ ou $[\sigma_l^2,\sigma_u^2]\times(0,\infty)$.
Rappelons que la variance asymptotique τ_n^2 dépend du choix des points d'observation $\{x_1,\dots,x_n\}$. En revanche, la variance asymptotique du MLE est indépendante du choix du tableau triangulaire des points d'observation [262].
Dans la proposition suivante, nous montrons que la quantité τ_n^2 du Théorème 2.11 est bornée inférieurement et supérieurement ce qui implique que le taux de convergence est toujours en $\sqrt{n}$.

Proposition 2.12. *Pour tout choix de tableau triangulaire de points d'observations $\{x_1,\dots,x_n\}$ pour lesquels (2.32) est satisfait, nous avons*

$$2 \leqslant \liminf_{n\to\infty}\tau_n^2 \leqslant \limsup_{n\to\infty}\tau_n^2 \leqslant 4. \tag{2.35}$$

La variance de la distribution limite de $\hat\sigma_{CV}^2\hat\alpha_{CV}-\sigma_0^2\alpha_0$ peut facilement être estimée. D'après la proposition précédente, cette variance asymptotique est toujours plus grande que celle du MLE. En effet, nous avons $(\sqrt{n}/(\alpha_0\sigma_0^2))(\hat\alpha_{ML}\hat\sigma_{ML}^2-\alpha_0\sigma_0^2)\xrightarrow[n\to+\infty]{\mathcal{L}}\mathcal{N}(0,2)$, voir [262]. Ce fait est tout à fait attendu puisque les estimations par MLE sont généralement les meilleures lorsque le modèle de covariance est bien spécifié, comme c'est le cas ici. Il est possible de vérifier facilement que la grille régulière $\Delta_i\equiv 1/(n-1)$ pour tout $i=2,\dots,n$, ne conduit pas à la variance limite du MLE. Pour cette grille, nous avons $\tau_n^2\underset{n\to+\infty}{\to}3$. Cependant, dans la Proposition 2.13, nous construisons une grille particulier qui conduit à la variance limite du MLE : $\tau_n^2\underset{n\to+\infty}{\to}2$. De plus, nous montrons que les bornes de (2.35) sont précises.

Proposition 2.13. *(i) Soit $\{x_1,\dots,x_n\}$ tel que $x_1=0$, for $i=2,\dots,n-1$,*

$$\Delta_i=\begin{cases}(1-\gamma_n)\frac{2}{n} & \text{si } i \text{ est pair},\\ \frac{2\gamma_n}{n} & \text{si } i \text{ est impair},\end{cases}$$

où $\gamma_n\in(0,1)$ et $\Delta_n=1-\sum_{i=2}^{n-1}\Delta_i$. Alors, en prenant $\gamma_n=1/n$, nous avons $\tau_n^2\underset{n\to+\infty}{\to}4$.

(ii) Soient $\{x_1,\dots,x_n\}$ et $0<\beta<1$ tels que $x_1=0$, $\Delta_i=1/(i\,!)$ pour $i=\lfloor n^\beta\rfloor+1,\dots,n$ et $\Delta_2=\cdots=\Delta_{\lfloor n^\beta\rfloor}\equiv(1-r_n)/(\lfloor n^\beta\rfloor-1)$ avec $r_n=\sum_{i=\lfloor n^\beta\rfloor+1}^{n}\Delta_i$. Alors $\sum_{i=2}^{n}\Delta_i=1$ et $\tau_n^2\underset{n\to+\infty}{\to}2$.

Pour prouver le Théorème 2.11, nous avons suivi le schéma général de la preuve de [262] pour le MLE. Cependant, nos preuves plus techniques contiennent de nouveaux éléments. En particulier, nous avons réalisé des développements de Taylor pour des couples de variables (les variables étant prises à la distance Δ_i) et nous avons eu recours à des théorèmes de la limite centrale pour des variables dépendantes. Nous renvoyons le lecteur à [J13] pour le détail des preuves et des simulations numériques dans lesquelles nous avons illustré la convergence du Théorème 2.11.

2.2.3 Variations quadratiques

Ici, le contexte est un peu différent. Nous ne considérons plus que le processus est stationnaire mais seulement que ses accroissements le sont. Le processus est observé aux temps $j\Delta$ pour $j=1,\dots,n$ avec $\Delta=\Delta_n$ tendant vers 0 lorsque n tend vers l'infini. Son variogramme est alors défini par

$$V(h)=\frac{1}{2}\mathbb{E}\left[(Y(t+h)-Y(t))^2\right]. \tag{2.36}$$

Posons $\Delta = n^{-\alpha}$ où $0 < \alpha \leqslant 1$. Notons que le cas $\alpha = 1$ correspond au cadre asymptotique de remplissage déjà défini plus haut [241]. Nous supposons encore que le processus Y est centré bien que nous ayons étendu les résultats qui suivent au cas non-centré dans la Section 3.4 de [S1].

Nous faisons les hypothèses suivantes.

($\mathcal{H}_0$) Le variogramme V est C^∞ sur $(0, +\infty]$.

($\mathcal{H}_1$) Le variogramme V est $2D$ fois différentiable avec $D \geqslant 0$ et il existe $C > 0$ et $0 < s < 2$ tels que pour tout $h \in \mathbb{R}$, nous avons

$$V^{(2D)}(h) = V^{(2D)}(0) + C(-1)^D |h|^s + r(h), \text{ avec } r(h) = o(|h|^s) \text{ et } |r(h)| \leqslant (Const)\,|h|^s\,. \quad (2.37)$$

($\mathcal{H}_2$) Le reste r dans ($\mathcal{H}_1$) est d fois différentiable en dehors de zéro et $\left|r^{(d)}(h)\right| \leqslant (Const)\,|h|^\beta$ avec $s - d < \beta < -1/2$. Lorsque $s < 3/2$, nous prendrons $d = 2$ et lorsque $s \geqslant 3/2$, $d = 3$.

($\mathcal{H}_3$) $\quad |r(h)| \leqslant (Const)|h|^{s+1/(2\alpha)}$.

Si la fonction de covariance appartient à une famille paramétrique de la forme $\{\sigma^2 k_\alpha; \sigma^2 \geqslant 0, \alpha \in \Theta\}$ où $\Theta \subset \mathbb{R}^p$, alors C est une fonction déterministe des paramètres σ^2 et α. Lorsque $D > 0$, la dérivée D-ème $Y^{(D)}$ en moyenne quadratique de Y est un processus gaussien stationnaire de fonction de covariance k_D donnée par $k_D(h) = \mathrm{Cov}(Y^{(D)}(t), Y^{(D)}(t+h)) = (-1)^{D+1} V^{(2D)}(h)$. Ceci implique que l'exposant de Hölder des trajectoires de $Y^{(D)}$ est $s/2$. Puisque $s < 2$, D est exactement l'ordre de différentiation des trajectoires de Y. Remarquons que dans le cadre asymptotique de remplissage ($\alpha = 1$), ($\mathcal{H}_2$) est quasi minimale. En effet, la condition $\beta < -1/2$ importe peu, puisque plus β est petit, plus la condition est faible. Par exemple, lorsque $s < 3/2$, la dérivée seconde du terme principal est de l'ordre de $|h|^{s-2}$ et nous supposons seulement $\beta > s - 2$. Quelques exemples de processus satisfaisant les hypothèses précédentes sont donnés dans la Section 2.2 de [S1].

Objectif de notre travail et état de l'art

Dans les applications, l'estimation du paramètre C est cruciale. Seule la connaissance de la constante C intervient dans la fonction de covariance et son estimation constitue alors une étape préliminaire nécessaire au krigeage. En effet, par exemple quand D=1, C fournit l'approximation au premier ordre de $\mathbb{E}\left[(Y(t+h) - Y(t))^2\right]$ quand h est petit. Lorsque la fonction de covariance du processus gaussien appartient à une famille paramétrique $\{\sigma^2 k_\alpha; \sigma^2 \geqslant 0, \alpha \in A\}$ où $A \subset \mathbb{R}^p$, C est une fonction de σ^2 et α. Dans ce cas, la plupart des logiciels (comme par exemple `DiceKriging` [218]) utilise le MLE pour estimer (σ^2, α) et donc C (voir [210, 226] pour plus de détails sur le MLE). Malheureusement, le MLE est connu pour être cher computationnellement. De plus, il peut diverger numériquement dans certaines applications pratiques. Enfin, il ne s'applique que lorsque l'ensemble paramétrique des fonctions de covariance est donné, alors que l'estimation de C est également pertinente dans le cas non paramétrique où aucune hypothèse paramétrique n'est faite sur le variogramme V défini par (2.36).

C'est pourquoi nous envisageons comme alternative un estimateur reposant sur les variations quadratiques du processus Y qui ne présuppose aucune hypothèse paramétrique. Les premiers résultats sur les variations quadratiques remontent à Levy dans [164] quantifiant les oscillations du mouvement Brownien. Citons aussi le théorème de Baxter (voir [28], [111, Chap. 5] et [105]). Plus récemment, Guyon et Léon [112] ont montré que si la fonction de covariance $k(h) = \mathrm{Cov}(Y(t+h), Y(t))$ est telle que $k(h) = 1 - |h|^s\, l(h)$ où $0 < s < 2$ et l est une fonction à variation lente en 0, alors, sous des hypothèses techniques additionnelles,

1) si $0 < s < 3/2$, $(V_{H,n}/n)$ défini par

$$V_{H,n} = \sum_{i=1}^{n} H\left(\frac{Y(i/n) - Y((i-1)/n)}{\sqrt{\mathrm{Var}(Y(i/n) - Y((i-1)/n))}}\right) \tag{2.38}$$

converge en loi vers la loi gaussienne à la vitesse $n^{1/2}$;

2) si $3/2 < s < 2$, la limite n'est pas gaussienne et la vitesse est donnée par n^{2-s}.

Coeurjolly [65] a montré que les variations quadratiques sont optimales. Dans [129], les auteurs considèrent des processus gaussiens à accroissements stationnaires (comme dans cette section) et généralisent les résultats sur les variations quadratiques en estimant à la fois l'exposant local de Hölder et le paramètre C. Plus récemment, Lang et Roueff [154] ont généralisé les résultats de [129] et [143] par un estimateur basé sur les incréments dans un contexte semi-paramétrique.

Différences discrètes et variations quadratiques

Dans nos travaux, nous avons considéré une suite non nulle à support fini a de nombres réels dont la somme est nulle. Sans perte de généralité, elle sera notée $a = (a_0, \ldots, a_{L(a)-1})$. Nous dirons qu'elle est d'ordre $M(a)$ si

$$\sum_{j=0}^{L(a)-1} a_j j^k = 0, \quad \text{pour} \quad 0 \leqslant k < M(a) \quad \text{et} \quad \sum_{j=0}^{L(a)-1} a_j j^{M(a)} \neq 0.$$

A chaque suite a, nous associons la différence discrète définie par

$$\Delta_{a,i}(Y) = \sum_{j=0}^{L(a)-1} a_j Y((i+j)\Delta), \quad i = 1, \ldots n - L(a) + 1. \tag{2.39}$$

Par conséquent, $\sum_{j=0}^{L(a)-1} a_j Y(j\Delta)$ est une approximation (au coefficient multiplicatif près) de la dérivée $M(a)$-ème (lorsqu'elle existe) de Y en 0. Il reste maintenant à introduire les variations quadratiques $V_{a,n}$ associées à a et Y définies par

$$V_{a,n} = \sum_{i=1}^{n-L(a)+1} (\Delta_{a,i}(Y))^2. \tag{2.40}$$

En comparaison avec le MLE, l'estimateur basé sur les variations quadratiques présente de nombreux avantages. Tout d'abord, il est plus flexible puisqu'il ne requiert pas que le noyau appartienne à une famille paramétrique, comme c'est le cas pour le MLE. Dans le même ordre d'idée, ajouter un drift n'impose aucune hypothèse supplémentaire (voir Section 3.4 dans [S1]). Du point de vue computationnel, le coût de l'estimateur par variations quadratiques est en $O(n)$ contrairement au MLE pour lequel le coût est en $O(n^3)$. Enfin, dans certains cas pratiques, le MLE peut s'avérer inutilisable car numériquement divergeant (voir Section 5.3 dans [S1]), ce qui ne risque pas de se produire avec l'estimateur par variations.

Résultats principaux

En utilisant l'identité suivante

$$\mathbb{E}[\Delta_{a,i}(Y)\Delta_{a',i'}(Y)] = -\Delta_{a*a',i-i'}(V), \tag{2.41}$$

pour toutes suites a et a', nous avons montré dans la Proposition 5 de [S1] que l'espérance de $V_{a,n}$ est asymptotiquement proportionnelle à la constante C que nous souhaitons estimer. Ceci suggère un estima-

teur $C_{a,n}$ de C basé sur la méthode des moments dont nous avons établi le comportement asymptotique.

Théorème 2.14 (Normalité asymptotique de $C_{a,n}$)**.** *Sous les hypothèses* $(\mathcal{H}_0)$ *à* $(\mathcal{H}_3)$ *et si* $M(a) > D + s/2 + 1/4$*, alors* $C_{a,n}$ *est asymptotiquement gaussien :*

$$\frac{C_{a,n} - C}{\sqrt{\mathrm{Var}(C_{a,n})}} \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0,1), \tag{2.42}$$

avec $\mathrm{Var}(C_{a,n}) = (Const)n^{-1}(1 + o(1))$.

Il est ensuite possible d'étendre ce résultat. Le corollaire suivant est particulièrement intéressant puisqu'il donne des résultats théoriques sur l'agrégation de différentes variations quadratiques.

Corollaire 2.15. *Sous les hypothèses du Théorème 2.14, considérons* k *suites* $a^{(1)}, \ldots, a^{(k)}$ *telles que pour tout* $i = 1, \ldots, k$, $M(a^{(i)}) > D + s/2 + 1/4$*. Suppsons de plus que la matrice de variance-covariance de* $(C_{a^{(i)},n}/\mathrm{Var}(C_{a^{(i)},n})^{1/2})_{i=1,\ldots,k}$ *converge vers une matrice inversible* Γ_∞ *lorsque* $n \to \infty$*. Alors,* $([C_{a^{(i)},n} - C]/\mathrm{Var}(C_{a^{(i)},n})^{1/2})_{i=1,\ldots,k}$ *converge en loi vers une loi* $\mathcal{N}(0, \Gamma_\infty)$.

Nous avons vu numériquement dans [S1] que l'agrégation de suites a semble être très prometteuse puisqu'elle fournit une piste pour régler le problème du choix de la suite a. Le lemme suivant reposant sur [157] ou [27] donne les poids optimaux lorsque l'on procède par agrégation de k suites $a^{(1)}, \ldots, a^{(k)}$ différentes et que l'on considère l'estimateur

$$\sum_{j=1}^{k} \lambda_j C_{a^{(j)},n}.$$

Lemme 2.16. *Supposons que les conditions du Corollaire 2.15 sont vérifiées. Soit* R *la matrice de variance-covariance asymptotique de taille* $k \times k$ *du vecteur de longueur* k *dont les coordonnées sont données par*$(n^{1/2}/C)C_{a^{(j)},n}$*,* $j = 1, \ldots, k$*. Alors, pour tous* $\lambda_1, \ldots, \lambda_k$ *tels que* $\lambda_1 + \cdots + \lambda_k = 1$*, nous avons*

$$(n^{1/2}/C)\left(\sum_{j=1}^{k} \lambda_j C_{a^{(j)},n} - C\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0, \lambda^T R \lambda).$$

Soit $\mathbf{1}_k$ *le vecteur colonne de longueur* k *composé de 1 et définissons*

$$\lambda^* = (\lambda_1^*, \ldots, \lambda_k^*)^\top = \frac{R^{-1}\mathbf{1}_k}{\mathbf{1}_k^T R^{-1} \mathbf{1}_k}.$$

Alors, $\sum_{j=1}^{k} \lambda_j^* = 1$ *et* $\lambda^{*T} R \lambda^* \leqslant \lambda^T R \lambda$.

Afin de valider notre procédure d'agrégation, nous avons comparé la variance asymptotique obtenue à la borne de Cramér-Rao. Plus précisément, dans la Section 4.2 de [S1], nous avons donné deux exemples de familles de processus pour lesquelles nous avons calculé explicitement la borne de Cramér-Rao.

Applications numériques

Dans la Section 5 de [S1], nous avons réalisé quelques simulations pour illustrer numériquement les résultats de convergence présentés ci-dessus. Nous avons montré qu'il est difficile de choisir une séquence optimale. Néanmoins, nous avons montré numériquement qu'agréger plusieurs séquences permet de se rapporcher de la borne de Cramér-Rao.

En outre, nous donnons une procédure pour généraliser l'estimateur par variations quadratiques à la dimension deux. Je présente dans ce manuscrit uniquement cette simulation numérique. Un autre jeu de données en dimension deux a été étudié dans [S1]. Pour cet exemple, la procédure par variations

quadratiques a prouvé une nouvelle fois son intérêt puisque dans ce contexte le MLE ne pouvait pas être implémenté en raison de son coût computationnel.
Venons-en à la simulation. Nous comparons deux méthodes d'estimation de la fonction de covariance d'un modèle gaussien séparable sur un jeu de données de spectroscopie de force atomique [2]. Les données sont constituées d'observations faites sur une grille de $[0,1]^2$ de pas $1/15$. On dispose donc de 256 points de la forme

$$Y(i/15, j/15) \;\; i = 0, \ldots, 15, \;\; j = 0, \ldots, 15.$$

La première méthode considérée est l'estimation par maximum de vraisemblance sur un modèle de krigeage obtenu en utilisant la fonction `km` du package `DiceKriging` de `R` [218]. Dans ce cas, nous avons supposé que la moyenne est donnée par $\mathbb{E}[Y(i/15, j/15)] = \mu$ et que les fonctions de covariance sont exponentielles :

$$\mathrm{Cov}(Y(i/15, j/15), Y(i'/15, j'/15)) = \sigma^2 e^{-\alpha_1 |i-i'|/15} e^{-\alpha_2 |j-j'|/15}. \tag{2.43}$$

Les paramètres $\mu, \sigma^2, \alpha_1, \alpha_2$ sont estimés par maximum de vraisemblance.
La second méthode fait la même hypothèse sur la covariance et procède comme suit.

(1) Nous estimons σ^2 par la variance empirique

$$\hat{\sigma}^2 = \frac{1}{256} \sum_{i,j=0}^{15} (Y(i/15, j/15) - \hat{\mu})^2$$

où $\hat{\mu} = (1/256) \sum_{i,j=0}^{15} Y(i/15, j/15)$.

(2) Pour chaque colonne j de $(Y(i/15, j/15))_{i,j=0,\ldots,15}$, le vecteur des 16 observations suit le modèle avec $s = 1$ et $C_1 = \sigma^2 \alpha_1$. Ainsi, nous pouvons estimer C_1 par $\hat{C}_1$ en faisant la moyenne des $\hat{C}_{1,j}$ pour $j = 0, \ldots, 15$ définis en dimension 1 en utilisant la suite élémentaire $a = (1, -1)$ d'ordre 1.

(3) Nous procédons de même ligne par ligne pour obtenir l'estimateur $\hat{C}_2$ de C_2.

(4) Pour $i = 1, 2$, α_i est estimé par $\hat{\alpha}_i = \hat{C}_i / \hat{\sigma}^2$.

La première méthode, utilisant le maximum de vraisemblance, conduit à des valeurs infinies pour α_1 et α_2, de sorte qu'elle considère les 256 valeurs observées comme complètement indépendantes sur le plan spatial. Par contre, la deuxième méthode donne les valeurs $\hat{\alpha}_1 = 14.72$ et $\hat{\alpha}_2 = 15.73$. Cela correspond à une corrélation d'environ $1/e \approx 0.36$ entre les voisins directs sur la grille. Par conséquent, la deuxième méthode, reposant sur notre estimateur quadratique est capable de détecter une faible corrélation (qui peut être vérifiée graphiquement), contrairement au MLE.

2.2.4 Maximum de vraisemblance sous contraintes d'inégalités

Nous supposons ici encore que la fonction de covariance du processus Y, définie sur $[0,1]^d$, appartient à la famille de fonctions de covariance de la forme $\{\sigma^2 k_\alpha, \; (\sigma^2, \alpha) \in \Theta\}$ où Θ est un compact de $]0, +\infty[\times\mathbb{R}^p$. Maintenant, nous considérons le cas où les trajectoires du processus gaussien Y sont supposées satisfaire des contraintes de bornitude, de monotonicité ou de convexité. En effet, les processus gaussiens avec des contraintes d'inégalités fournissent des modèles de régression appropriés dans les domaines d'application tels que le travail en réseau (monotonicité) [108], l'analyse de réseaux sociaux (monotonicité) [213] et l'économétrie (monotonicité ou positivité) [69]. En outre, il a été démontré que la prise en compte des contraintes peut considérablement améliorer les prévisions et les intervalles prédictifs [76, 108, 213].

2. Communication personnelle de C. Gales et J. M. Senard.

Récemment, un estimateur par maximum de vraisemblance contrainte (cMLE) pour les paramètres de covariance a été suggéré dans [171]. Contrairement au MLE classique (sans contrainte) évoqué ci-dessus, le cMLE prend explicitement en compte les informations supplémentaires apportées par les contraintes d'inégalités. Dans [171], il est montré que la consistance du MLE implique celle du cMLE sous contraintes de bornitude, de monotonicité ou de convexité.

Nous considérons ici encore un tableau triangulaire $\left(x_i\right)_{n\in\mathbb{N},i=1,\dots,n}$ de points d'observation dans $[0,1]^d$, dense dans $[0,1]^d$, *i.e.*, tel que $\sup_{x\in[0,1]^d}\inf_{i=1,\dots,n}|x-x_i|\to 0$ lorsque $n\to\infty$. Nous supposons de plus que l'information $\{Y\in\mathcal{E}_\kappa\}$ est disponible, où $\mathcal{E}_\kappa$ est un ensemble convexe de fonctions défini par des contraintes d'inégalités. Nous considérerons les ensembles suivants

$$\begin{aligned}
\mathcal{E}_0 &= \{f\in\mathcal{C}([0,1]^d,\mathbb{R}) \quad \text{t.q. } \ell\leqslant f(x)\leqslant u,\ \forall x\in[0,1]^d\},\\
\mathcal{E}_1 &= \{f\in\mathcal{C}^1([0,1]^d,\mathbb{R}) \quad \text{t.q. } \partial f(x)/\partial x_i\geqslant 0,\ \forall x\in[0,1]^d, i\in\{1,\dots,d\}\},\\
\mathcal{E}_2 &= \{f\in\mathcal{C}^2([0,1]^d,\mathbb{R}) \quad \text{t.q. } f \text{ est convexe}\},
\end{aligned}$$

correspondant respectivement à la bornitude, à la monotonicité et à la convexité. Pour $\mathcal{E}_0$, les bornes $-\infty\leqslant\ell<u\leqslant+\infty$ sont fixées et connues.

Tout d'abord, nous étudions la distribution asymptotique du MLE (non contraint) conditionné à $\{Y\in\mathcal{E}_\kappa\}$. Cependant, l'inconvénient de cette approche est que l'information contenue dans $\{Y\in\mathcal{E}_\kappa\}$ n'est pas exploitée. Par conséquent, nous étudions ensuite le cMLE introduit dans [171]. Cet estimateur est obtenu en maximisant le logarithme de la vraisemblance de y sous la fonction de covariance σ^2k_α conditionnellement à $\{Y\in\mathcal{E}_\kappa\}$. Le logarithme de cette vraisemblance conditionnelle est donné par

$$\mathcal{L}_{n,c}(\sigma^2,\alpha) = \mathcal{L}_n(\sigma^2,\alpha) - \ln\mathbb{P}_{(\sigma^2,\alpha)}(Y\in\mathcal{E}_\kappa) + \ln\mathbb{P}_{(\sigma^2,\alpha)}(Y\in\mathcal{E}_\kappa|y) \tag{2.44}$$

où $\mathcal{L}_n(\sigma^2,\alpha)$ est la log vraisemblance de y sous la fonction de covariance σ^2k_α donnée en (2.17) et les deux derniers termes dépendent des contraintes d'inégalités. Dans [171], il est prouvé que le cMLE est consistant dès que le MLE l'est. Le but des travaux de [S19] a été de donner le comportement asymptotique du MLE conditionné à $\{Y\in\mathcal{E}_\kappa\}$ et du cMLE.

Estimation du paramètre de variance uniquement

En premier lieu, de même que dans la Section 2.2.1, nous considérons la famille de fonctions de covariance de la forme $\{\sigma^2k_1,\ \sigma^2\in[\sigma_l^2,\sigma_u^2]\}$ où la fonction de covariance k_1 est connue et $0<\sigma_l^2<\sigma_u^2<+\infty$. Nous supposons de plus que le vrai paramètre de variance σ_0^2 est dans $[\sigma_l^2,\sigma_u^2]$. Dans ce cas, la vraisemblance $\mathcal{L}_n(\sigma^2)$ définie en (2.17) s'écrit

$$\mathcal{L}_n(\sigma^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2}\ln(\det(R_{n,1})) - \frac{1}{2\sigma^2}y^\top R_{n,1}^{-1}y,$$

où $R_{n,1}=(k_1(x_i,x_j))_{1\leqslant i,j\leqslant n}$. Le MLE $\hat\sigma^2_{ML}$ donné par

$$\hat\sigma^2_{ML}\in\operatorname*{argmax}_{\sigma^2>0}\mathcal{L}_n(\sigma^2)$$

est alors asymptotiquement gaussien :

$$\sqrt{n}(\hat\sigma^2_{ML}-\sigma_0^2)\to\mathcal{N}(0,2\sigma_0^4).$$

Dans un premier temps, nous avons montré que la distribution asymptotique de $\hat\sigma^2_{ML}$ n'est pas affectée par le conditionnement par $\{Y\in\mathcal{E}_\kappa\}$.

Avant d'énoncer le résultat, précisons les notations. Pour toute suite de v.a. ou vecteurs aléatoires $(X_n)_{n\in\mathbb{N}}$ de $\mathbb{R}^l$ ($l \geqslant 1$) qui dépendent de Y et pour toute loi de probabilité μ sur $\mathbb{R}^l$, nous écrirons

$$X_n \xrightarrow[n\to+\infty]{\mathcal{L}|Y\in\mathcal{E}_\kappa} \mu$$

si, pour tout fonction continue bornée $g : \mathbb{R}^l \to \mathbb{R}$,

$$\mathbb{E}[g(X_n)|Y \in \mathcal{E}_\kappa] \underset{n\to+\infty}{\to} \int_{\mathbb{R}^l} g(x)\mu(\mathrm{d}x).$$

Théorème 2.17 (Normalité asymptotique du MLE conditionné pour la variance)**.** *Supposons que le noyau k et la suite des points d'observation vérifient des conditions techniques données dans [S19]. Alors pour $\kappa = 0, 1, 2$,*

$$\sqrt{n}\left(\hat{\sigma}^2_{ML} - \sigma_0^2\right) \xrightarrow[n\to+\infty]{\mathcal{L}|Y\in\mathcal{E}_\kappa} \mathcal{N}(0, 2\sigma_0^4).$$

Les hypothèses sur k et sur le plan d'échantillonnage ne sont pas en réalité très contraignantes. Nous montrons ensuite que le cMLE, donné par

$$\hat{\sigma}^2_{cML} \in \operatorname*{argmax}_{\sigma^2>0} \mathcal{L}_{n,c}(\sigma^2)$$

où $\mathcal{L}_{n,c}$ est défini dans (2.44), a le même comportement asymptotique que le MLE conditionné à $\{Y \in \mathcal{E}_\kappa\}$.

Théorème 2.18 (Normalité asymptotique du cMLE pour la variance)**.** *Supposons que le noyau k et la suite des points d'observation vérifient des conditions techniques données dans [S19]. Alors pour $\kappa = 0, 1, 2$,*

$$\sqrt{n}\left(\widehat{\sigma}^2_{n,c} - \sigma_0^2\right) \xrightarrow[n\to+\infty]{\mathcal{L}|Y\in\mathcal{E}_\kappa} \mathcal{N}(0, 2\sigma_0^4).$$

Estimation jointe pour le modèle de Matérn

Nous supposons maintenant que $d = 1$, 2 ou 3 et nous considérons la famille de fonctions de covariance isotropiques de Matérn sur $[0, 1]^d$. Voir par exemple [241] pour plus de détails. Ici $k_{\sigma^2,\alpha} = k_{\sigma^2,\rho,\nu}$ est donné par, pour tous $x, x' \in [0, 1]^d$,

$$k_{\sigma^2,\rho,\nu}(x, x') = \sigma^2 K_\nu\left(\frac{\|x - x'\|}{\rho}\right) = \frac{\sigma^2}{\Gamma(\nu)2^{\nu-1}}\left(\frac{\|x - x'\|}{\rho}\right)^\nu \kappa_\nu\left(\frac{\|x - x'\|}{\rho}\right).$$

Le paramètre $\sigma^2 > 0$ est simplement la variance du processus, $\rho > 0$ est un paramètre de longueur de corrélation qui contrôle la vitesse de décroissance de la fonction de covariance en fonction de la distance et $\nu > 0$ est le paramètre de régularité du processus. La fonction κ_ν est la fonction de Bessel modifiée de seconde espèce d'ordre ν [2]. Nous supposons par la suite que le paramètre de régularité ν est connu. Il s'agit donc d'estimer (σ^2, ρ) et $p = 2$.

La vraisemblance définie par (2.17) pour σ^2 et ρ sous le modèle de Matérn en supposant connu ν s'écrit

$$\mathcal{L}_n(\sigma^2, \rho) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2}\ln(\det(R_{n,\rho,\nu})) - \frac{1}{2\sigma^2}y^\top R^{-1}_{n,\rho,\nu}y, \tag{2.45}$$

où $R_{n,\rho,\nu} = (K_\nu(\|x_i - x_j\|/\rho))_{1\leqslant i,j\leqslant n}$. Posons $\Theta = [\sigma_l^2, \sigma_u^2] \times [\rho_l, \rho_u]$ où $0 < \sigma_l^2 < \sigma_u^2 < \infty$ et $0 < \rho_l < \rho_u < \infty$. Supposons de plus que les vrais paramètres sont tels que $\sigma_l^2/(\rho_l^{2\nu}) < \sigma_0^2/(\rho_0^{2\nu}) < \sigma_u^2/(\rho_u^{2\nu})$. Alors le MLE est donné par

$$(\widehat{\sigma}^2_n, \widehat{\rho}_n) \in \operatorname*{argmax}_{(\sigma^2,\rho)\in\Theta} \mathcal{L}_n(\sigma^2, \rho). \tag{2.46}$$

Rappelons qu'il est montré dans [264] que les paramètres σ_0^2 et ρ_0 ne peuvent pas être estimés de façon consistante contrairement au paramètre microergodique $\sigma_0^2/\rho_0^{2\nu}$. De plus, [141] a montré que $\sqrt{n}\left(\widehat{\sigma}_n^2/\widehat{\rho}_n^{2\nu} - \sigma_0^2/\rho_0^{2\nu}\right)$ converge vers une loi $\mathcal{N}\big(0, 2\left(\sigma_0^2/\rho_0^{2\nu}\right)^2\big)$. Nous établissons dans le théorème suivant que cette normalité asymptotique est encore satisfaite conditionnellement à $\{Y \in \mathcal{E}_\kappa\}$.

Théorème 2.19 (Normalité asymptotique du MLE conditionné et du cMLE pour le modèle de Matérn)**.** *Supposons que le noyau k et la suite des points d'observation vérifient des conditions techniques données dans [S19]. Alors pour $\kappa = 0, 1, 2$,*

$$\sqrt{n}\left(\frac{\widehat{\sigma}_n^2}{\widehat{\rho}_n^{2\nu}} - \frac{\sigma_0^2}{\rho_0^{2\nu}}\right) \xrightarrow[n\to+\infty]{\mathcal{L}|Y\in\mathcal{E}_\kappa} \mathcal{N}\left(0, 2\left(\frac{\sigma_0^2}{\rho_0^{2\nu}}\right)^2\right)$$

et

$$\sqrt{n}\left(\frac{\widehat{\sigma}_{n,c}^2}{\widehat{\rho}_{n,c}^{2\nu}} - \frac{\sigma_0^2}{\rho_0^{2\nu}}\right) \xrightarrow[n\to+\infty]{\mathcal{L}|Y\in\mathcal{E}_\kappa} \mathcal{N}\left(0, 2\left(\frac{\sigma_0^2}{\rho_0^{2\nu}}\right)^2\right).$$

Les preuves techniques des résultats précédents font intervenir des outils issus de la statistique spatiale asymptotique, des extrema de processus gaussiens et des espaces de Hilbert à noyaux reproduisants.

Nous avons aussi généralisé les résultats précédents à l'estimation du paramètre microergodique de fonctions de covariance isotropiques de Wendland. Dans ce contexte, $k_\theta = k_{\theta,s,\mu}$, avec $\theta = (\sigma^2, \rho)$, est donné par

$$k_{\theta,s,\mu}(x) = \sigma^2 \phi_{s,\mu}\left(\frac{\|x\|}{\rho}\right),$$

pour $x \in \mathbb{R}^d$ avec, pour $t \geqslant 0$,

$$\phi_{s,\mu}(t) = \begin{cases} \frac{1}{B(2s,\mu+1)} \int_{\|x\|}^1 u(u^2 - \|x\|^2)^{s-1}(1-u)^\mu du & \text{si} \quad \|x\| < 1, \\ 0 & \text{sinon.} \end{cases}$$

Les paramètres $s > 0$ et $\mu \geqslant (d+1)/2+s$ sont supposés fixés et connus. Le paramètre s traduit la régularité des fonctions de covariance de Wendland, de même que pour les fonctions de covariance de Matérn [30]. Les paramètres $\sigma^2 > 0$ et $\rho > 0$ s'interprètent aussi de la même façon que pour les finctions de covariance de Matérn et doivent être estimés. Remarquons que sous certaines relations sur les paramètres ν, s and μ, les mesures gaussiennes obtenues à partir des fonctions de covariance de Matérn et de Wendland sont équivalentes [30]. Notons que la fonction de covariance de Wendland est à support compact, ce qui constitue un avantage computationel notable [30].
Dans [30], il est montré que les paramètres σ_0^2 et ρ_0 ne peuvent pas être estimés de façon consistante. En revanche, le paramètre σ_0^2/ρ_0^{1+2s} est microergodique. Dans [P19], nous avons établi l'analogue du Théorème 2.19 pour l'estimation du paramètre microergodique des fonctions de covariance de Wendland.

Applications numériques

Les résultats asymptotiques énoncés dans cette section ont été testés numériquement. Pour ce faire, nous avons considéré la fonction de covariance de Matérn 5/2 et deux cas de figure : ρ_0 connu et inconnu. Les codes ont été implémentés en utilisant le package R `LineqGPR` [170]. Nous observons numériquement dans les deux cas que pour de grandes tailles d'échantillon les distributions empiriques du cMLE et du MLE conditionné sont très similaires. Par contre, pour des tailles d'échantillon moyennes, la convergence du cMLE est plus rapide que celle du MLE conditionné et fournit donc des résultats plus précis. Ainsi, prendre en compte les contraintes semble être pertinent et profitable.

Résultats sur la prédiction

Nous nous sommes aussi intéressés à la prédiction. Nous avons montré que, conditionnellement aux contraintes d'inégalité, les prédictions obtenues en prenant en compte les contraintes sont asymptotiquement égales aux prédictions standard par kriegage (non contraintes). Il en va de même pour la comparaison des écarts conditionnels obtenus avec et sans prise en compte des contraintes.

En outre, lorsque que le processus n'est pas contraint, des résultats significatifs sur l'utilisation de fonctions de covariance mal spécifiées qui sont asymptotiquement équivalentes à la vraie ont été obtenus dans [237, 238, 240]. Plus précisément, les prédictions et les variances conditionnelles obtenues à partir de mesures gaussiennes équivalentes sont asymptotiquement équivalentes. Nous avons montré que cette équivalence reste vraie lorsque les prédictions et les variances conditionnelles sont calculées en prenant en compte les contraintes d'inégalités.

Enfin, une question importante pour les processus gaussiens consiste à évaluer l'exactitude asymptotique des prévisions obtenues à partir des paramètres de covariance estimés (éventuellement de façon uniforme). Nous avons limité notre étude à l'analyse asymptotique de la prédiction à des paramètres de covariance fixes (éventuellement mal spécifiés).

Lorsqu'aucune contrainte n'est prise en compte et que nous nous plaçons dans le cadre asymptotique par expansion, les prédictions obtenues à partir d'estimateurs consistants des paramètres de covariance sont généralement optimales asymptotiqueùent [18, 19]. Dans le cadre asymptotique par remplissage, sans tenir compte des contraintes, les prédictions obtenues à partir des estimateurs des paramètres de covariance peuvent être asymptotiquement égales à celles obtenues à partir des paramètres de covariance vraie [205]. Il serait intéressant, dans de futurs travaux, d'étendre les résultats donnés dans [205], au cas des contraintes d'inégalité. Cela pourrait se faire en uniformisant les preuves des résultats précédents sur les sous-espaces des paramètres de covariance.

Extension aux processus discontinus et aux observations bruitées

Les résultats précédents sont valables pour les processus gaussiens continus observés exactement. Il est donc naturel de se demander si des résultats similaires sont envisageables pour les processus gaussiens discontinus ou pour les processus gaussiens observés avec des erreurs.

Tout d'abord, nous avons considéré le modèle standard de processus gaussien discontinu avec un effet pépite de la forme :

$$Y = Y_c + Y_\delta,$$

où Y_c est un processus gaussien continu sur $[0,1]^d$ et Y_δ est un processus gaussien sur $[0,1]^d$ de fonction moyenne nulle et de fonction de covariance k_δ donné par

$$k_\delta(u,v) = \delta \mathbb{1}_{\{u=v\}},$$

pour $u, v \in [0,1]^d$. De plus, nous supposons que Y_c et Y_δ sont indépendants. Nous avons montré qu'il a une probabilité nulle de satisfaire les contraintes de bornitude. Par conséquent, il ne semble pas possible de définir, de manière significative, un processus gaussien discontinu conditionné par des contraintes de bornitude. Ce résultat peut être étendu aux contraintes de monotonicité et de convexité.

Dans le cas des observations bruitées, il semble difficile mais intéressant d'obtenir des résultats asymptotiques dans le cadre asymptotique par remplissage sur le MLE (non contraint) des paramètres de covariance et de la variance du bruit. À notre connaissance, les seuls modèles de covariance qui ont été étudiés théoriquement, dans le cadre asymptotique par remplissage et avec des erreurs de mesure, sont le mouvement Brownien [239] et le modèle exponentiel [61, 56]. Nous nous sommes intéressés au modèle

exponentiel pour lequel nous disposons de n observations données par

$$y_i = Y(x_i) + \epsilon_i,$$

pour $i = 1, \dots, n$ où $(x_1, \dots, x_n) = (0, 1/(n-1), \dots, 1)$, les $\epsilon_1, \dots, \epsilon_n$ sont indépendants, indépendants de Y, et sont distribués selon la loi $\mathcal{N}(0, \delta_0^2)$.
Dans [61], il est montré que le MLE $\widehat{\sigma}_n^2/\widehat{\rho}_n$ du paramètre microergodique et le MLE $\widehat{\delta}_n^2$ de la variance du bruit satisfont conjointement le théorème de la limite centrale :

$$\begin{pmatrix} n^{1/4}\left(\widehat{\sigma}_n^2/\widehat{\rho}_n - \sigma_0^2/\rho_0\right) \\ n^{1/2}\left(\widehat{\delta}_n^2 - \delta_0^2\right) \end{pmatrix} \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 4\sqrt{2}\delta_0(\sigma_0^2/\rho_0)^{3/2} & 0 \\ 0 & 2\delta_0^4 \end{pmatrix}\right). \tag{2.47}$$

Par conséquent, le taux de convergence du MLE du paramètre microergodique est réduit de $n^{1/2}$ à $n^{1/4}$, en raison des erreurs de mesure. Le taux de convergence du MLE de l'écart de bruit, quant à lui, est de $n1/2$. Dans [P19], nous avons montré que ces taux sont inchangés par conditionnement par $\{Y \in \mathcal{E}_0\}$.
Il serait intéressant de voir si le théorème de limite centrale (2.47) reste valable conditionnellement à $\{Y \in \mathcal{E}_0\}$. Ce serait une extension au cas bruité des Théorèmes 2.17 et 2.19. Néanmoins, pour prouver le Théorème 2.17, nous avons observé que dans le cas non bruité, le MLE de σ_0^2 est une somme normalisée des variables indépendantes $W_{n,1}^2, \dots, W_{n,n}^2$, avec

$$W_{n,i} := \frac{y_i - \mathbb{E}[y_i | y_1, \dots, y_{i-1}]}{\sqrt{\mathrm{Var}(y_i | y_1, \dots, y_{i-1})}},$$

pour $i = 1, \dots, n$. Nous avons tiré profit du fait que le conditionnement par $W_{n,1}$,...,$W_{n,k}$ permet de conditionner par $Y(x_1)$,...,$Y(x_k)$ et conditionner approximativement par l'événement $\{Y \in \mathcal{E}_0\}$, tout en laissant la distribution de $W_{n,k+1}$,...,$W_{n,n}$ inchangée.
En revanche, dans le cas bruité, les auteurs de [61] montrent que le MLE de σ_0^2/ρ_0 est également une somme normalisée des variables indépendantes $W_{n,1}^2$,...,$W_{n,n}^2$, mais chaque $W_{n,i}$ dépend du vecteur complet d'observations $y = (y_1, ldots, y_n)$ (voir [61, Equations (3.40) et (3.42)]). Par conséquent, il semble beaucoup plus difficile de s'attaquer à la normalité asymptotique du MLE de σ_0^2/ρ_0 et δ_0^2, conditionné à $\{Y \in \mathcal{E}_0\}$. Nous laissons cette question ouverte aux travaux futurs.
La vraisemblance conditionnelle et le cMLE peuvent être naturellement étendus au cas bruité. Néanmoins, l'étude asymptotique du cMLE, dans le contexte de la fonction de covariance exponentielle, semble nécessiter un travail supplémentaire substantiel. En effet, pour analyser le MLE dans le cas non bruité pour les fonctions de covariance Matérn, nous nous sommes appuyés sur les résultats de [141] et [255], qui sont spécifiques au cas non bruité. De plus, les arguments de martingale, utilisés dans la preuve du théorème 2.18, nécessitent que les points d'observation soient extraits d'une suite. Par conséquent, ces arguments de martingale ne sont pas disponibles dans ce cadre, dans lequel les points d'observation sont pris sur des grilles régulières. Enfin, les arguments RKHS, utilisés dans la preuve du Théorème 2.19, nécessitent de travailler avec des fonctions de covariance qui sont au moins deux fois différenciables, ce qui n'est pas le cas avec les fonctions de covariance exponentielle.

2.3 Quelques éléments sur le maximum d'excursions browniennes

Dans cette section, je présente un travail probabiliste mené avec Sabine Mercier (IMT-UT2J), Pierre Vallois (Université de Lorraine, Nancy) et pour partie avec Claudie Chabriac (IMT-UT2J). Cette collaboration a abouti à la publication des quatre articles : [J8], [J9], [J11] et [J16]. Les résultats présentés

dans cette section sont développés dans ces articles.

Etat de l'art

Le score local est un outil probabiliste souvent utilisé par les biologistes pour étudier les séquences d'acides aminés ou de nucléotides comme l'ADN. En particulier, ses propriétés permettent de déterminer le segment le plus significatif dans une séquence donnée, voir par exemple [139, 257]. A chaque position i de la séquence, on associe une variable aléatoire notée ϵ_i qui représente un score. Par exemple, ϵ_i peut mesurer la propriété physique ou chimique du i-ème acide aminé ou du i-ème nucléotide de la séquence. Il peut également coder la similarité entre deux composants de deux séquences. On suppose que $(\epsilon_i)_{i\geqslant 1}$ est une suite de v.a. i.i.d. de même loi que ε. Il convient ensuite d'introduire la suite des scores cumulés jusqu'à la position n :

$$S_n := \epsilon_1 + \ldots + \epsilon_n \quad \text{pour } n \geqslant 1 \; ; \quad S_0 = 0. \tag{2.48}$$

Il est clair que $(S_n)_n$ est une marche aléatoire issue de 0 et à incréments indépendants. Définissons, de plus,

$$\underline{S}_n := \min_{0\leqslant i\leqslant n} S_i, \quad n \geqslant 0, \tag{2.49}$$

ainsi que

$$U_n := S_n - \underline{S}_n = S_n - \min_{i\leqslant n} S_i, \quad n \geqslant 0, \tag{2.50}$$

et

$$\overline{U}_n := \max_{0\leqslant k\leqslant n} U_k, \quad n \geqslant 0. \tag{2.51}$$

Les deux derniers processus jouent un rôle important dans l'étude des séquences biologiques. Le premier, toujours positif, est appelé processus de Lindley (voir, e.g., [10, Chap. III] ou [42, Chap. I] pour les propriétés de ce processus). Le score local $\overline{U}_n$ est simplement le supremum du processus de Lindley jusqu'au temps n. Les biologistes moléculaires s'intéressent aux grandes valeurs "inattendues" de U_n [257] et mettent ainsi en exergue les séquences dites *atypiques*.

La distribution exacte de $\overline{U}_n$ a été déterminée dans [180], en utilisant l'exponentiation d'une matrice appropriée et des outils classiques de la théorie des chaînes de Markov. Le résultat est le suivant :

$$\mathbb{P}(\overline{U}_n \geqslant a) = (1, 0, \ldots, 0) \cdot \Pi^n \cdot (0, \ldots, 0, 1)\top, \tag{2.52}$$

où Π est une matrice carrée de taille $(a+1)$ dépendant de la distribution de ε. La formule donnée dans [180] est valable quel que soit le signe de $\mathbb{E}[\epsilon]$ mais, numériquement, elle ne peut être appliquée que pour de courtes séquences. Cependant, en pratique, on est souvent confrontés à de longues séquences ayant généralement un score moyen négatif, $\mathbb{E}[\epsilon] < 0$. Sous cette hypothèse, le score local $\overline{U}_n$ croît en $\ln n$ [256] et une approximation asymptotique de la distribution de $\overline{U}_n$ lorsque n est grand a été donnée dans [139] et [87] en utilisant la théorie du renouvellement :

$$\lim_{n\to\infty} \mathbb{P}\left(\overline{U}_n \leqslant a + \frac{\ln n}{\lambda}\right) = \exp\big\{ - K^* \cdot e^{-\lambda a}\big\}, \tag{2.53}$$

où λ est l'unique racine positive de $\mathbb{E}[e^{x\varepsilon}] = 1$ et K^* dépend de la loi de ε. Dans le cas de la comparaison de deux séquences, une généralisation de ce résultat a été implémentée dans le logiciel BLAST [6]. Les auteurs de [179] ont proposé une autre approximation de $\mathbb{P}(\overline{U}_n \leqslant a + \ln n/\lambda)$ s'exprimant comme une somme et dont le premier terme donne la limite de Karlin *et al.*

Lorsque $\mathbb{E}[\epsilon] = 0$, le comportement asymptotique de la queue de $\overline{U}_n$ a été déterminé dans [81] :

$$\mathbb{P}(\overline{U}_n \leqslant a\sqrt{n}) \underset{n\to\infty}{\sim} \frac{2}{\pi} \sum_{k\in\mathbb{Z}} \frac{(-1)^k}{2k+1} \exp\Big\{ - \frac{(2k+1)^2\pi^2}{8a^2} \Big\}, \tag{2.54}$$

et la vitesse de convergence est donnée dans [92].

Enfin quel que soit le signe de $\mathbb{E}[\varepsilon]$, [245] propose, pour les grandes valeurs de a, l'approximation suivante

$$\mathbb{P}(\overline{U}_n \geqslant a\sqrt{n}) \underset{n\to\infty}{\sim} 2\sqrt{\frac{2}{\pi}} \frac{\sigma}{a} \exp\Big\{ - \frac{(\delta_n - a)^2}{2\sigma^2} \Big\}, \tag{2.55}$$

où $\delta_n = \sqrt{n} \cdot \mathbb{E}[\varepsilon]$ et $\sigma = \sqrt{\mathrm{Var}(X)}$.

Motivation

Remarquons maintenant que le score local peut aussi se réécrire de la façon suivante :

$$\overline{U}_n = \max_{0\leqslant i\leqslant j\leqslant n} \sum_{k=i}^{j} \varepsilon_k \tag{2.56}$$

La longueur θ_n du (dernier) segment qui réalise le score local est alors définie comme $j - i$ où i et j sont les deux entiers qui réalisent le maximum dans (2.56). Cette variable aléatoire est également d'intérêt (voir [9]) et son comportement asymptotique lorsque n tend vers l'inifni a été déterminé dans [86, 87] dans le cas où $\mathbb{E}[\varepsilon] < 0$. Dans un autre contexte, les auteurs de [140] ont établi une loi limite pour la longueur des mots communs parmi un ensemble de séquences aléatoires. Dans [211], un résultat sur la distribution de la plus longue correspondance entre deux séquences a été donné.

Comme dit précédemment, les biologistes utilisent la variable $\overline{U}_n$ pour déterminer les séquences atypiques, *i.e.* celles ayant un score local élevé et innattendu. Cependant, à la lumière des illustrations numériques qui suivent, il semble primordial de prendre en compte à la fois le score local et sa longueur pour déterminer les séquences atypiques. La connaissance de la distribution conjointe du couple $(\overline{U}_n, \theta_n)$ devrait également permettre aux tests statistiques bidimensionnels associés d'être plus puissants que les tests habituels basés uniquement sur la première composante $\overline{U}_n$. Pour justifier ces affirmations et motiver notre étude théorique, nous avons réalisé quelques illustrations numériques présentées dans [J11].

Illustrations numériques

Tout d'abord, nous considérons les 606 sequences de la base de données Structural Classification of Proteins (SCOP2) [3] et utilisons le score hydrophobique défini dans [130]. Pour toute séquence i ($1 \leqslant i \leqslant 606$), n_i représente sa longueur, u_{n_i} son score local et ℓ_{n_i} la longueur qui réalise ce score.

D'une part, la probabilité $\mathbb{P}(\overline{U}_{n_i} \geqslant u_{n_i})$ est calculée pour tout i via la méthode exacte basée sur (2.52). Nous déterminons ensuite les dix séquences $i_1, \ldots, i_{10}$ ayant les plus faibles probabilités pour le score local :

$$\mathbb{P}(\overline{U}_{n_{i_1}} \geqslant u_{n_{i_1}}) \leqslant \ldots \leqslant \mathbb{P}(\overline{U}_{n_{i_{10}}} \geqslant u_{n_{i_{10}}}).$$

Les caractéristiques des dix séquences sont résumées dans la Table 2.1. Trois séquences, d'indices i_{11}, i_{12} et i_{13}, sont ajoutées en bas de la Table 2.1.

D'autre part, pour tout k ($1 \leqslant k \leqslant 10$), nous générons $N = 10^5$ séquences i.i.d. de longueurs n_{i_k} afin d'approximer $\mathbb{P}(\overline{U}_{n_{i_k}} \geqslant u_{n_{i_k}}, \theta_{n_{i_k}} \leqslant \ell_{n_{i_k}})$. Nous avons ensuite ordonné $\mathbb{P}(\overline{U}_{n_{i_k}} \geqslant u_{n_{i_k}}, \theta_{n_{i_k}} \leqslant \ell_{n_{i_k}})$ pour tout $1 \leqslant k \leqslant 13$ et indiqué le rang de chaque séquence dans la dernière colonne de la Table 2.1.

3. SCOP2 : CF scop2dom 20140205aa. *http ://scop2.mrc-lmb.cam.ac.uk/downloads/*

n_i	u_{n_i}	ℓ_{n_i}	Proba. sur $\overline{U}_n$	Rang score seul	Estimation proba. sur $(\overline{U}_n, \theta_n)$	Rang couple score et longueur
173	185	169	10^{-6}	1	$.< 10^{-6}$	1
103	106	88	$3.13 \cdot 10^{-4}$	2	$5 \cdot 10^{-5}$	2
80	93	76	$4.17 \cdot 10^{-4}$	3	$3.10 \cdot 10^{-4}$	4
94	100	85	$4.03 \cdot 10^{-4}$	4	$2.50 \cdot 10^{-4}$	3
93	88	86	$1.68 \cdot 10^{-4}$	5	$1.24 \cdot 10^{-3}$	5
111	82	107	$6.41 \cdot 10^{-3}$	6	$5.81 \cdot 10^{-3}$	9
129	76	127	$1.75 \cdot 10^{-2}$	7	$1.69 \cdot 10^{-2}$	13
227	93	102	$1.84 \cdot 10^{-2}$	8	$2.94 \cdot 10^{-3}$	8
145	73	130	$2.98 \cdot 10^{-2}$	9	$2.64 \cdot 10^{-2}$	12
109	67	79	$2.56 \cdot 10^{-2}$	10	$1.37 \cdot 10^{-2}$	11
113	49	22	$1.26 \cdot 10^{-1}$	33	$1.37 \cdot 10^{-3}$	6
133	44	18	$2.28 \cdot 10^{-1}$	67	$1.53 \cdot 10^{-3}$	7
227	40	19	$4.96 \cdot 10^{-1}$	192	$8.21 \cdot 10^{-3}$	10

TABLE 2.1 – Dix séquences les plus significatives pour le score seul pour les données SCOP2 (haut) et trois séquences faussement négatives (bas). "Proba. sur $\overline{U}_n$" donne la valeur de $\mathbb{P}(\overline{U}_{n_i} \geqslant u_{n_i})$, "Estimation proba sur $(\overline{U}_n, \theta_n)$" est une estimation de $\mathbb{P}(\overline{U}_{n_i} \geqslant u_{n_i}, \theta_{n_i} \leqslant \ell_{n_i})$.

Nous remarquons que le classement obtenu en prenant en compte le score local et sa longueur diffère du premier pour lequel seul le score local a été pris en compte.

Nous nous intéressons dans un second temps aux séquences faussement négatives. Selon la signification statistique basée sur la p-valeur, une séquence avec $\overline{U}_n = u$ et $\theta_n = \ell$ est dite α-faux-négative si elle est α-$(\overline{U}_n, \theta_n)$ significative mais pas α-$\overline{U}_n$ significative, *i.e.*,

$$\mathbb{P}(\overline{U}_n \geqslant u) > \alpha > \mathbb{P}(\overline{U}_n \geqslant u, \theta_n \leqslant \ell),$$

pour un niveau donné $\alpha \in [0, 1]$.
Nous approchons maintenant $\mathbb{P}(\overline{U}_n \geqslant u, \theta_n \leqslant \ell)$ et $\mathbb{P}(\overline{U}_n \geqslant u)$ pour différentes valeurs de u et ℓ par un schéma Monte-Carlo. La simulation d'un échantillon de N séquences de longueur n donne lieu à $(u_{n,i}, \ell_{n,i})_{1 \leqslant i \leqslant N}$. Naturellement, $\mathbb{P}(\overline{U}_n \geqslant u)$ et $\mathbb{P}(\overline{U}_n \geqslant u, \theta_n \leqslant \ell))$ sont estimés par

$$\frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{\{u \geqslant u_{n,i}\}} \quad \text{et} \quad \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{\{u \geqslant u_{n,i}, \ell \leqslant \ell_{n,i}\}}.$$

Nous avons représenté dans la Figure 2.6 le nuage de points $(\mathbb{P}(\overline{U}_n \geqslant u_{n,i}), \mathbb{P}(\overline{U}_n \geqslant u_{n,i}, \theta_n \leqslant \ell_{n,i}))$ pour $i = 1, \ldots, N$. Puisque $\mathbb{P}(\overline{U}_n \geqslant u) \geqslant \mathbb{P}(\overline{U}_n \geqslant u, \theta_n \leqslant \ell)$ alors tous ces points sont naturellement au-dessus de la première bissectrice. Nous pouvons voir d'une part qu'il y a un grand pourcentage de faux-négatifs (environ 1/3 des séquences pour $\alpha = 5\%$ pour différentes valeurs de n) et d'autre part, que les probabilités du score local seul et du couple peuvent être très différentes (beaucoup de points sont loin de la première bissectrice).
Ces résultats et conclusions motivent le travail qui va suivre concernant l'étude théorique de la distribution du couple $(\overline{U}_n, \theta_n)$.

2.3.1 Cadre de notre étude

Dans nos travaux, nous avons considéré uniquement le cas où les variables aléatoires $(\epsilon_i)_{i \geqslant 1}$ sont centrées et réduites. Il est clair que la trajectoire de U_n peut alors être décomposée en une succession de 0 et

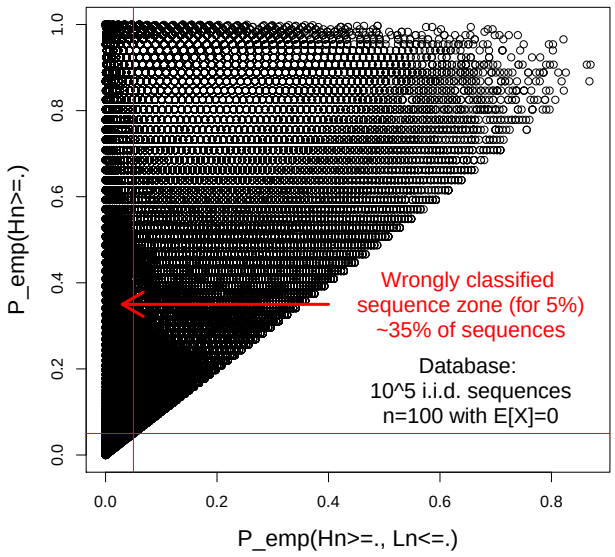


FIGURE 2.6 – Chaque point est associé à l'une des $N = 10^5$ séquences simulées de longueur commune $n = 100$ et ses coordonnées sont données par $\mathbb{P}(\overline{U}_n \geqslant u_{n,i}, \theta_n \leqslant \ell_{n,i})$ (en abscisses) et $\mathbb{P}(\overline{U}_n \geqslant u_{n,i})$ (en ordonnées).

d'excursions supérieures à 0. Ici pour des raisons techniques, nous ne considérons que les excursions complètes jusqu'à un temps fixé. Cela nous amène à introduire le maximum U_n^* des hauteurs de toutes les excursions complètes jusqu'à n. La deuxième variable qui joue un rôle important est θ_n^*, le temps nécessaire pour atteindre la hauteur maximale U_n^*, pendant de θ_n pour l'ensemble des excursions.

Les illustrations numériques précédentes ont montré que la connaissance de la distribution jointe de (U_n^*, θ_n^*) est capitale. Malheureusement, il est difficile de déterminer explicitement cette loi pour un n donné. Cependant, en pratique, les séquences biologiques sont généralement longues et il est donc pertinent d'étudier la distribution asymptotique de (U_n^*, θ_n^*). Le théorème de convergence de Donsker établit alors que la marche aléatoire initiale $(S_k)_{0 \leqslant k \leqslant n}$ normalisée par le facteur $1/\sqrt{n}$ converge en distribution lorsque $n \to \infty$ vers le mouvement Brownien (MB) $(B(s),\ 0 \leqslant s \leqslant 1)$. Il est alors facile d'en déduire que le processus de Lindley normalisé $\left(U_k/\sqrt{n}\right)_{0 \leqslant k \leqslant n}$ peut être approximé par $(\widehat{U}(s),\ 0 \leqslant s \leqslant 1)$ où :

$$\widehat{U}(t) := B(t) - \inf_{0 \leqslant s \leqslant t} B(s) \overset{(d)}{=} |B(t)|, \quad t \geqslant 0. \tag{2.57}$$

Il s'ensuit que le comportement asymptotique de (U_n^*, θ_n^*) pour n grand est étroitement lié à la distribution de $\left(U^*(1), \theta^*(1)\right)$ où $U^*(1)$ et $\theta^*(1)$ sont les analogues en temps continu de U_n^* et θ^*. Par conséquent, la connaissance de la loi de (U_n^*, θ_n^*) pour les n grands revient à l'étudie de $\left(U^*(1), \theta^*(1)\right)$.

Notations

Définissons maintenant précisément les variables d'intérêt dans le cadre continu. Soit $(B(t),\ t \geqslant 0)$ le MB standard issu de 0 et $U(t)$ le MB réfléchi :

$$U(t) := |B(t)|, \quad t \geqslant 0. \tag{2.58}$$

L'excursion (au-dessus de 0) enjambant t démarre de $g(t)$ et se termine à $d(t)$, où

$$g(t) := \sup\{s \leqslant t,\ U(s) = 0\}, \quad d(t) := \inf\{s \geqslant t,\ U(s) = 0\}, \quad t \geqslant 0. \tag{2.59}$$

Soit $\overline{U}(t)$ le supremum de U sur $[0, t]$:

$$\overline{U}(t) := \sup_{0 \leqslant s \leqslant t} U(s), \quad t \geqslant 0. \tag{2.60}$$

Alors la plus grande valeur $U^*(t)$ sur toutes les excursions complètes du processus $(U(r),\ 0 \leqslant r \leqslant t)$ est définie par

$$U^*(t) := \overline{U}(g(t)) = \sup_{0 \leqslant s \leqslant g(t)} U(s), \quad t \geqslant 0. \tag{2.61}$$

Soit $f^*(t)$ l'unique temps qui réalise le maximum de U sur $[0,\ g(t)]$:

$$f^*(t) := \sup\{r \leqslant g(t)\ ;\ U(r) = U^*(t)\}, \quad t \geqslant 0. \tag{2.62}$$

Il est utile d'introduire aussi $g^*(t)$ le début de l'excursion enjambant $f^*(t)$:

$$g^*(t) := g(f^*(t)) = \sup\{r \leqslant f^*(t)\ ;\ U(r) = 0\}, \quad t \geqslant 0 \tag{2.63}$$

de même que la fin $d^*(t)$ de cette excursion :

$$d^*(t) := d(f^*(t)) = \inf\{r \geqslant f^*(t)\ ;\ U(r) = 0\}, \quad t \geqslant 0. \tag{2.64}$$

Les différentes notations sont illustrées dans la Figure 2.7.

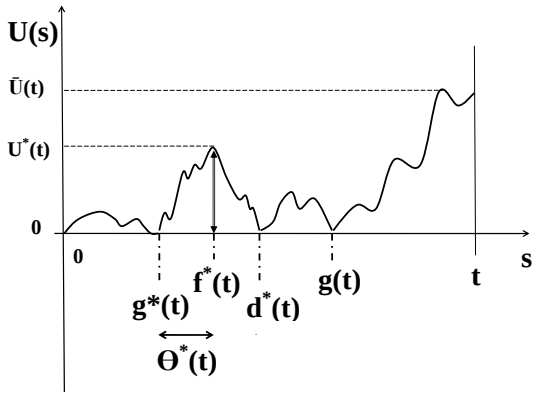


FIGURE 2.7 – Notations dans le cadre continu

Nous nous intéressons alors à la loi jointe du score local $U^*(t)$ et de sa longueur $\theta^*(t)$ sur les excursions complètes où

$$\theta^*(t) := f^*(t) - g^*(t), \quad t \geqslant 0. \tag{2.65}$$

Soit t un nombre réel fixé. La densité de $\overline{U}(t)$ est connue (voir [34, Section 2.11] ou [219, Lemma 3.2]). Bien que $U^*(t) = \overline{U}\big(g(t)\big)$ et que $g(t)$ ne soit pas un temps d'arrêt, il est quand même t facile de déterminer la densité de $U^*(t)$. En effet, le processus $\big(g(t)^{-1/2} B(g(t)s),\ 0 \leqslant s \leqslant 1\big)$ est distribué comme

$\big(b(s),\ 0 \leqslant s \leqslant 1\big)$ et il est indépendant de $g(t)$, où b est le pont Brownien [29]). Ainsi

$$U^*(t) \overset{(d)}{=} \sqrt{g(t)} \sup_{0\leqslant s\leqslant 1} |b(s)|. \tag{2.66}$$

Finalement, on conclut en utilisant le fait que $g(t)$ est distribué selon la loi de l'arcsinus [29] et la loi de $\sup_{0\leqslant s\leqslant 1} |b(s)|$ est donnée par la formule de Kolmogorov-Smirnov [203] (cf. Théorème 2.22 pour le résultat final).

Cependant, à notre connaissance, la distribution de $\big(U^*(t), \theta^*(t)\big)$ n'a jamais été déterminée. C'est ce que nous nous proposons de faire dans la section suivante (Théorèmes 2.20 et 2.21) en utilisant la théorie des excursions du MB dans $\mathbb{R}$. Nous établissons aussi les distributions marginales de $U^*(t)$ et $\theta^*(t)$ (Théorème 2.22).

2.3.2 Résultats théoriques

Avant d'établir les résultats, introduisons quelques notations supplémentaires.
1) $(\xi_n)_{n\geqslant 1} \cup \{\xi, \xi'\}$ est une famille de v.a. i.i.d. telles que

$$\xi \overset{(d)}{=} \xi' \overset{(d)}{=} \xi_n \overset{(d)}{=} T_1(R) \tag{2.67}$$

avec

$$T_x(R) = \inf\{s \geqslant 0\ ;\ R(s) = x\}, \quad x > 0 \tag{2.68}$$

et $(R(s), s \geqslant 0)$ est le processus de Bessel de dimension 3 issu de 0. La densité p_ξ de ξ est connue et donnée par

$$p_\xi(u) = \frac{1}{\sqrt{2\pi}u^{3/2}} \sum_{k\in\mathbb{Z}} \left(-1 + \frac{(1+2k)^2}{u}\right) \exp\left(-\frac{(1+2k)^2}{2u}\right) \tag{2.69}$$

$$= \frac{d}{du}\left(\sum_{k\in\mathbb{Z}} (-1)^k \exp\left(-\frac{k^2\pi^2 u}{2}\right)\right) \tag{2.70}$$

(voir par exemple [32] p 8 et 24). En vue de la simulation, un algorithme pour simuler rapidement et efficacement les v.a. ξ a été développé dans [82].
2) e_0', $(e_n)_{n\geqslant 0}$ est une suite de v.a. i.i.d. de loi exponentielle de paramètre 1.
3) $(\lambda(x), x \geqslant 0)$ est le processus défini par

$$\lambda(x) := x^2(\xi_1 + \xi_2) + \sum_{k\geqslant 1} \frac{\xi_{2k+1} + \xi_{2k+2}}{\big(\frac{1}{x} + e_1 + \ldots + e_k\big)^2}, \quad x \geqslant 0. \tag{2.71}$$

La somme ci-dessus converge p.s. et dans L^1.
4) α_1 et α_2 sont deux v.a. sur $[0,1]$; α_2 suit la loi uniforme tandis que la densité de α_1 est $\frac{2}{\pi}\frac{1}{\sqrt{1-s^2}}\mathbb{1}_{[0,1]}(s)$.

Dans la suite, nous supposons que

$$e_0',\ (e_n)_{n\geqslant 0},\ (\xi_n)_{n\geqslant 1},\ \xi,\ \xi',\ \alpha_1,\ \alpha_2 \text{ et } (U(t),\ t \geqslant 0) \text{ sont indépendantes.} \tag{2.72}$$

Dans le théorème suivant, nous déterminons la densité de $(U^*(t), \theta^*(t))$. Sa preuve se fonde sur la théorie des excursions du mouvement Brownien [212, Chapter XII]. Soit $(L(t),\ t \geqslant 0)$ le temps local en 0 du mouvement Brownien. La fonction aléatoire $t \mapsto L(t)$ est continue et croissante. Soit $(\tau_s,\ s \geqslant 0)$ son

inverse à droite. La preuve procède alors selon deux étapes : tout d'abord, on exprime la densité de $(U^*(t), \theta^*(t))$ en fonction de $\left(\overline{U}(\tau_1),\ \tau_1\right)$ puis on détermine la loi de $\left(\overline{U}(\tau_1), \tau_1\right)$. Nous obtenons alors le résultat suivant.

Théorème 2.20 (Densité du couple $\left(U^*(t), \theta^*(t)\right)$)**.** *Pour tout $t > 0$, la densité du couple $\left(U^*(t), \theta^*(t)\right)$ est donnée par*

$$p_{(U^*(t),\theta^*(t))} = \sqrt{\frac{2}{\pi}}\rho(x,y)p_\xi\left(\frac{y}{x^2}\right)\frac{1}{x^4}, \quad 0 < x, 0 < y < t \tag{2.73}$$

où p_ξ est la densité de ξ (cf. (2.69)-(2.70)*),*

$$\rho(x,y) = \int_{\mathbb{R}_+^2} \mathbb{E}\left[\frac{1}{\lambda(v)}\left\{\sqrt{(\rho_1)_+} - \sqrt{\left(\rho_1 - \frac{x^2}{v^2}\lambda(v)\right)_+}\right\}\right] p_\xi(u)\frac{e^{-1/v}}{v^2}\,\mathrm{d}u\mathrm{d}v, \tag{2.74}$$

$x_+ := \sup\{x, 0\}$ *et* $\rho_1 := t - y - x^2u$.

La formule (2.74) n'est pas complétement explicite et ne permet pas le calul direct de $\mathbb{E}\left[f\left(U^*(t), \theta^*(t)\right)\right]$ pour toute fonction Borélienne bornée f. Par exemple, en vue d'une application aux séquences biologiques comme expliqué en introduction, il serait intéressant de déterminer $\mathbb{P}\left(U^*(t) \leqslant a,\ \theta^*(t) \leqslant b\right)$ pour tout $a > 0$ et tout $b > 0$. Cependant, en outre, il est possible d'établir une formulation équivalente au Théorème 2.20 qui donne lieu à un résultat plus utile en pratique. Notamment, la quantité $\mathbb{E}\left[f\left(U^*(t), \theta^*(t)\right]\right.$ pourra être approchée par une procédure Monte-Carlo.

Théorème 2.21. *Soit $f : \mathbb{R}_+^2 \to \mathbb{R}$ une fonction Borélienne bornée. Alors*

$$\mathbb{E}\left[f(U^*(t), \theta^*(t))\right] = \sqrt{\frac{\pi}{2}}\mathbb{E}\left[f\left(\frac{\alpha_1\sqrt{t}}{\sqrt{Z}}, \frac{t\alpha_1^2\xi}{Z}\right)\frac{\alpha_2 e_0'^2}{\sqrt{Z}}\right] \tag{2.75}$$

où $Z = \xi + \xi' + e_0'^2\alpha_2^2\lambda(1/e_0')$.

Les distributions marginales du couple $(U^*(t), \theta^*(t))$ sont données par le théorème suivant.

Théorème 2.22 (Distributions marginales de $U^*(t)$ et $\theta^*(t)$)**.** *Soit $t > 0$. Les variables $U^*(t)$ et $\theta^*(t)$ ont pour densité respective :*

$$f_{U^*(t)}(x) = 4\sqrt{\frac{2}{\pi t}}\left(\sum_{k\geqslant 1}(-1)^{k-1}ke^{-\frac{2k^2x^2}{t}}\right)\mathbb{1}_{]0,+\infty[}(x), \tag{2.76}$$

et

$$f_{\theta^*(t)}(x) = \frac{1}{x}\sum_{k\geqslant 1}(-1)^{k+1}\frac{\sinh\left(\pi k\sqrt{\frac{x}{t-x}}\right)}{\cosh^2\left(\pi k\sqrt{\frac{x}{t-x}}\right)}\mathbb{1}_{[0,t]}(x). \tag{2.77}$$

A noter que la preuve de (2.76) n'utilise pas le Théorème 2.20.

Rappelons que pour des raisons techniques, nous avons considéré le score local $U^*(t)$ sur les montagnes complètes au lieu du score local $\overline{U}(t)$ sur l'ensemble de la séquence. Dans ce qui suit, nous souhaitons étudier la différence entre $\overline{U}(t)$ et $U^*(t)$. Lorsque $\mathbb{E}[\varepsilon] < 0$, la dernière excursion (incomplète) n'est généralement pas très longue. Cependant, lorsque $\mathbb{E}[\varepsilon] = 0$, ce n'est plus le cas. En effet, des simulations présentées en Section 2.3.3 et réalisées dans le cadre discret ont montré que pour de nombreuses séquences, $\overline{U}_n$ est réalisé lors de la dernière excursion incomplète. Naturellement, le nombre d'excursions augmente lorsque la longueur de la séquence augmente. Cependant, la proportion de séquences qui atteignent leur maximum sur une excursion complète reste étonnamment constante. L'objectif principal de la suite de cette étude est d'expliquer ces observations et de calculer cette proportion lorsque n est grand.

Pour ce faire, nous introduisons donc la probabilité p_c que le maximum de U sur $[0,t]$ soit atteint sur une excursion complète :

$$p_c := \mathbb{P}\left(\overline{U}(t) = U^*(t)\right). \tag{2.78}$$

Théorème 2.23 (Calcul de p_c)**.** *La probabilité p_c vaut $\psi(1/4) - \psi(1/2) + 1 + \pi/2 \approx 0.3069$.*

Présentons quelques éléments de preuve. Il est évident que la connaissance de $\overline{U}(t)$ et celle de $U^*(t)$ ne suffisent pas pour déterminer p_c. La trajectoire $(U(s), 0 \leqslant s \leqslant t)$ est naturellement divisée en deux parties, avant et après le temps aléatoire $g(t)$ qui n'est pas un temps d'arrêt comme nous l'avons déjà mentionné. Bien que $\big(U(s), 0 \leqslant s \leqslant g(t)\big)$ et $\big(U(s), g(t) \leqslant s \leqslant t\big)$ ne sont pas indépendants, le changement d'échelle par $g(t)$ conduit à de l'indépendance :

$$\left(\frac{1}{\sqrt{g(t)}} U(sg(t)),\ 0 \leqslant s \leqslant 1\right),\ \left(\frac{1}{\sqrt{t-g(t)}} \left|B(g(t) + s(t-g(t)))\right|,\ 0 \leqslant s \leqslant 1\right),\ g(t) \tag{2.79}$$

sont indépendants et la distribution de chacune des coordonnées est connue (voir [29, 203, 31]). Comme

$$\left(U^*(t), \max_{g(t) \leqslant s \leqslant t} U(s)\right) \overset{(d)}{=} \left(\sqrt{tg(1)}\, b^*, \sqrt{t(1-g(1))} \max_{0 \leqslant u \leqslant 1} m(u)\right) \tag{2.80}$$

où $b^* = \sup_{0 \leqslant s \leqslant 1} |b(s)|$ et m est le méandre Brownien, on en déduit que p_c ne dépend pas de t et que

$$p_c = \sqrt{\frac{\pi}{2}} \mathbb{E}\left[F\left(b^* \sqrt{\frac{g(1)}{1-g(1)}}\right)\right] \quad \text{où} \quad F(x) = \mathbb{E}\left[\frac{1}{R(1)} \mathbb{1}_{\left\{\max_{0 \leqslant u \leqslant 1} R(u) < x\right\}}\right], \tag{2.81}$$

avec $(R(u), 0 \leqslant u \leqslant 1)$ le processus de Bessel de dimension 3 issu de 0. Il suffit ensuite de déterminer la loi de $b^* \sqrt{g(1)/(1-g(1))}$, de calculer F puis l'espérance ci-dessus.

Pour conclure cette section, nous nous intéressons maintenant à la distribution de $g^*(t)$. Ce temps aléatoire s'interprète facilement dans le contexte du score local, comme nous l'avons vu précédemment. Rappelons que d'après la propriété de changement d'échelle du mouvement Brownien, $g^*(t)$ est distribué comme $tg^*(1)$. Nous considérons donc dans la suite $t = 1$.

Théorème 2.24 (Densité de $g^*(1)$)**.** *Introduisons la fonction suivante :*

$$h(x) = \sum_{k \geqslant 1} (-1)^{k+1} \frac{k}{\cosh^2(kx)}. \tag{2.82}$$

(i) La densité de $g^*(1)$ est donnée par

$$p_{g^*(1)}(y) = \frac{1}{2\pi\sqrt{y(1-y)}} \int_0^{+\infty} \ln\left|1 - \frac{\pi^2(1-y)}{4ys}\right| \frac{h(\sqrt{s})}{\sqrt{s}} \mathrm{d}s,\ 0 < y < 1. \tag{2.83}$$

(ii) Elle s'écrit encore sous la forme :

$$p_{g^*(1)}(y) = \frac{1}{\pi y} \int_0^{+\infty} \ln|\cot s| \frac{\mathrm{d}s}{\cosh^2\left(s\sqrt{\frac{1-y}{y}}\right)},\ 0 < y < 1. \tag{2.84}$$

Le schéma de la preuve est le suivant et se base sur l'identité en loi donnée par

$$\left(\frac{1}{\sqrt{g(1)}} B(sg(1)),\ 0 \leqslant s \leqslant 1\right) \overset{(d)}{=} (b(s),\ 0 \leqslant s \leqslant 1) \tag{2.85}$$

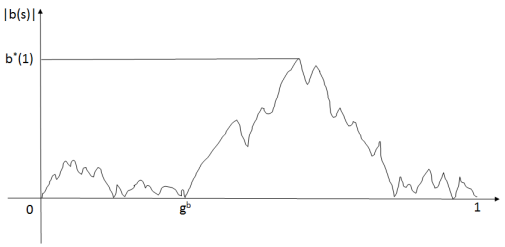


FIGURE 2.8 – Les v.a. g^b et $b^*(1)$

où

$$\left(\frac{1}{\sqrt{g(1)}}B(sg(1)),\ 0 \leqslant s \leqslant 1\right) \quad \text{est indépendant de} \quad g(1). \tag{2.86}$$

En remplaçant U par $|b|$ dans les équations (2.59)-(2.63) (resp. (2.59)-(2.61)), nous obtenons g^b (resp. $b^*(1)$) que nous avons représentés dans la Figure 2.8 Nous montrons ensuite que

$$g^*(1) \overset{(d)}{=} g(1)g^b \tag{2.87}$$

où $g(1)$ et g^b sont des v.a. indépendantes. Ainsi, il reste à déterminer les lois de $g(1)$ et de g^b. La FR de $g(1)$ est donnée par

$$\mathbb{P}\left(g(1) \in dx\right) = \frac{1}{\pi\sqrt{x(1-x)}}\,dx,\ 0 < x < 1, \tag{2.88}$$

tandis que la densité de g^b est donnée par

$$p_{g^b}(u) = \frac{\sqrt{2\pi}}{4}\frac{1}{\sqrt{u}(1-u)}\int_0^{+\infty}\mathbb{E}\left[\frac{1}{\sqrt{\tau_1}}\mathbb{1}_{\{u(4t\overline{U}(\tau_1)^2+\tau_1)<\tau_1\}}\right]p_\xi(t)\mathrm{d}t. \tag{2.89}$$

2.3.3 Application au cas discret des séquences biologiques

Commençons par définir l'interpolation linéaire de $(U_k)_k$. Soit $M > 0$ un paramètre d'échelle. Le processus continu $\left(B^M(t),\ t \geqslant 0\right)$ associé à $(S_k)_k$ (défini par (2.48)) et de facteur de normalisation M est défini par $B^M\left(\frac{k}{M}\right) = \frac{1}{\sqrt{M}}S_k$ et, pour tout k tel que $\frac{k}{M} \leqslant t \leqslant \frac{k+1}{M}$,

$$B^M(t) = B^M\left(\frac{k}{M}\right) + M\left(t - \frac{k}{M}\right)\left(B^M\left(\frac{k+1}{M}\right) - B^M\left(\frac{k}{M}\right)\right).$$

Nous introduisons ensuite le processus $(U^M(t),\ t \geqslant 0)$

$$U^M(t) = B^M(t) - \min_{s \leqslant t} B^M(s), \quad t \geqslant 0. \tag{2.90}$$

Notons que

$$U^M\left(\frac{k}{M}\right) = \frac{1}{\sqrt{M}}U_k, \quad k \geqslant 0 \tag{2.91}$$

où $(U_k)_k$ est le processus de Lindley associé à $(S_k)_k$ via (2.50). Enfin, nous définissons les analogues continus associés à $(B^M(t))$ des v.a. discrètes définies par les équations (2.51) dans le cadre du processus

de Lindley.

$$\begin{aligned}
\overline{U}^M(t) &:= \sup_{0\leqslant s\leqslant t} U^M(s), \qquad g^M(t) := \sup\left\{s \leqslant t \ ; \ U^M(s) = 0\right\}, \\
U^{M,*}(t) &:= \overline{U}^M(g^M(t)) = \sup_{0\leqslant s\leqslant g^M(t)} U^M(s), \\
f^{M,*}(t) &:= \sup\left\{r \leqslant g^M(t);\ U^M(r) = U^{M,*}(t)\right\}, \\
g^{M,*}(t) &:= g^M(f^{M,*}(t)) = \sup\left\{r \leqslant f^{M,*}(t) \ ; \ U^M(r) = 0\right\}, \\
d^{M,*}(t) &:= \inf\left\{s \geqslant f^{M,*}(t) \ ; \ U(s) = 0\right\}, \qquad \theta^{M,*}(t) := f^{M,*}(t) - g^{M,*}(t).
\end{aligned} \tag{2.92}$$

Il est alors clair que

$$\frac{\theta^*_M}{M} = \theta^{M,*}(1) \quad \text{et} \quad \frac{U^*_M}{\sqrt{M}} = U^{M,*}(1). \tag{2.93}$$

L'ingrédient clef intervenant pour montrer la convergence de $\big(U^{M,*}(t), \theta^{M,*}(t),\ t \geqslant 0\big)$ vers $\big(U^*(t), \theta^*(t),$ $t \geqslant 0\big)$ est le théorème de Donsker [34, Section 2.10] établissant que le processus $\big(B^M(t),\ t \geqslant 0\big)$ converge en loi vers le MB $\big(B(t),\ t \geqslant 0\big)$ lorsque $M \to +\infty$. En utilisant de plus (2.57), il s'ensuit que

$$\big(U^M(t),\ t \geqslant 0\big) \xrightarrow[M\to\infty]{\mathcal{L}} \big(U(t),\ t \geqslant 0\big). \tag{2.94}$$

A ce stade, il n'est pas clair que l'application $\omega \mapsto \big(g^{M,*}(t), f^{M,*}(t), d^{M,*}(t), \theta^{M,*}(t)\big)$ soit continue. Ainsi la convergence en loi de $\big(g^{M,*}(t), f^{M,*}(t), d^{M,*}(t), \theta^{M,*}(t), U^{M,*}(t)\big)$ lorsque $M \to \infty$ n'est pas une conséquence directe de (2.94). Cependant, il est malgré tout possible de démontrer le résultat suivant.

Théorème 2.25. *Soit $t > 0$. Le vecteur $\big(g^{M,*}(t), f^{M,*}(t), d^{M,*}(t), \theta^{M,*}(t), U^{M,*}(t)\big)$ converge en loi vers $\big(g^*(t), f^*(t), d^*(t), \theta^*(t), U^*(t)\big)$ lorsque $M \to \infty$ où les v.a. $g^*(t)$, $f^*(t)$, $d^*(t)$, $\theta^*(t)$ et $U^*(t)$ sont définies par les relations* (2.61)-(2.65)*.*

Introduisons maintenant $p_c(n)$ la probabilité que le maximum de $(U_k)_{0\leqslant k\leqslant n}$ soit atteint sur une excursion complète, *i.e.*

$$p_c(n) := \mathbb{P}(\overline{U}_n = U^*_n). \tag{2.95}$$

Proposition 2.26. *$p_c(n)$ converge vers p_c lorsque $n \to \infty$.*

Applications numériques

Dans [J11], nous avons réalisé quelques simulations numériques. D'une part, nous avons tracé des graphiques log-log très souvents utilisés dans la communauté des biologistes pour tester l'exactitude des approximations proposées. D'autre part, nous avons proposé des tests d'adéquation pour le score local (test classique de Kolmogorov-Smirnov) et pour le couple en considérant la variable scalaire $U^*_n/\sqrt{\theta^*_n}$. Enfin, nous avons étudié numériquement comment les valeurs de U^*_n et du score local classique $\overline{U}_n$ diffèrent lorsque le score local est réalisé sur la dernière excursion incomplète. Nous avons ainsi considéré des données simulées pour lesquelles nous avons vérifié que $p_c(n) \approx 70\%$ ainsi que les jeux de données SCOP1 (SP scop2dom 20140205aa. *http ://scop2.mrc-lmb.cam.ac.uk/downloads/*) et SCOP2 (SP scop2dom 20140205aa. *http ://scop2.mrc-lmb.cam.ac.uk/downloads/*) et l'échelle hydrophobique proposée dans [130]. Les résultats pour ces deux derniers jeux sont présentés dans la Table 2.2.

2.4 Processus autorégressifs

Dans cette section, je présente rapidement le cadre d'étude et les résultats obtenus en collaboration avec Thi Mong Ngoc Nguyen (Université des Sciences d'Ho Chi Minh, Vietnam) et Frédéric Proïa (Université

Données	Nb de séquences	Lg moyenne (max)	$\widehat{\mathbb{E}[\varepsilon]}$	Pourcentage
SCOP1	780	292 (1506)	-0.02	40%
SCOP2	606	115 (404)	-0.23	64%

TABLE 2.2 – Pourcentage de séquences réalisant leur score local score sur une excursion complète (données SCOP).

d'Angers) et publiés dans [J15].

Pour les séquences i.i.d. de v.a., il existe un large éventail de tests d'adéquation en relation avec la distribution sous-jacente. On peut penser au test de Kolmogorov-Smirnov déjà évoqué dans la Section 2.3.3, au critère de Cramér-von Mises, au test du chi-deux de Pearson, ou encore à des tests plus spécifiques comme les tests de normalité. La plupart d'entre eux sont fréquemment utilisés en pratique et directement mis en œuvre pour tester les résidus de modèles de régression. Pour ces applications, l'hypothèse d'indépendance n'est pas pertinente, en particulier pour les séries chronologiques pour lesquelles il existe une dépendance temporelle. Ainsi, la question cruciale qui se pose naturellement consiste à avoir un aperçu de leur sensibilité face à certaines hypothèses affaiblies. Nous nous concentrons dans [J15] sur une telle généralisation pour la statistique de Bickel-Rosenblatt, introduite par les statisticiens éponymes en 1973 [33], qui ont établi sa normalité asymptotique et donné leurs noms à la procédure de test associée. La statistique est étroitement liée à la distance L^2 entre l'estimateur à noyau de Parzen-Rosenblatt et une distribution paramétrique (ou une version lissée). Plus précisément, il prend la forme

$$\int_{\mathbb{R}} \big(\widehat{f}_n(x) - f(x)\big)^2 a(x)\,\mathrm{d}x.$$

Les notations seront précisées plus loin. Quelques améliorations intéressantes ont été apportées. Premièrement, dans [244] puis plus tard dans [190], les auteurs ont étendu le résultat à des séquences faiblement dépendantes (sous des conditions de mélange ou de régularité). Comme ils l'ont remarqué, ces hypothèses sont satisfaites par de nombreux processus temporels. Lee et Na [162] ont ensuite montré qu'il est aussi valable pour les résidus d'un processus autorégressif d'ordre 1 tant qu'il ne contient pas de racine d'unité. Ceci conduit à un test d'adéquation pour la distribution des innovations du processus. Bachmann et Dette [15] ont mis en pratique les résultats obtenus par Lee et Na. Leur étude permet également d'obtenir la normalité asymptotique de la statistique correctement renormalisée sous certaines alternatives fixes. Même s'il n'est pas directement lié à notre cadre d'étude, nous mentionnons également le travail de Horváth et Zitikis [123] qui donne quelques résultats sur les tests d'adéquation en norme r ($r \geqslant 1$) pour les résidus de processus autorégressifs de premier ordre.

L'enjeu de notre travail a été double. Premièrement et principalement, nous avons généralisé les résultats de Lee et Na aux processus autorégressifs d'ordre p ($p \geqslant 1$) tout en affinant l'ensemble des hypothèses et en discutant de l'effet des racines de l'unité sur la statistique d'intérêt. Deuxièmement, nous en avons déduit un test d'adéquation dont l'efficacité a été testé sur une simulation numérique. D'un point de vue théorique, il n'était pas évident d'étendre les résultats de premier ordre bien connus à des cas plus généraux et il était encore plus difficile de traiter les racines de l'unité dans le contexte des séries chronologiques. Les modèles autorégressifs sont très répandus dans des domaines d'application tels que l'économétrie, la finance mathématique, la météorologie et la prévision énergétique, l'ingénierie... Ainsi, proposer quelques avancées sur l'étude des processus autorégressifs, en termes d'inférence, de prédiction, de signification statistique ou, dans notre cas, d'adéquation, était très motivant et relevait d'un grand défi.

Précisons le contexte de notre étude. Nous considérons un processus autorégressif $(X_t)_t$ d'ordre p (AR(p))

défini par

$$X_t = \theta_1 X_{t-1} + \ldots + \theta_p X_{t-p} + \varepsilon_t \tag{2.96}$$

pour tout $t \geqslant 1$ ou de façon équivalente par

$$X_t = \theta^\top \Phi_{t-1} + \varepsilon_t$$

où $\theta = (\theta_1, \ldots, \theta_p)^\top$ est le vecteur des paramètres, Φ_0 est un vecteur aléatoire initial, $\Phi_t = (X_t, \ldots, X_{t-p+1})^\top$ et (ε_t) est un bruit blanc fort de variance $0 < \sigma^2 < +\infty$ et de densité marginale f (positive sur la droite réelle). Le polynôome caractéristique associé est défini, pour tout $z \in \mathbb{C}$, par

$$\Theta(z) = 1 - \theta_1 z - \ldots - \theta_p z^p \tag{2.97}$$

et la matrice compagnon associée à Θ (voir, *e.g.*, [91, Sec. 4.1.2]) est donnée par

$$C_\theta = \begin{pmatrix} \theta_1 & \theta_2 & \ldots & \theta_{p-1} & \theta_p \\ 1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & 1 & 0 \end{pmatrix}. \tag{2.98}$$

Il est bien connu que la stabilité du processus p-dimensionnel dépend des valeurs propres de C_θ que nous notons et réordonnons de la façon suivante :

$$\rho(C_\theta) = |\lambda_1| \geqslant |\lambda_2| \geqslant \ldots \geqslant |\lambda_p|.$$

Supposons que nous ayons observé $X_{-p+1}, \ldots, X_0, X_1, \ldots, X_n$ pour $n \gg p$ et notons $\widehat{\theta}_n$ défini par

$$\widehat{\theta}_n = \left(\sum_{t=1}^{n} \Phi_{t-1} \Phi_{t-1}^\top \right)^{-1} \sum_{t=1}^{n} \Phi_{t-1} X_t \tag{2.99}$$

l'estimateur par moindres carrés de θ pour $p > 0$. Le processus résiduel associé est le suivant :

$$\widehat{\varepsilon}_t = X_t - \widehat{\theta}_n^T \Phi_{t-1} \tag{2.100}$$

pour tout $1 \leqslant t \leqslant n$, ou simplement $\widehat{\varepsilon}_t = X_t$ lorsque $p = 0$. Dans la suite, $\mathbb{K}$ est un noyau et $(h_n)_n$ est la largeur de la fenêtre. Ceci sous-entend que $\mathbb{K}$ est une fonction positive satisfaisant

$$\int_{\mathbb{R}} \mathbb{K}(x)\,\mathrm{d}x = 1, \qquad \int_{\mathbb{R}} \mathbb{K}^2(x)\,\mathrm{d}x < +\infty \qquad \text{et} \qquad \int_{\mathbb{R}} x^2\,\mathbb{K}(x)\,\mathrm{d}x < +\infty,$$

et $(h_n)_n$ est une suite positive décroissant vers 0. L'estimateur de Parzen-Rosenblatt [196, 215] de la densité f est donné, pour tout $x \in \mathbb{R}$, par

$$\widehat{f}_n(x) = \frac{1}{n\,h_n} \sum_{t=1}^{n} \mathbb{K}\left(\frac{x - \widehat{\varepsilon}_t}{h_n} \right). \tag{2.101}$$

Les comportements local et global de cette densité empirique ont été étudiés dans la littérature. Cependant, pour un test d'adéquation, nous nous concentrons sur l'adéquation globale de $\widehat{f}_n$ à f sur la droite réelle. De ce point de vue, nous considérons la statistique de Bickel-Rosenblatt que nous définissons

comme

$$\widehat{T}_n = n\,h_n \int_{\mathbb{R}} \big(\widehat{f}_n(x) - (\mathbb{K}_{h_n} * f)(x)\big)^2 a(x)\,\mathrm{d}x \tag{2.102}$$

où $\mathbb{K}_{h_n} = h_n^{-1}\,\mathbb{K}(\cdot/h_n)$, a est une fonction positive continue et intégrable par morceaux et $*$ représente l'opérateur de convolution, *i.e.* $(g*h)(x) = \int_{\mathbb{R}} g(x-u)\,h(u)\,\mathrm{d}u$. Une statistique probablement plus intéressante et plus facile à implémenter est la suivante :

$$\widetilde{T}_n = n\,h_n \int_{\mathbb{R}} \big(\widehat{f}_n(x) - f(x)\big)^2 a(x)\,\mathrm{d}x. \tag{2.103}$$

Bickel et Rosenblatt montrent dans [33] que sous certaines conditions, si T_n est la statistique donnée par (2.102) mais construite sur le bruit blanc fort (ε_t) au lieu des résidus, alors, lorsque n tend vers l'infini,

$$\frac{T_n - \mu}{\sqrt{h_n}} \xrightarrow[N\to\infty]{\mathcal{L}} \mathcal{N}(0,\tau^2) \tag{2.104}$$

où le terme de centrage μ est donné par $\mu = \int_{\mathbb{R}} f(s)\,a(s)\,\mathrm{d}s \int_{\mathbb{R}} \mathbb{K}^2(s)\,\mathrm{d}s$ et la variance asymptotique par

$$\tau^2 = 2\int_{\mathbb{R}} f^{\,2}(s)\,a^2(s)\,\mathrm{d}s \int_{\mathbb{R}} \left(\int_{\mathbb{R}} \mathbb{K}(t)\,\mathbb{K}(t+s)\,\mathrm{d}t\right)^2 \mathrm{d}s. \tag{2.105}$$

Les hypothèses requises sont données dans [104, Sec. 2], elles découlent du travail originel de Bickel et Rosenblatt ensuite amélioré par Rosenblatt dans [216]. De plus, notons qu'en faisant quelques hypothèses aditionnelles et comme dans les résultats de [162], la normalité asymptotique (2.104) est satisfaite

— lorsque $\mathbb{K}$ est borné, a un support compact borné et la largeur de la fenêtre est donnée par $h_n = h_0\,n^{-\kappa}$ avec $0 < \kappa < 1$;

— lorsque $\mathbb{K}$ est un noyau continu, positif défini sur $\mathbb{R}$, la largeur de la fenêtre est donnée par $h_n = h_0\,n^{-\kappa}$ avec $0 < \kappa < 1/4$ et

$$\int_{|x|\geqslant 3} |x|^{3/2}\,(\ln\ln|x|)^{1/2}\,|\mathbb{K}'(x)|\,\mathrm{d}x < +\infty \quad \text{et} \quad \int_{\mathbb{R}} (\mathbb{K}'(x))^2\,\mathrm{d}x < +\infty.$$

Ultérieurement dans [244], [190] ou encore [15], nous pouvons trouver des preuves alternatives de la normalité asymptotique (2.104) avec $a(x) = 1$ sous des hypothèses appropriées. Cependant, dans notre étude, nous avons gardé la fonction a intégrable afin de suivre le cadre originel de Bickel et Rosenblatt. Ce cadre rend également les calculs plus faciles et reste approprié pour les applications puisqu'il est toujours possible de définir un support compact incluant toute masse d'une densité donnée f_0 que l'on souhaite tester (c'est d'ailleurs comme cela que nous avons procédé pour les simulations faites dans la Section 3 de [J15]).

Pour résumer, nous avons établi le comportement asymptotique de la statistique de Bickel-Rosenblatt $\widehat{T}_n$ définie par (2.102) et basée sur les résidus de processus autorégressifs stables et explosifs d'ordre p ($p \geqslant 1$). De façon analogue, la statistique $\widetilde{T}_n$ définie par (2.103) a aussi été étudiée. En outre, nous avons aussi établi des résultats pour des processus autorégressifs instables particuliers, comme la marche aléatoire et le processus saisonnier intégré. Enfin, nous avons abordé la question des processus généraux instables ou mixtes. En effet, par exemple, le processus ARIMA($p-1$,1,0) (instable) mérite une attention particulière en raison de son utilisation intensive en économétrie. Pour finir, nous avons construit un test d'adéquation.

Sur des échantillons finis de taille moyenne, le comportement gaussien limite est difficile à atteindre. Qui plus est, une asymétrie se produit pour les cadres dépendants (voir, par exemple, [247]). Dans [J15], nous avons mené des simulations reprenant les contextes numériques proposés dans [247] et [93] et Ghosh et

Huang [104] pour essayer de minimiser cet effet.
Ajoutons que le lecteur pourra aussi trouver dans [J15] une revue des résultats existants sur l'estimation par moindres carrés des paramètres autorégressifs, en fonction des racines de son polynôme caractéristique, puisqu'elle présente un intérêt crucial pour les preuves de nos résultats.

Chapitre 3

Principes de grandes déviations et bornes de Berry-Esseen

Le travail de ce chapitre a été mené en collaboration avec Pierre Petit (IMT) et Thierry Klein (IMT-ENAC) principalement mais aussi avec Franck Barthe (IMT) et Fabien Brosset (Doctorant IMT) que Franck et moi co-encadrons en thèse. Il fait l'objet de trois articles dont un accepté [J17] proposant des résultas de type Berry-Esseen et deux en cours de rédaction [P5] et [P4], le premier concernant des résultats limites sur les sommes non conditionnées de v.a. à queue lourde et le second proposant des résultats limites sur les sommes de v.a. conditionnées dans le cas particulier du hachage avec essais linéaires.

Comme l'a souligné Svante Janson dans son travail précurseur [132], dans de nombreux problèmes aléatoires combinatoires, la statistique d'intérêt peut se réécrire comme la somme de v.a. i.i.d. conditionnées par une v.a. à valeurs entières. En général, cette dernière est elle-même une somme de v.a. à valeurs entières. Plus précisément, nous nous intéressons à la loi de $N_n^{-1}(Y_{n,1}+\cdots+Y_{n,N_n})$ conditionnée par une valeur fixée de $X_{n,1}+\cdots+X_{n,N_n}$, *i.e.*, nous souhaitons étudier la distribution de

$$\mathcal{L}_n := \mathcal{L}(Y_{n,1}+\cdots+Y_{n,N_n} \mid X_{n,1}+\cdots+X_{n,N_n}=m_n),$$

où m_n et N_n sont des entiers et $(X_{n,i},Y_{n,i})_{n\in\mathbb{N}^*,1\leqslant i\leqslant N_n}$ sont des copies i.i.d. d'un couple (X_n,Y_n) de v.a. telles que X_n est à valeurs entières. Le cadre de travail concerne donc les tableaux triangulaires de v.a.
Le travail de Janson [132] a été en partie motivé par le modèle de hachage avec essais linéaires. Ce dernier est issu de l'informatique théorique, où il modélise le coût de stockage de données dans la mémoire. Il a été introduit dans un cadre mathématique par Knuth [147]. En raison de sa forte connexion avec les fonctions de parking, les distributions d'Airy (*i.e.*, la zone sous l'excursion Brownienne) et les promenades aléatoires de Lukasiewicz [158], ce modèle a été étudié par de nombreux auteurs (par exemple, Flajolet *et al.* [97], Janson [131, 133, 134], Chassaing *et al.* [58, 59, 60], et Marckert [174]).
Dans son travail, Janson prouve un théorème central limite général (avec la convergence de tous les moments) pour ce type de distribution conditionnelle sous certaines hypothèses et donne plusieurs applications combinatoires classiques : allocation d'urnes, hachage avec essais linéaires, forêts aléatoires, processus de ramification, etc. Suite à ce travail, au moins deux questions naturelles se posent.

1) Est-il possible d'obtenir des bornes de Berry-Esseen pour ce type de modèle afin de préciser le théorème central limite ?
2) Est-il possible d'obtenir des résultats de grandes déviations ?

Une réponse complète à la première question

Le premier théorème de Berry-Esseen pour des modèles conditionnés est dû à Quine et Robinson [207]. Dans leur travail, les auteurs étudient le problème d'occupation, *i.e.* le cas où les v.a. X_n suivent la loi de Poisson et $Y_n = \mathbb{1}_{\{X_n=0\}}$. A notre connaissance, c'est le seul résultat dans cette direction. Dans notre travail, nous prouvons des bornes générales de Berry-Esseen (Théorème 3.3 de la Section 3.1) qui sont valables pour tous les modèles présentés par Janson [132]. Ces modèles sont rappelés dans la Section 3.2.

Une réponse partielle à la seconde question

Lorsque la distribution de (X_n, Y_n) ne dépend pas de n, le principe de conditionnement de Gibbs ([248, 74, 88]) établit que $\mathcal{L}_n$ converge en loi vers une distribution dégénérée concentrée sur un point dépendant de la valeur de conditionnement [101, Corollary 2.2]. Autour du principe de conditionnement de Gibbs, des théorèmes limites généraux donnant le comportement asymptotique de sommes conditionnées ont été établis dans [236, 121, 152] et des développements asymptotiques ont été prouvés dans [119, 214].

Une extension aux tableaux triangulaires a été proposée par Gamboa, Klein et Prieur [101]. Les auteurs prouvent un principe de déviations modérées et un principe de grandes déviations sous certaines hypothèses dont la plus contraignante est que la transformée de Laplace jointe de (X_n, Y_n) est finie au moins dans un voisinage de l'origine $(0, 0)$. Cette hypothèse est satisfaite pour les exemples considérés dans [132], à l'exception du hachage avec essais linéaires, qui semble être le plus intéressant et qui est à l'origine des travaux de Janson. En effet, dans ce cas, la transformée de Laplace jointe n'est définie que sur $]-\infty\,, a] \times]-\infty\,, 0]$ pour une constante strictement positive a.

Plus généralement, il serait intéressant d'obtenir des résultats de grandes déviations pour une classe plus large de modèles pour lesquels la transformée de Laplace n'est pas définie. Dans [186, 187], Nagaev établit des résultats de grandes déviations pour les sommes de v.a. i.i.d. absolument continues par rapport à la mesure de Lebesgue et dont la transformée de Laplace n'est pas définie au voisinage de 0. En travaillant dans ce contexte, nous étendons son résultat et prouvons un résultat de grandes déviations à l'échelle logarithmique (Théorème 3.7 de la Section 3.3) pour des tableaux. Il est alors naturel de considérer le comportement asymptotique des sommes conditionnées et d'étendre le travail de [101] aux modèles pour lesquels la transfomée de Laplace n'est pas définie. Prouver un théorème pour une classe générale de modèles semble être une tâche très difficile. C'est pourquoi, nous nous limitons à l'étude du hachage avec essais linéaires (voir Section 3.4).

Cadre de notre étude

Commençons par préciser le cadre de notre travail. Pour tout $n \geqslant 1$, nous considérons les v.a. (X_n, Y_n) telles que X_n est une v.a. à valeurs entières et Y_n est une v.a. réelle. Soit N_n un entier tel que $N_n \to \infty$ lorsque n tend vers l'infini. Soit $(X_{n,i}, Y_{n,i})_{1 \leqslant i \leqslant N_n}$ un échantillon i.i.d. distribué comme (X_n, Y_n) et définissons

$$S_{n,k} := \sum_{i=1}^{k} X_{n,i} \quad \text{et} \quad T_{n,k} := \sum_{i=1}^{k} Y_{n,i},$$

pour $k \in [\![1\,, N_n]\!]$. Remarquons que nous considérons ici des tableaux triangulaires $\mathbf{X} = (X_{n,i})_{n \in \mathbb{N}^*, i=1,\ldots,N_n}$ et $\mathbf{Y} = (Y_{n,i})_{n \in \mathbb{N}^*, i=1,\ldots,N_n}$ tels que sur chaque ligne, les v.a. sont i.i.d. Evidemment, les tableaux ne sont pas indépendants ; sinon, la loi conditionnelle $\mathcal{L}_n$ serait triviale.

De façon à alléger les notations, nous poserons $S_n := S_{n,N_n}$ et $T_n := T_{n,N_n}$. Soit $m_n \in \mathbb{Z}$ une suite d'entiers relatifs telle que $\mathbb{P}(S_n = m_n) > 0$ pour tout n. L'objectif est d'étudier le comportement asymptotique des distributions conditionnelles suivantes

$$\mathcal{L}_n = \mathcal{L}(T_n | S_n = m_n).$$

Introduisons enfin la variable U_n distribuée comme T_n conditionné à $\{S_n = m_n\}$.

Plan du chapitre

Je présente dans la Section 3.1 les bornes de Berry-Esseen obtenues pour les sommes de v.a. conditionnées. Quelques exemples classiques sont ensuite décrits dans la Section 3.2. Des résultats de type Nagaev à l'échelle logarithmique sont établis dans la Section 3.3. Ils concernent les sommes non conditionnées de v.a. à queue lourde et seront ensuite utilisés pour établir les résultats limites de la Section 3.4 dédiée au modèle de hachage avec essais linéaires.

3.1 Bornes de type Berry-Esseen

L'objectif de cette section est de prouver des bornes de type Berry-Essen pour les distributions conditionnelles suivantes

$$\mathcal{L}(U_n) = \mathcal{L}(T_n | S_n = m_n).$$

Hypothèses 3.1. *Supposons qu'il existe des constantes positives c_1, $\tilde{c}_2$, c_2, c_3, $\tilde{c}_4$, c_4, c_5, c_6, c_7 et η_0, telles que :*

(H1) $\gamma_n := 2\pi\sigma_{X_n} N_n^{1/2} \mathbb{P}(S_n = m_n) \geqslant c_1$;

(H2) $\tilde{c}_2 \leqslant \sigma_{X_n} := \mathrm{Var}\,(X_n)^{1/2} \leqslant c_2$;

(H3) $\rho_{X_n} := \mathbb{E}\big[|X_n - \mathbb{E}[X_n]|^3\big] \leqslant c_3 \sigma_{X_n}^3$;

(H4) $\tilde{c}_4 \leqslant \sigma_{Y_n} := \mathrm{Var}\,(Y_n)^{1/2} \leqslant c_4$;

(H5) $\rho_{Y_n} := \mathbb{E}\big[|Y_n - \mathbb{E}[Y_n]|^3\big] \leqslant c_5 \sigma_{Y_n}^3$;

(H6) les corrélations $r_n := \mathrm{Cov}\,(X_n, Y_n)\, \sigma_{X_n}^{-1} \sigma_{Y_n}^{-1}$ *vérifient* $|r_n| \leqslant c_6 < 1$;

(H7) pour $Y_n' := Y_n - \mathbb{E}[Y_n] - \mathrm{Cov}(X_n, Y_n)\sigma_{X_n}^{-2}(X_n - \mathbb{E}[X_n])$, *pour tout* $s \in [-\pi\,, \pi]$ *et pour tout* $t \in [-\eta_0\,, \eta_0]$,

$$\Big|\mathbb{E}\big[e^{i(sX_n + tY_n')}\big]\Big| \leqslant 1 - c_7\big(\sigma_{X_n}^2 s^2 + \sigma_{Y_n'}^2 t^2\big).$$

Les Hypothèses 3.1 sont très proches de celles du TCL établi dans [132, Theorem 2.3]. En particulier, (H1) est une conséquence de $m_n = N_n \mathbb{E}[X_n] + O\big(\sigma_{X_n} N_n^{1/2}\big)$, (H3) et (H7) (voir la preuve du Théorème 2.3 de [132]). Par [132, Lemma 4.1.], $\sigma_{X_n}^2 \leqslant 4\mathbb{E}[|X - \mathbb{E}[X]|^3]$, on peut prendre $\tilde{c}_2$ égal à $1/(4c_3)$. (H6) n'est pas très restrictive et est satisfaite pour les exemples présentés dans la Section 3.2. En suivant [132], nous introduisons Y_n' dans (H7) de façon à travailler avec une v.a. centrée décorrélée de X_n. La v.a. Y_n' vérifie bien les mêmes hypothèses que Y_n et T_n' conditionné à $\{S_n = m_n\}$ a bien la même distribution que T_n conditionné à $\{S_n = m_n\}$. Si (X, Y') est un vecteur de v.a. centrées et décorrélées, alors

$$\Big|\mathbb{E}\big[e^{i(sX + tY')}\big]\Big| = 1 - \frac{1}{2}\big(\sigma_X^2 s^2 + \sigma_{Y'}^2 t^2\big) + o(s^2 + t^2).$$

Ainsi (H7) est raisonnable à partir du moment où les vecteurs (X_n, Y_n') sont identiquement distribués.

Proposition 3.2. *Supposons que*

$$m_n = N_n \mathbb{E}[X_n] + O\big(\sigma_{X_n} N_n^{1/2}\big),$$

que (X_n, Y_n) converge en loi vers (X, Y) lorsque $n \to \infty$ et que pour tout $r > 0$,

$$\limsup_{n \to +\infty} \mathbb{E}\left[|X_n|^r\right] < \infty \quad \textit{et} \quad \limsup_{n \to +\infty} \mathbb{E}\left[|Y_n|^r\right] < \infty.$$

Supposons en outre que la distribution de X a un span 1 (i.e. $\sup\{m \in \mathbb{N},\ \exists b \in \mathbb{N},\ Supp(X) \subset m\mathbb{Z}+b\}$ *où $Supp(X)$ est le support de X) et que Y n'est pas p.s. égale à une fonction affine $c+dX$ de X. Alors les Hypothèses 3.1 sont satisfaites.*

Nous omettons la preuve puisqu'elle repose seulement sur le Corollaire 2.1 et le Théorème 2.3 de [132]. Nous établissons maintenant les bornes de Berry-Esseen pour la suite de v.a. U_n.

Théorème 3.3. *Sous les Hypothèses 3.1, nous avons*

$$\sup_{x\in\mathbb{R}}\left|\mathbb{P}\left(\frac{U_n - N_n\mathbb{E}\left[Y_n\right] - r_n\sigma_{Y_n}\sigma_{X_n}^{-1}(m_n - N_n\mathbb{E}\left[X_n\right])}{N_n^{1/2}\tau_n} \leqslant x\right) - \Phi(x)\right| \leqslant \frac{(Cste)}{N_n^{1/2}}, \tag{3.1}$$

avec $\tau_n^2 = \sigma_{Y_n}^2(1-r_n^2) > 0$ et où Φ représente la FR de la distribution gaussienne standard et $(Cste)$ est une constante positive qui ne dépend que de $\tilde{c}_2$, c_2, c_3, $\tilde{c}_4$, c_4, c_5, c_6, c_8, η_0 et c_1.

Remarquons que la renormalisation des variables U_n intervenant dans (3.1) n'est pas la plus naturelle. La proposition suivante permet d'établir le Théorème 3.5 qui résout ce défaut de normalisation.

Proposition 3.4. *En supposant (H1), (H3), (H4), (H5) et (H7), il existe deux constantes positives d_1 et d_2 dépendant seulement de c_3, c_4, c_5, c_8 et c_1 telles que, pour $N_n \geqslant 3$,*

$$\left|\mathbb{E}\left[U_n\right] - N_n\mathbb{E}[Y_n] - r_n\sigma_{Y_n}\sigma_{X_n}^{-1}(m_n - N_n\mathbb{E}[X_n])\right| \leqslant d_1 \tag{3.2}$$

et

$$\left|\mathrm{Var}\left(U_n\right) - N_n\tau_n^2\right| \leqslant d_2 N_n^{1/2}. \tag{3.3}$$

Théorème 3.5. *Sous les Hypothèses 3.1, nous avons*

$$\sup_{x\in\mathbb{R}}\left|\mathbb{P}\left(\frac{U_n - \mathbb{E}\left[U_n\right]}{\mathrm{Var}\left(U_n\right)^{1/2}} \leqslant x\right) - \Phi(x)\right| \leqslant \frac{(Cste)}{N_n^{1/2}}, \tag{3.4}$$

où $(Cste)$ est une constante qui ne dépend que de $\tilde{c}_2$, c_2, c_3, $\tilde{c}_4$, c_4, c_5, c_6, c_8, η_0 et c_1.

De plus, comme dans [132], les résultats des Théorèmes 3.3 et 3.5 se simplifient considérablement dans le cas particulier où le vecteur (X_n, Y_n) ne dépend pas de n, ceci signifie que nous considérons une suite de v.a. i.i.d. au lieu d'un tableau triangulaire. C'est une conséquence de la Proposition 3.2.

3.2 Exemples classiques

Dans cette section, nous décrivons les exemples mentionnés dans [132] et [121] qui ont motivé leurs études. Il est montré dans [132] que chacun d'eux satisfait les hypothèses de la Proposition 3.2. Les bornes de Berry-Esseen sont donc satisfaites pour ces exemples.

Problème d'occupation

Dans le problème d'occupation, m balles sont réparties de façon aléatoire uniformément dans N urnes. Le vecteur $(Z_1, \ldots, Z_N)$ des nombres de balles dans chaque urne suit la loi multinomiale. Il est alors bien connu que $(Z_1, \ldots, Z_N)$ est aussi distribué comme le vecteur $(X_1, \cdots, X_N)$ conditionné à $\{\sum_{i=1}^N X_i = m\}$, où les v.a. X_i sont i.i.d. de loi de Poisson de paramètre $\lambda > 0$ arbitraire. Le problème classique d'occupation s'intéresse au nombre d'urnes vides donné par $U = \sum_{i=1}^N \mathbb{1}_{\{Z_i=0\}}$ et distribué comme $\sum_{i=1}^N \mathbb{1}_{\{X_i=0\}}$ conditionné à $\{\sum_{i=1}^N X_i = m\}$.

Maintenant, si $m = m_n \to \infty$ et $N = N_n \to \infty$ de telle sorte que $m_n/N_n \to \lambda \in]0\,,\infty[$, nous pouvons prendre $X_n \sim \mathcal{P}(\lambda_n)$ avec $\lambda_n = m_n/N_n$, $Y_n = \mathbb{1}_{\{X_n=0\}}$ et appliquer la Proposition 3.2 pour obtenir une inégalité de type Berry-Esseen pour $U_n = \sum_{i=1}^{N_n} \mathbb{1}_{\{Z_i=0\}}$.

Remarque 3.6. *Dans [207], les auteurs prouvent une inégalité de type Berry-Esseen pour le problème d'occupation dans un cadre plus général : la probabilité d'atterrir dans chaque urne peut être différente. Les outils développés dans [207] ont été utilisés dans la suite pour prouver nos résultats.*

Ici, nous avons besoin du résultat pour les tableaux triangulaires et pas seulement pour les suites de v.a. i.i.d. En effet, en considérant $X_n = X$ avec $X \sim \mathcal{P}(\lambda)$, nous aurions seulement

$$m_n = N_n(\lambda + o(1)) = N_n\mathbb{E}[X_n] + o(N_n).$$

Pourtant, la Proposition 3.2 requiert

$$m_n = N_n\mathbb{E}[X] + O(N_n^{1/2}),$$

qui est plus fort. Cette remarque vaut aussi pour les exemples qui suivent.

Statistique de Bose-Einstein

Cet exemple a été emprunté à [121] (voir aussi [95]). Ici m balles indiscernables sont jetées aléatoirement dans N urnes de telle sorte que chaque configuration résultante a la même probabilité $1/\binom{m+N-1}{m}$. Il est bien connu que le vecteur $(Z_1, \ldots, Z_N)$ des nombres de balles dans chaque urne. est distribué comme le vecteur $(X_1, \ldots, X_N)$ conditionné à $\{\sum_{i=1}^N X_i = m\}$, où les v.a. X_i sont i.i.d. de loi géométrique $\mathcal{G}(p)$ pour tout $p \in]0\,,1[$ arbitraire. Si $m = n$, $N = N_n \to \infty$ avec $N_n/n \to p$, en prenant $X_n \sim \mathcal{G}(p_n)$ où $p_n = N_n/n$, nous obtenons une inégalité de type Berry-Esseen pour toute suite de v.a. de la forme $U_n = \sum_{i=1}^{N_n} f(Z_i)$.

Processus de branchement

Considérons un processus de Galton-Watson ayant un seul ancètre et où le nombre d'enfants d'un individu est donné par une variable aléatoire X ayant des moments finis. Supposons de plus que $\mathbb{E}[X] = 1$. Nous numérotons les individus selon leur ordre d'apparition. Soit X_i le nombre d'enfants du i-ème individu et $S_k = \sum_{i=1}^k X_i$. Il est alors bien connu (voir [132, Example 3.4] et les références y figurant) que la taille totale de la population $S_N + 1$ vaut N ($N \geqslant 1$) si, et seulement si,

$$\forall k \in \{0, \ldots, N-1\} \quad S_k \geqslant k \quad \text{et} \quad S_N = N - 1. \tag{3.5}$$

Ce type de conditionnement est différent de celui considéré dans ce chapitre. Cependant, par [259, Corollary 2] et [132, Example 3.4], si nous ignorons l'ordre cyclique de $X_1, \ldots, X_N$, il est possible de prouver que $(X_1, \ldots, X_N)$ a la même distribution conditionnée à (3.5) que conditionnée à $\{S_N = N - 1\}$. En appliquant la Proposition 3.2 avec $N = n$ et $m = n-1$, nous obtenons une inégalité de type Berry-Esseen pour toute suite U_n distribuée comme $T_n = \sum_{i=1}^n f(X_i)$ conditionné à $\{S_n = n - 1\}$. Par exemple, si $f(x) = \mathbb{1}_{\{x=3\}}$, U_n est alors le nombre d'individus ayant trois enfants sachant que la population totale est n.

Forêts aléatoires

Considérons une forêt étiquetée uniformément distribuée avec m sommets et N racines $N < m$. Sans perte de généralité, nous pouvons supposer que les sommets sont numérotés $1, \ldots, m$ et, par symétrie,

que les racines sont les N premiers sommets. D'après [132], ce modèle peut être réalisé comme suit. Les tailles des N arbres dans la forêt sont distribuées comme $(X_1, \ldots, X_N)$ conditionné à $\{\sum_{i=1}^N X_i = m\}$, où les v.a. X_i sont i.i.d. et suivent la distribution de Borel distribués de paramètre $\mu \in]0\,,1[$ arbitraire (voir, e.g., [97] ou [131] pour plus de détails). Alors, le i-ème arbre est choisi uniformément parmi les arbres de taille X_i. La Proposition 3.2 fournit alors une inégalité de type Berry-Esseen pour toute suite de variables de type $U_n = \sum_{i=1}^{N_n} f(Z_i)$ où $N_n \to \infty$ et Z_1, ..., Z_{N_n} sont les tailles des arbres de la forêt. Par exemple, si $f(x) = \mathbb{1}_{\{x=K\}}$, U_n est le nombre d'arbres de taille K dans la forêt (voir [149, 199, 200]).

Hachage avec essais linéaires

Le hachage avec essais linéaires est un modèle classique en informatique théorique qui est apparu dans les années 60. Il a été étudié d'un point de vue mathématique tout d'abord dans [146]. Pour plus de détails sur le modèle, nous renvoyons le lecteur à [97, 131, 174, 59, 57, 133]. Le modèle décrit l'expérience suivante. On lance n balles séquentiellement dans m urnes au hasard avec $m > n$; les urnes sont disposées en cercle et numérotées dans le sens trigonométrique. Une balle qui atterrit dans une urne occupée est déplacée à l'urne vide suivante (dans le sens trigonométrique). La longueur du décalage est appelée le *déplacement* de la balle et nous sommes intéressés par la somme (aléatoire) $d_{m,n}$ de tous les déplacements. Après avoir lancé toutes les balles, il y a $N = m - n$ urnes vides. Celles-ci séparent les urnes occupées en blocs d'urnes consécutives. Nous considérons que l'urne vide qui suit un bloc appartient à ce bloc. Dans la suite de [148, 97], Janson [131] prouve que les longueurs des blocs et le déplacement total à l'intérieur de chaque bloc sont distribués comme (X_1, Y_1), ..., (X_N, Y_N) conditionnés à $\{\sum_{i=1}^N X_i = m\}$, où les vecteurs aléatoires (X_i, Y_i) sont des copies i.i.d. d'un vecteur (X, Y) de v.a. telles que X suit la distribution de Borel de paramètre $\mu \in]0\,,1[$ arbitraire et Y sachant $\{X = l\}$ est distribué comme $d_{l,l-1}$. En particulier, $d_{m,n}$ est distribué comme $\sum_{i=1}^N Y_i$ conditionné à $\{\sum_{i=1}^N X_i = m\}$. Si $m = m_n \to \infty$ et $N = N_n = m_n - n \to \infty$ avec $n/m_n \to \mu \in]0\,,1[$, nous prenons X_n suivant la loi de Borel de paramètre $\mu_n = n/m_n$ pour obtenir une inégalité de type Berry-Esseen pour $d_{m_n,n}$, d'après la Proposition 3.2.
Nous reviendrons plus en détails sur le modèle de hachage dans la Section 3.4 et donnerons des résultats de déviations pour différents régimes.

3.3 Résultats de type Nagaev à l'échelle logarithmique

Le premier résultat de grandes déviations sur les sommes de v.a. indépendantes est dû à Kinchin [144] en 1929 et concerne le cas particulier de v.a. de Bernoulli. Ce résultat a été complété par Smirnov [232] en 1933 puis généralisé par Cramér [71] en 1938 aux v.a. satisfaisant la condition éponyme, *i.e.* pour lesquelles la transformée de Laplace existe. Des améliorations du résultat précédent ont ensuite été proposées dans [94] et [201]. Il faut attendre un travail de Linnik en 1961 [166] pour avoir des résultats de grandes déviations lorsque la condition de Cramér n'est pas satisfaite. Linnik se place dans le cas où la décroissance de la queue de X est polynomiale. Le cas intermédiaire, où la queue de X décroît plus vite que toute puissance mais malgré tout pas assez vite pour que la transformée de Laplace soit définie, a été partiellement traité par Petrov [201] et S.V. Nagaev [188]. C'est finalement A.V. Nagaev dans [186, 187] qui résout le problème en 1969.
Dans cette section, nous réécrivons à l'échelle logarithmique les résultats de [186, 187] pour les sommes de v.a. à queue lourde, *i.e.* telles que

$$\log \mathbb{P}(X \geqslant x) \sim -x^{1-\varepsilon} \tag{3.6}$$

pour $\varepsilon \in]0\,,1[$ arbitraire lorsque x tend vers l'infini. Dans ce cas, la v.a. X est dite *sur-exponentielle*. Sous

cette hypothèse, la transformée de Laplace n'est pas définie partout à droite et la condition de Cramér n'est donc pas satisfaite. Nous supposerons en outre que la v.a. X est centrée, de variance σ^2 et rappelons que $S_n = \sum_{i=1}^n X_i$.

Théorème 3.7. *(i) Lorsque* $x_n \leqslant c_\varepsilon n^{1/(1+\varepsilon)}$*, alors*

$$\lim_{n\to+\infty} \frac{n}{x_n^2} \log \mathbb{P}(S_n \geqslant x_n) = -\frac{1}{2\sigma^2}. \tag{3.7}$$

(ii) Lorsque $c_\varepsilon n^{1/(1+\varepsilon)} \leqslant x_n \leqslant cn^{1/(1+\varepsilon)}$*, alors*

$$\log \mathbb{P}(S_n \geqslant x_n) \underset{n\to+\infty}{\sim} -(1-\alpha_1)^{1-\varepsilon} x_n^{1-\varepsilon} - \frac{\alpha_1^2 x_n^2}{2\sigma^2 n}, \tag{3.8}$$

où c_ε *et* α_1 *sont donnés par*

$$c_\varepsilon = (1+\varepsilon)\frac{\sigma^{2/(1+\varepsilon)}}{(2\varepsilon)^{\varepsilon/(1+\varepsilon)}} \quad \textit{et} \quad \alpha_1 = \frac{1-\varepsilon}{1+\varepsilon}. \tag{3.9}$$

(iii) Lorsque $x_n \gg n^{1/(1+\varepsilon)}$*, alors*

$$\lim_{n\to+\infty} \frac{1}{x_n^{1-\varepsilon}} \log \mathbb{P}(S_n \geqslant x_n) = -1. \tag{3.10}$$

Ainsi, à l'échelle logarithmique, nous distinguons trois régimes. Dans le régime exponentiel, toutes les variables comptent, *i.e.* toutes les variables contribuent pour atteindre les grandes déviations. L'autre régime extrémal est le régime du saut maximal où une seule variable est très grande et réalise les grandes déviations. Enfin, le régime intermédiaire combine les deux précédents. Dans ce régime, $x_n^{1-\varepsilon}$ et x_n^2/n ont un comportement limite identique. Remarquons qu'en régime précis (pas à l'échelle logarithmique), Nagaev distingue cinq régimes différents. Remarquons aussi que (3.9) correspond à la version limite de (6) dans [186]. Enfin, il est à noter que notre preuve utilise deux résultats classiques de grandes déviations : le théorème de Gärtner-Ellis et la principe de contraction à la différence de celle de Nagaev qui n'est pas analytique et fait intervenir de nombreuses étapes analytiques très techniques.

Schéma de la preuve du Théorème 3.7

Pour prouver le Théorème 3.7, nous avons suivi les étapes des preuves de Nagaev mais, comme nous venons de le signaler, en utilisant des résultats connus tels que le théorème de Gärtner-Ellis et le principe de contraction. Nous avons aussi réalisé les mêmes découpages ; à savoir, la probabilité d'intérêt $\mathbb{P}(S_n \geqslant x_n)$ se décompose en

$$P_n(x_n) := \mathbb{P}(S_n \geqslant x_n) = P_{n,0}(x_n) + P_{n,1}(x_n) + R_{n,2}(x_n), \tag{3.11}$$

où $P_{n,m}(x_n) = \mathbb{P}(S_n \geqslant x_n, X_1 > x_n, \ldots, X_m > x_n, X_{m+1} \leqslant x_n, \ldots, X_n \leqslant x_n)$ pour $m = 0, \ldots, n$ et

$$R_{n,2}(x_n) = \sum_{m=2}^{n} \binom{n}{m} P_{n,m}(x_n).$$

Il est immédiat de montrer que

$$R_{n,2}(x_n) \leqslant \sum_{m=1}^{n} n^m \mathbb{P}(X > x_n)^m = n\mathbb{P}(X > x_n) \sum_{m=0}^{n-1} n^m \mathbb{P}(X > x_n)^m = O\left(n\mathbb{P}(X > x_n)\right). \tag{3.12}$$

Ensuite, d'une part, il est évident que

$$P_n(x_n) \geqslant P_{n,1}(x_n)$$

et d'autre part, en utilisant (3.12), nous avons

$$\limsup_{n\to+\infty} a_n \log P_n(x_n) \sim \max\{\limsup_{n\to+\infty} a_n \log P_{n,0}(x_n), \limsup_{n\to+\infty} a_n \log P_{n,1}(x_n)\},$$

où a_n correspond aux différentes vitesses étudiées. Enfin dans les trois régimes, nous avons

$$a_n \log P_{n,1}(x_n) \leqslant a_n \log n + a_n \log \mathbb{P}(X_1 > x_n)$$

et $\limsup_{n\to+\infty} a_n \log P_{n,1}(x_n) \leqslant -1$. Il reste donc à traiter $P_{n,0}(x_n)$. Pour ce faire, nous procédons au découpage suivant

$$P_{n,0}(x_n) = \mathbb{P}(S_n \geqslant x_n, X_1 \leqslant x_n, \ldots, X_n \leqslant x_n) = \Pi_{n,0}(x_n,\varepsilon) + n\Pi_{n,1}(x_n,\varepsilon) + R'_{n,2}(x_n,\varepsilon), \tag{3.13}$$

où $\Pi_{n,m}(x_n,\varepsilon) = \mathbb{P}(S_n \geqslant x_n, x_n^\varepsilon < X_1 < x_n, \ldots, x_n^\varepsilon < X_m < x_n, X_{m+1} \leqslant x_n^\varepsilon, \ldots, X_n \leqslant x_n^\varepsilon)$ pour $m = 0, \ldots, n$ et

$$R'_{n,2}(x_n) = \sum_{m=2}^{n} \binom{n}{m} \Pi_{n,m}(x_n).$$

Par le théorème de Gärtner-Ellis unilatéral (voir Plachky-Steinebach [204]), nous montrons le résultat suivant.

Proposition 3.8. *(i) Lorsque $x_n \leqslant (Const) n^{1/(1+\varepsilon)}$, alors*

$$\lim_{n\to+\infty} \frac{n}{x_n^2} \log \Pi_{n,0}(x_n,\varepsilon) = -\frac{1}{2\sigma^2}. \tag{3.14}$$

(ii) Lorsque $x_n \gg n^{1/(1+\varepsilon)}$, alors

$$\frac{1}{x_n^{1-\varepsilon}} \log \Pi_{n,0}(x_n,\varepsilon) \underset{n\to+\infty}{\sim} -M_n, \tag{3.15}$$

où $M_n \to +\infty$ quand $n \to +\infty$.

En appliquant le principe de contraction (unilatéral) et en utilisant la Proposition 3.8 appliquée à $\Pi_{n-1,0}(x_n,\varepsilon)$, nous établissons le comportement asymptotique de $\Pi_{n,1}(x_n,\varepsilon)$.

Proposition 3.9. *(i) Lorsque $x_n \leqslant c_\varepsilon n^{1/(1+\varepsilon)}$, alors*

$$\log \Pi_{n,1}(x_n,\varepsilon) \underset{n\to+\infty}{\sim} -\frac{1}{2\sigma^2}\frac{x_n^2}{n}. \tag{3.16}$$

(ii) Lorsque $c_\varepsilon n^{1/(1+\varepsilon)} \leqslant x_n \leqslant c_n n^{1/(1+\varepsilon)}$, alors

$$\log \Pi_{n,1}(x_n,\varepsilon) \underset{n\to+\infty}{\sim} -(1-\alpha_1)^{1-\varepsilon} x_n^{1-\varepsilon} - \frac{\alpha_1^2 x_n^2}{2\sigma^2 n}, \tag{3.17}$$

où $c_n \to +\infty$ lorsque $n \to +\infty$ et c_ε et α_1 sont donnés par (3.9).
(iii) Lorsque $x_n \gg n^{1/(1+\varepsilon)}$, alors

$$\log \Pi_{n,1}(x_n,\varepsilon) \underset{n\to+\infty}{\sim} -x_n^{1-\varepsilon}. \tag{3.18}$$

Enfin, il reste à établir le comportement asymptotique de $R'_{n,2}(x_n)$. C'est l'équivalent du Lemme 5 technique de [187]. La preuve se fait en considérant tout d'abord les m tels que $m \geqslant 2x_n^{1-\varepsilon} =: \hat{m}_n$ (cas facile) pour lesquels on montre que

$$\Pi_{n,m}(x_n, \varepsilon) \leqslant e^{-2x_n^{1-\varepsilon}}$$

puis que

$$\sum_{m=\hat{m}_n}^{n} \binom{n}{m} \Pi_{n,m}(x_n, \varepsilon) \leqslant (Const) e^{-x_n^{1-\varepsilon}} \leqslant (Const) P_{n,1}(x_n)$$

dans les différents régimes. Après quelques calculs techniques et un peu d'analyse convexe, nous montrons enfin que $\sum_{m=\hat{m}_n}^{n} \binom{n}{m} \Pi_{n,m}(x_n, \varepsilon) \leqslant (Const) P_{n,1}(x_n)$.

3.4 Le hachage avec essais linéaires

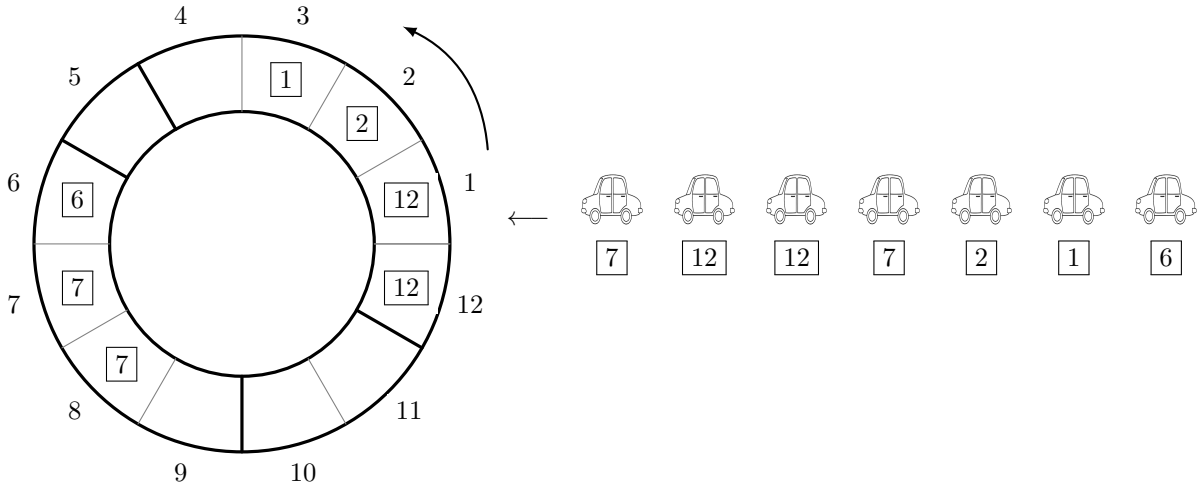
3.4.1 Le modèle de hachage

Dans cette section, je présente plus précisément le modèle de hachage avec essais linéaires. J'utilise ici la terminologie du problème de parking introduite initialement par Rényi et formalisée par Knuth [146] de la façon suivante :

> " A certain one-way street has m parking spaces in a row numbered 1 to m. A man and his dozing wife drive by, and suddenly, she wakes up and orders him to park immediately. He dutifully parks at the first available space [...]. "

La question est alors de déterminer la distance moyenne entre l'endroit où l'on se gare et l'endroit où l'on souhaitait se garer initialement. Naturellement, on s'attend à ce que se garer dans une rue quasiment déserte ($n = o(m)$) soit facile. En revanche, cela devient compliqué lorsque la rue se remplit ($n \approx m$).

Pour fixer les idées, considérons l'exemple suivant. Supposons que $n = 7$, $m = 12$, et (7,12,12,7,2,1,6) sont les adresses où les voitures arrivent. Cette suite d'adresses est appelée *suite de hachage* de longueur m et de taille n. Soit d_i le déplacement de la voiture i. Alors $d_1 = d_2 = 0$. La voiture 3 souhaite arriver sur la place 12 qui est occupée par la deuxième voiture ; ainsi, elle est déplacée d'un pas en avant et se gare sur la place 1 de sorte que $d_3 = 1$. La voiture 4 arrive sur la place 7 qui est occupée, de sorte que $d_4 = 1$. Et ainsi de suite : $d_5 = 0$, $d_6 = 2$, $d_7 = 0$. Ici, le déplacement total $d_{12,7}$ est donc égal à $1 + 1 + 2 = 4$. Rappelons qu'une fois toutes les voitures garées, il y a $N = m - n$ places vides. Celles-ci séparent les places occupées en blocs de places consécutives. Nous considérons que la place vide qui suit un bloc appartient à ce bloc. Dans notre exemple, il y a deux blocs : le premier contenant les places $12, 1, 2, 3$ (occupées) et la place 4 (libre) et le second contenant les places $6, 7, 8$ (occupées) et la place 9 (libre). Cet exemple est illustré dans le graphique suivant.



Comme nous l'avons vu précédemment, les longueurs des blocs et le déplacement total à l'intérieur de chaque bloc sont distribués comme (X_1, Y_1), ..., (X_N, Y_N) conditionnés à $\{\sum_{i=1}^N X_i = m\}$, où les vecteurs aléatoires (X_i, Y_i) sont des copies i.i.d. d'un vecteur (X, Y) de v.a. telles que X suit la distribution de Borel de paramètre $\mu \in]0\,,1[$ arbitraire et Y sachant $\{X = l\}$ distribué comme $d_{l,l-1}$. Remarquons que la distribution conditionnelle de Y sachant X ne dépend pas du paramètre μ. La distribution de Borel est donnée par

$$\mathbb{P}(X = l) = e^{-\mu l}\frac{(\mu l)^{l-1}}{l!},\ \mu \in]0\,,1[\,. \tag{3.19}$$

L'espérance et la variance de X sont donnés par

$$\mathbb{E}[X] = \frac{1}{1-\mu} \quad \text{et} \quad \mathrm{Var}(X) = \frac{\mu}{(1-\mu)^3}.$$

Dans certains cas, pour faciliter les calculs, nous pourrons aussi utiliser la paramétrisation $\lambda = e^{-\mu}\mu$ donnant la définition équivalente de la distribution de Borel :

$$\mathbb{P}(X = l) = \frac{1}{T(\lambda)}\frac{l^{l-1}\lambda^l}{l!}, \tag{3.20}$$

où T est la fonction arbre, "tree function" en anglais, (see, e.g., [**?** , p.127]).
En particulier, le déplacement total $d_{m,n}$ est distribué comme $\sum_{i=1}^N Y_i$ conditionné à $\{\sum_{i=1}^N X_i = m\}$; ce qui nous permettra par la suite d'avoir un peu de latitude et de se placer éventuellement à la moyenne. Nous présentons dans le lemme suivant quelques propriétés sur le déplacement total $d_{m,n}$.

Lemme 3.10.

1) Le nombre de séquences de hachage de longueur m et de taille n est m^n.

2) Nous avons $0 \leqslant d_{m,n} \leqslant \frac{n(n-1)}{2}$.

3) Le déplacement total de toute suite de hachage $(h_1, \ldots, h_n)$ est invariant par permutation des h_i. Plus précisément, pour toute permutation σ de $\{1, \ldots, n\}$, le déplacement total associé à la suite de hachage $(h_1, \ldots, h_n)$ est le même que le déplacement total associé à la suite de hachage $(h_{\sigma(1)}, \ldots, h_{\sigma(n)})$.

4) $d_{l,l} = d_{l,l-1} + U$ où U est une v.a. uniforme sur $\{1, \ldots, l-1\}$.

Les deux premiers points et le dernier sont évidents tandis que le troisième est une conséquence directe de [131, Lemma 2.1].

3.4.2 Résultats théoriques pour les tables pleines

Dans ce contexte,

$$\mathbb{E}[d_{n,n}] \underset{n}{\sim} \frac{\sqrt{2\pi}}{4} n^{3/2} \quad \text{et} \quad \mathrm{Var}(d_{n,n}) \underset{n}{\sim} \frac{10-3\pi}{24} n^3,$$

par [97, Theorem 2].

Théorème 3.11 (Déviations standard - Theorem 3 [97]). *Pour les tables pleines, la distribution normalisée du déplacement total $d_{n,n}/(n/2)^{3/2}$ est asymptotiquement distribué comme la loi d'Airy ; à savoir que, ponctuellement pour chaque $x \geqslant 0$, nous avons*

$$\mathbb{P}(d_{n,n} \leqslant n^{3/2}x) \underset{n\to+\infty}{\to} \mathbb{P}(A \leqslant x)$$

où la v.a. A suit la loi d'Airy définie dans [97]. Naturellement,

$$\mathbb{P}(d_{n,n} \geqslant n^{3/2}x) \underset{n\to+\infty}{\to} 1 - \mathbb{P}(A \leqslant x).$$

Nous nous intéressons maintenant aux déviations supérieures non standard [1] pour lesquelles nous avons établis les deux théorèmes qui suivent.

Théorème 3.12 (Déviations modérées). *Pour tout $\alpha \in]3/2\,,2[$ et pour tout $x \in [0,+\infty[$, nous avons*

$$\frac{1}{n^{2\alpha-3}} \log \mathbb{P}(d_{n,n} \geqslant n^{\alpha}x) \underset{n\to+\infty}{\to} -6x^2.$$

Théorème 3.13 (Grandes déviations). *Pour tout $x \in [0\,,1/2[$, nous avons*

$$\frac{1}{n} \log \mathbb{P}(d_{n,n} \geqslant n^{2}x) \underset{n\to\infty}{\to} -I(x),$$

où I est défini par $I(x) = \lambda_x(1/2-x) + \log(1-\lambda_x(1/2+x))$ avec λ_x solution de

$$\left(\frac{1}{\lambda_x} - x - \frac{1}{2}\right)(e^{\lambda_x} - 1) = 1.$$

Notons que les déviations modérées et les grandes déviations inférieures sont triviales : pour tout $\alpha \in]3/2\,,2]$ et tout n assez grand,

$$\mathbb{P}(d_{n,n} - \mathbb{E}[d_{n,n}] \leqslant -n^{\alpha}x) = 0.$$

en raison de la positivité de $d_{n,n}$ et du comportement asymptotique de $\mathbb{E}[d_{n,n}]$ en $n^{3/2}$. En ce qui concerne les très grandes déviations, le même phénomène se produit : pour tout $\alpha > 2$ et tout n assez grand,

$$\mathbb{P}(d_{n,n} - \mathbb{E}[d_{n,n}] \geqslant n^{\alpha}x) = 0 \quad \text{et} \quad \mathbb{P}(d_{n,n} - \mathbb{E}[d_{n,n}] \leqslant -n^{\alpha}x) = 0.$$

puisque $d_{n,n} \leqslant n(n-1)/2$.

Remarque 3.14. *Clairement, les conclusions des Théorèmes 3.11, 3.12 et 3.13 sont toujours valables en remplaçant $d_{n,n}$ par $d_{m,n}$ du moment que $m - n \ll n^{\alpha-1}$. En effet, en couplant naturellement $d_{m,n}$ à*

1. Comme nous le verrons à la suite des deux théorèmes suivants, les déviations inférieures non standard sont triviales.

$d_{m,m}$ en ajoutant $m-n$ voitures, nous avons

$$\begin{aligned}|d_{m,n}-d_{m,m}| &\leqslant (m-1)+(m-2)+\ldots+n\\ &= \frac{(m+n-1)(m-n)}{2}\\ &\sim n(m-n)\\ &\ll n^{\alpha}.\end{aligned}$$

Ainsi nos résultats complètent les déviations standard établies dans [131], dans le cas dense.

Quelques éléments de preuves des Théorèmes 3.12 et 3.13

Pour tout $n \geqslant 1$, soit $(V_{n,i})_{1\leqslant i\leqslant n}$ une suite de v.a. i.i.d. uniformément distribuées sur $[\![1\,,n]\!]$. La v.a. $V_{n,i}$ correspond à l'adresse de hachage de la voiture i. Nous définissons ensuite

$$S_n(k)=\sum_{i=1}^{n}\mathbb{1}_{V_{n,i}\leqslant k}.$$

Maintenant soit $(U_i)_{1\leqslant i\leqslant n}$ une suite de v.a. i.i.d. de distribution uniforme sur $[0\,,1]$. En suivant [261], nous introduisons la mesure empirique L_n associée aux v.a. U_i.

Par [131, Equation 2.1 and Lemma 2.1], le déplacement total est donné par

$$\begin{aligned}d_{n,n} &= \sum_{k=1}^{n}\left(S_n(k)-k-\min_{1\leqslant l\leqslant n}\{S_n(l)-l\}\right)\\ &= n\sum_{k=1}^{n}(L_n-\mathcal{U})([0\,,k/n])-n^2\min_{1\leqslant l\leqslant n}\{(L_n-\mathcal{U})([0\,,l/n])\}\\ &= n^2\left(\int_0^1(L_n-\mathcal{U})([0\,,y])dy-\inf_{\substack{0\leqslant y\leqslant 1,\\ y\in\mathbb{Q}}}\{(L_n-\mathcal{U})([0\,,y])\}+\frac{a}{n}\right)\end{aligned}$$

où $a\in[-1\,,1]$.

Traitons les déviations modérées et appliquons [261, Equation (1.3)] :

$$\begin{aligned}&\frac{1}{n^{2\alpha-3}}\log\mathbb{P}\left(d_{n,n}\geqslant xn^{\alpha}\right)\\ &\quad=\frac{1}{n^{2\alpha-3}}\log\mathbb{P}\left(\frac{\sqrt{n}}{n^{\alpha-3/2}}\left(\int_0^1(L_n-\mathcal{U})([0\,,y])dy-\min_{0\leqslant y\leqslant 1}\{(L_n-\mathcal{U})([0\,,y])\}+\frac{a}{n}\right)\geqslant x\right)\\ &\quad=\frac{1}{n^{2\alpha-3}}\log\mathbb{P}\left(\varphi\left(\frac{\sqrt{n}}{n^{\alpha-3/2}}\left(L_n-\mathcal{U}\right)\right)\geqslant x-\frac{a}{n^{\alpha-1}}\right)\\ &\quad=\frac{1}{n^{2\alpha g-3}}\log\mathbb{P}\left(\varphi\left(\frac{\sqrt{n}}{n^{\alpha-3/2}}\left(L_n-\mathcal{U}\right)\right)\geqslant x\right)\\ &\quad\to-\inf\left\{\frac{1}{2}\int_0^1\left(\frac{d\nu}{dy}(y)\right)^2dy\ ;\ \nu\in\mathcal{M}_b([0\,,1]),\nu\ll\mathcal{U},\varphi(\nu)\geqslant x,\nu([0\,,1])=0\right\}\\ &\quad=-\inf\left\{\frac{1}{2}\int_0^1G'(y)^2dy\ ;\ G\in AC([0\,,1]),\int_0^1G(y)dy-\min G\geqslant x,G(0)=G(1)=0\right\}\end{aligned}\tag{3.21}$$

où φ est une fonction mesurable, $\mathcal{M}_b([0\,,1])$ est l'ensemble des mesures signées à variation finie sur $[0\,,1]$ muni de la τ-topologie et AC est l'ensemble des fonctions absolument continues. Il suffit ensuite de résoudre le problème variationnel (3.21) en utilisant les multiplicateurs de Lagrange et en appliquant le lemme de Du Bois-Reymond [64, p.184].

Pour les grandes déviations, nous appliquons le théorème de Sanov [261, Equation (1.1)] et nous sommes conduits à résoudre le problème variationnel suivant :

$$\begin{aligned}
&-\inf\left\{\int_0^1 \left(\frac{d\nu}{dy}(y)\right)\log\left(\frac{d\nu}{dy}(y)\right)dy\ ;\ \nu\in\mathcal{M}_1^+([0\,,1]),\nu\ll\mathcal{U},\varphi(\nu-\mathcal{U})\geqslant x\right\}\\
&=-\inf\Bigg\{\int_0^1 F'(y)\log F'(y)dy\ ;\ F\in AC([0\,,1]),F'\geqslant 0,\int_0^1(F-\mathrm{id})(y)dy-\min(F-\mathrm{id})\geqslant x,\\
&\qquad\qquad F(0)=0,F(1)=1\Bigg\},
\end{aligned}\tag{3.22}$$

où $\mathcal{M}_1^+([0\,,1])$ est l'espace des mesures de probabilités sur $[0\,,1]$. Sa résolution se fait en utilisant des outils d'optimisation convexe. □

3.4.3 Résultats intermédiaires

Dans cette section, nous déterminons les comportements asymptotiques des queues de X_n et de Y_n qui nous seront utiles par la suite ainsi qu'un résultat de limite locale pour S_n. De la même façon que dans la Section 3.2, nous supposons dorénavant que $m=m_n\to\infty$ et $N=N_n=m_n-n\to\infty$ avec $\mu_n=n/m_n\in\,]0\,,1[\,\to\mu\in\,]0\,,1[$. Soient $(X_{n,i},Y_{n,i})_{1\leqslant i\leqslant N_n}$ des copies i.i.d. de (X_n,Y_n), où X_n suit la loi de Borel de paramètre μ_n (voir (3.19) pour la definition) et Y_n sachant $X_n=l$ est distribué comme $d_{l,l-1}$. Notons que, naturellement, la convergence de μ_n vers μ entraîne celle du paramètre λ_n vers $\lambda:=e^{-\mu}\mu$ dans la définition équivalente de la distribution de Borel (voir (3.20)).

Le déplacement total $d_{m_n,n}$ est donc distribué comme la distribution conditionnelle de T_n sachant $S_n=m_n$.

Nous commençons par déterminer le comportement asymptotique de la queue de X_n dont la distribution est donnée par (3.19).

Proposition 3.15 (Queue de X_n)**.** *Pour tout $l\geqslant 1$,*

$$\log\mathbb{P}(X_n=l)\leqslant-\kappa_n l\tag{3.23}$$

où $\kappa_n=\mu_n-\log(\mu_n)-1\in\,]0\,,\infty[$. De plus, si $l_n\to+\infty$, alors

$$\log\mathbb{P}(X_n\geqslant l_n)\sim\log\mathbb{P}(X_n=l_n)\sim-\kappa l_n\tag{3.24}$$

avec $\kappa=\lim\kappa_n=\mu-\log(\mu)-1\in\,]0\,,\infty[$.

Poursuivons avec le comportement asymptotique de la queue de Y_n. Nous commençons par montrer la borne supérieure grossière suivante

$$\limsup_{n\to+\infty}\frac{1}{\sqrt{y_n}}\log\mathbb{P}(Y_n\geqslant y_n)\leqslant-\kappa\sqrt{2}.\tag{3.25}$$

qui nous permettra ensuite d'obtenir le résultat précis de la Proposition 3.16. La preuve repose sur la décomposition triviale suivante :

$$\mathbb{P}(Y_n\geqslant y)=\sum_{n=l_n}^{+\infty}\mathbb{P}(d_{n+1,n}\geqslant y)\mathbb{P}(X_n=n+1)\tag{3.26}$$

où l_n est défini par

$$l_n = \left\lceil \sqrt{2y_n + \frac{1}{4}} + \frac{1}{2} \right\rceil . \tag{3.27}$$

Le résultat vient directement en majorant $\mathbb{P}(d_{n+1,n} \geqslant y)$ par 1 et en utilisant la Proposition 3.15.
La proposition suivante donne le comportement exact à l'échelle logarithmique de la queue de Y_n. Elle est basée sur la décomposition (3.26), la majoration (3.25), la propriété 4) du Lemme 3.10 qui permet de se ramener aux tables pleines ($m = n = l$) à partir des tables quasi-pleines ($m = l$, $n = l - 1$) et le Théorème 3.13 traitant les grandes déviations pour les tables pleines justement.

Proposition 3.16 (Comportement asymptotique de la queue de Y_n)**.** *Si $y_n \to +\infty$, alors*

$$\frac{1}{\sqrt{y_n}} \log \mathbb{P}(Y_n \geqslant y_n) \underset{n\to+\infty}{\to} - \inf_{\delta>0} \sqrt{\frac{1}{\delta}}\big(\kappa + I(\delta)\big)$$

où $\kappa = \mu - \log(\mu) - 1 \in]0\,,\infty[$ et I a été défini dans le Théorème 3.13.

Proposition 3.17 (Théorème de la limite locale pour S_n)**.** *Supposons qu'il existe une constante $c_3 > 0$ telles que (H3) dans les Hypothèses 3.1 est satisfaite et que*

(H8) il existe une constante $c_8 > 0$ telle que, pour tout $s \in [-\pi\,, \pi]$,

$$\left|\mathbb{E}\left[e^{isX_n}\right]\right| \leqslant 1 - c_8\sigma_{X_n}^2 s^2;$$

(H9) $m_n = N_n\mathbb{E}[X_n] + O(\sigma_{X_n}N_n^{1/2})$ (rappelons que $m_n \in \mathbb{Z}$ et $\mathbb{P}(S_n = m_n) > 0$);

Alors il existe $c_9 > 0$ tel que

$$\mathbb{P}(S_n = m_n) \geqslant \frac{c_9}{2\pi\sigma_{X_n}N_n^{1/2}}.$$

L'hypothèse (H8) concerne X uniquement ; c'est l'analogue de (H7) pour le couple (X, Y'). Le théorème limite local établi dans la Proposition 3.17 est crucial pour les preuves des résultats de grandes déviations (Théorème 3.21 (iii)) et pour celle de la borne de Berry-Esseen (Théorème 3.3).
Notons aussi que dans la preuve des déviations modérées supérieures du Théorème 3.21 (i), nous aurons besoin du comportement exact (pas seulement d'une minoration). De même que pour la minoration pour les grandes déviations supérieures du Théorème 3.21 (ii) où il faudra en outre le comportement "loin" de la moyenne.

3.4.4 Résultats théoriques pour les tables creuses

Supposons encore que $n/m_n \to \mu \in]0\,,1[$ et que $N_n = m_n - n$. Alors d'après [97, Theorem 5],

$$\mathbb{E}[d_{m_n,n}] \underset{n}{\sim} \frac{\mu^2}{2(1-\mu)^2}N_n \quad \text{et} \quad \mathrm{Var}(d_{m_n,n}) \underset{n}{\sim} \sigma^2(\mu)N_n,$$

où (cf. [97, Theorem 5]).

$$\sigma^2(\mu) := \frac{6\mu^2 - 6\mu^3 + 4\mu^4 - \mu^5}{12(1-\mu)^5}.$$

Théorème 3.18 (Déviations standard - Theorem 6 [97])**.** *La distribution du déplacement total $d_{m,n}$ est asymptotiquement gaussiennes ; à savoir que, ponctuellement pour chaque x,*

$$\mathbb{P}\left(d_{m_n,n} - \mathbb{E}[d_{m_n,n}] \leqslant N_n^{1/2}x\right) \underset{n\to+\infty}{\to} \mathbb{P}(Z \leqslant x)$$

où Z est une gaussienne centrée de variance $\sigma^2(\mu)$). Naturellement,

$$\mathbb{P}\left(d_{m_n,n} - \mathbb{E}[d_{m_n,n}] \geqslant N_n^{1/2} x\right) \underset{n\to+\infty}{\to} 1 - \mathbb{P}(Z \geqslant x)$$

Nous nous intéressons maintenant aux déviations inférieures et supérieures non standard pour lesquelles nous avons établis les trois théorèmes qui suivent.

Théorème 3.19 (Déviations modérées inférieures)**.** *Pour tout $\alpha \in]1/2\,,1[$ et tout $y > 0$,*

$$\frac{1}{N_n^{2\alpha-1}} \log \mathbb{P}(d_{m_n,n} - \mathbb{E}[d_{m_n,n}] \leqslant -N_n^\alpha x) \underset{n\to+\infty}{\to} -\frac{x^2}{2\sigma^2(\mu)}. \tag{3.28}$$

Théorème 3.20 (Grandes déviations inférieures)**.** *Pour tout $y > 0$,*

$$\frac{1}{N_n} \log \mathbb{P}(d_{m_n,n} - \mathbb{E}[d_{m_n,n}] \leqslant -N_n x) \underset{n\to+\infty}{\to} -\Lambda^*_{(X,Y)}\left(\frac{1}{1-\mu}, x\right), \tag{3.29}$$

où Λ^* est la transformée de Fenchel-Legendre de

$$\Lambda_{(X,Y)}(\lambda,\rho) = \log \sum_{n=1}^{\infty} \frac{e^{(\lambda-\mu)n}(\mu n)^{n-1}}{n!} \mathbb{E}[e^{\rho d_{n,n-1}}]$$

Pour tout $\alpha > 1$ et tout n assez grand,

$$\mathbb{P}(d_{m_n,n} - \mathbb{E}[d_{m_n,n}] \leqslant -N_n^\alpha x) = 0$$

puisque $d_{m_n,n} \geqslant 0$ et $\mathbb{E}[d_{m_n,n}]$ est linéaire en N_n.

Théorème 3.21 (Déviations supérieures)**.**
(i) Pour tout $\alpha \in]1/2\,,2/3[$ et tout $y > 0$,

$$\frac{1}{N_n^{2\alpha-1}} \log \mathbb{P}(d_{m_n,n} - \mathbb{E}[d_{m_n,n}] \geqslant N_n^\alpha x) \underset{n\to+\infty}{\to} -\frac{x^2}{2\sigma(\mu)^2}. \tag{3.30}$$

(ii) Pour tout $\alpha \in]2/3\,,2[$ et tout $y > 0$,

$$\frac{1}{N_n^{\alpha/2}} \log \mathbb{P}(d_{m_n,n} - \mathbb{E}[d_{m_n,n}] \geqslant N_n^\alpha x) \underset{n\to+\infty}{\to} -x^{1/2} \inf_{\delta>0} \frac{1}{\sqrt{\delta}}\big(\kappa + I(\delta)\big) \tag{3.31}$$

où $\kappa = \mu - \log(\mu) - 1 \in]0\,,+\infty[$ et I a été défini dans le Théorème 3.13.

Pour tout $\alpha > 2$ et n assez grand,

$$\mathbb{P}(d_{m_n,n} - \mathbb{E}[d_{m_n,n}] \geqslant N_n^\alpha x) = 0$$

puisque $d_{m_n,n} \leqslant n(n-1)/2$ et $\mathbb{E}[d_{m_n,n}]$ est linéaire en N_n.

Remarquons que nous obtenons comme pour le Théorème 3.7 trois régimes principaux : le cas (i), le cas (ii) et le cas intermédiaire $\alpha = 2/3$ (non traité pour l'instant). Dans les preuves, nous utilisons une fois de plus l'égalité en distribution de $d_{m_n,n}$ et de T_n conditionné à $\{S_n = m_n\}$ et nous notons que le conditionnement par $\{S_n = m_n\}$ ne joue finalement pas dans les grandes déviations (cas (ii) du Théorème 3.21). En revanche, il contribue pour les déviations modérées (cas (i) du Théorème 3.21).
Il reste à prouver les cas frontières. Le cas $\alpha = 2/3$ est l'équivalent pour les sommes conditionnées du cas

intermédiaire et difficile (ii) dans le Théorème 3.7. Quant à $\alpha = 2$, il est facile de montrer que

$$\mathbb{P}(d_{m_n,n} - \mathbb{E}[d_{m_n,n}] \geqslant N_n^2 x) = \Theta\left(e^{-(Const)N_n \log N_n}\right)$$

et il reste à déterminer la constante.

Quelques éléments de preuve pour les déviations inférieures (Théorèmes 3.19 et 3.20)
Les preuves des Théorèmes 3.19 et 3.20 se montrent en utilisant l'égalité en distribution de $d_{m_n,n}$ et de T_n conditionné à $\{S_n = m_n\}$, le théorème de Plachky-Steinebach [204] (version unilatérale du Théorème de Gärtner-Ellis) et en adaptant au cas unilatéral les calculs de [101]. □

Quelques éléments de preuve pour les déviations modérées supérieures (Théorème 3.21 (i))
Remarquons tout d'abord que la liberté dans le choix du paramètre de la loi de Borel, mentionnée dans l'introduction de ce chapitre, va nous permettre de nous placer à la moyenne pour tout n. Concrètement cela signifie que nous avons choisi μ_n de telle sorte que $\mathbb{E}[S_n] = m_n$ et que nous étudions le conditionnement à la moyenne $\{S_n = m_n\}$. De la sorte, les calculs seront plus simples à mener. Par exemple, le changement de loi introduit pour cette preuve sera trivial (voir pour comparaison la preuve de [101, Théorème 2.2] dont nous suivons le déroulement ici).

Plus précisément, pour la minoration, nous suivons la preuve de [101, Theorem 2.2] et en vue d'appliquer le théorème de Gärtner-Ellis, nous introduisons la fonction

$$g_n(u) = \frac{1}{N_n^{2\alpha-1}} \log \mathbb{E}\left[e^{u(T_n - \mathbb{E}[T_n])/N_n^{1-\alpha}} \mid \forall i,\ Y_{n,i} < N_n^{\alpha/2},\ S_n = m_n\right]$$

et les v.a. conditionnées $Y_n^< = Y_n | Y_n < N_n^{\alpha/2}$. Par la formule de Bartlett [26, Equation 16], une étude de fonction fine reposant sur le Lemme 3.10 et en prouvant que $(X_n, Y_n)_{n\geqslant 1}$ converge en distribution avec tous les moments croisés, nous montrons que

$$g_n(u) \to \frac{u^2}{2}\sigma(\mu)^2. \tag{3.32}$$

Pour la majoration, guidés par [132], nous introduisons sans perte de généralité les v.a.

$$Y_n' = (Y_n - \mathbb{E}[Y_n]) + \frac{\mathrm{Cov}(X_n, Y_n)}{\mathrm{Var}(X_n)}(X_n - \mathbb{E}[X_n])$$

centrées et décorrélées des X_n. Ces variables avaient été déjà été considérées dans l'Hypothèse (H7) du le jeu d'Hypothèses nécessaires pour montrer les résultats de type Berry-Essen (Théorèmes 3.3 et 3.5 de la Section 3.1). Remarquons que la variance de Y_n' converge vers $\sigma^2(\mu)$ qui apparaît dans (3.32) et naturellement aussi dans la variance de $d_{m_n,n}$ rappelée en début de section. Nous vérifions que

$$\frac{1}{\sqrt{y_n}} \log \mathbb{P}(Y_n \geqslant y_n) \underset{n\to\infty}{\sim} \frac{1}{\sqrt{y_n}} \log \mathbb{P}(Y_n' \geqslant y_n)$$

pour toute suite $y_n \to +\infty$. Ensuite, par des minorations et l'utilisation de la Proposition 3.17 concernant la limite locale de S_n et du comportement asymptotique de la queue de Y_n donné dans la Proposition 3.16, nous obtenons la majoration. □

Quelques éléments de preuve pour les grandes déviations supérieures (Théorème 3.21 (ii))
Pour la majoration, il suffit d'écrire

$$\mathbb{P}(d_{m_n,n} - \mathbb{E}[d_{m_n,n}] \geqslant N_n^\alpha y) \leqslant \frac{\mathbb{P}(T_n - \mathbb{E}[T_n] \geqslant N_n^\alpha y_n)}{\mathbb{P}(S_n = m_n)},$$

où $y_n = y + \frac{1}{n}(\mathbb{E}[T_n|S_n = m_n] - \mathbb{E}[T_n])$. La conclusion découle des Propositions 3.16, d'une version du Théorème 3.7 (cas (iii)) adaptée aux tableaux de v.a. discrètes et du résultat de limite locale pour S_n établi dans la Proposition 3.17.
Pour la minoration, nous supposons que l'infimum dans le membre de droite de (3.31) est atteint en δ_0. Soit $\varepsilon > 0$ et $l_n = \lceil (n^\alpha(y_n + 2\varepsilon)/\delta_0)^{1/2} \rceil$ de sorte que

$$\begin{aligned}\mathbb{P}(d_{m_n,n} - \mathbb{E}[d_{m_n,n}] \geqslant N_n^\alpha y) \geqslant \mathbb{P}\big(T_{n,N_n-1} - \mathbb{E}[T_{n,N_n-1}] \geqslant -N_n^\alpha \varepsilon,\ S_{n,N_n-1} = m_n - l_n\big)\\ \mathbb{P}\big(Y_n - \mathbb{E}[Y_n] \geqslant N_n^\alpha (y_n + \varepsilon),\ X_n = l_n\big).\end{aligned}$$

Il suffit ensuite d'utiliser la propriété 4) du Lemme 3.10 pour se ramener aux tables pleines et d'appliquer le Théorème 3.13 pour les tables pleines afin de minorer le second terme du membre de droite de l'inégalité précédente. Quant au premier terme, il est minoré par

$$\begin{aligned}&\mathbb{P}\big(T_{n,N_n-1} - \mathbb{E}[T_{n,N_n-1}] \geqslant -N_n^\alpha \varepsilon,\ S_{n,N_n-1} = m_n - l_n\big)\\ &\qquad \geqslant \mathbb{P}\big(S_{n,N_n-1} = m_n - l_n\big) - \mathbb{P}\big(T_{n,N_n-1} - \mathbb{E}[T_{n,N_n-1}] < -N_n^\alpha \varepsilon\big).\end{aligned}$$

Pour conclure, on utilise le théorème de la limite locale ("loin" de la moyenne) établi dans le Lemme 3.3 de [101] pour traiter la première probabilité et une version unilatérale du Théorème 2.2 [101] pour traiter la seconde. □

Soulignons les principales différences entre les résultats de grandes déviations de ce chapitre et ceux de [101, Theorem 2.1]. D'abord, la preuve de [101, Theorem 2.1] est basée sur un contrôle précis de la transformée de Fourier-Laplace $\Phi_{X_n,Y_n}(t,u) = \mathbb{E}\left[\exp\{itX_n + uY_n\}\right]$ de (X_n, Y_n). La partie Fourier permet de traiter le conditionnement tandis que la partie Laplace permet d'appliquer le théorème de Gärtner-Ellis. Dans notre travail, la preuve suit des idées empruntées à [186, 187]. Contrairement au cas où la transformée de Laplace est définie, les déviations importantes de la somme de v.a. à queue lourde sont dues à des valeurs exceptionnelles prises par quelques variables aléatoires. Ensuite, contrairement aux vitesses classiques en N_n obtenues dans le théorème de Cramér ou dans le Théorème 2.1 de [101], les vitesses que nous obtenons ne sont pas nécessairement N_n.

Chapitre 4

Analyse de sensibilité et quantification des incertitudes

Ce chapitre est consacré à l'analyse de sensibilité et à la quantification d'incertitudes. J'y présente les résultats de collaborations fructueuses avec Jean-Claude Fort (Paris V), Fabrice Gamboa (IMT), Alexandre Janon (Université Paris Sud), Thierry Klein (IMT-ENAC), Béatrice Laurent (IMT-INSA), Leonardo Moreno (Université Montevideo, Uruguay), Maëlle Nodet (LJK) et Clémentine Prieur (LJK), initiées par l'ANR Costa Brava (2009-2014) dont le porteur était Fabrice Gamboa. Le détail ainsi que les preuves des théorèmes et propositions se trouvent dans les publications [N2], [J5], [J6], [J7], [J10], [J14] et [P1].

4.1 La problématique de la quantification des incertitudes

De nombreux modèles mathématiques rencontrés dans le monde industriel font intervenir un grand nombre de paramètres parfois mal connus ou erronés (lors de la saisie par exemple). Plus précisément, la sortie du modèle, notée y, est l'image par une fonction f de variables d'entrées $x_1, \ldots, x_p$. Ces variables d'entrées peuvent être scalaires, vectorielles, fonctionnelles,... Le code de calcul f, bien que déterministe, n'est généralement pas connu (ou du moins est très complexe) et est pour cette raison appelé *boîte noire*. Concrètement, pour un jeu de paramètres $x_1, \ldots, x_p$, l'utilisateur peut déterminer la valeur de $y = f(x_1, \ldots, x_p)$; cependant, il n'a pas un accès direct à la fonction f. L'objectif est de connaître parfaitement la fonction f ; ce qui paraît difficile, d'autant que dans de nombreuses applications industrielles, l'évaluation d'une seule valeur de y peut s'avérer très coûteuse (plusieurs heures voire plusieurs jours). En pratique, l'utilisateur travaille en général à budget de calcul fixé, budget qui se traduit en termes d'un nombre d'appels au code limité. Il va donc être amené à choisir ses points d'expériences astucieusement de façon à obtenir le maximum d'information sur la fonction f. Une fois les points d'expérience choisis, il pourra être judicieux de construire une approximation du modèle, moins coûteuse. Ce modèle approché s'appelle un *métamodèle*. Dans le même ordre d'idée, une autre problématique consiste à déterminer les variables d'entrées les plus "importantes" sur la sortie du code afin de se concentrer ensuite sur celles-ci et éventuellement, procéder ensuite à la réduction du modèle. Concrètement on fixe les autres variables à leur valeur nominale et on consacre la majeure partie du budget à leur apprentissage. Toutes ces problématiques sont bien évidemment liées. Dans ce chapitre, je me concentre sur l'évaluation de l'impact de l'incertitude des paramètres d'entrée sur la sortie du modèle. L'analyse de sensibilité, qui a pour but d'identifier les paramètres les plus sensibles ; en d'autres termes, les paramètres ayant la plus grande influence sur la sortie du modèle.

Analyse de sensibilité locale

L'approche la plus naturelle pour analyser l'influence des paramètres d'entrée est de comparer les variations de la sortie du code y induites par des perturbations des entrées les unes après les autres autour de leur valeur nominale. L'exploitation des dérivées partielles en analyse de senbilité remonte aux années 60 (voir, par exemple, [114] pour des considérations historiques et [47] pour plus de détails techniques). Cette méthode appelée *analyse de sensibilité différentielle* est basée sur un développement de Taylor au premier ordre de la sortie. Plus précisément, si

$$y = f(x_1, \ldots, x_p) \quad \text{et} \quad y_\varepsilon = f(x_1(1+\varepsilon), \ldots, x_p(1+\varepsilon)),$$

alors

$$\frac{y_\varepsilon - y}{y} \approx \varepsilon \sum_{i=1}^{p} \frac{\partial f}{\partial x_i}(x_i)\frac{x_i}{y}.$$

La variation relative de la sortie y au point x s'écrit donc comme la somme de mesures d'importance, chacune quantifiant la variation due à un seul paramètre d'entrée. Pour des modèles relativement simples, le calcul des dérivées au point x peut être obtenu en même temps en utilisant son modèle adjoint. Pour des modèles plus complexes, des approximations par différences finies peuvent être utilisées.

Analyse de sensibilité globale

Une autre approche consiste à traduire l'incertitude sur les entrées x du modèle en considérant que x est aléatoire (dès lors noté X). Il s'ensuit que la sortie du modèle y devient elle aussi aléatoire. On la notera désormais Y. C'est l'expertise de l'utilisateur qui guidera le choix des lois de probabilités régissant les entrées du code. Lorsque les entrées du code sont indépendantes, la variance totale de la sortie du code peut alors être divisée en différentes variances partielles : il s'agit de la décomposition dite de Hoeffding, voir [249]. Chacune de ces variances partielles mesure l'incertitude sur la sortie induite par celle de la variable d'entrée correspondante. En considérant le rapport entre chaque variance partielle et la variance totale, nous obtenons une mesure de l'importance pour chaque variable d'entrée appelée *indice de Sobol au premier ordre* de la variable [235]. Les paramètres les plus sensibles peuvent alors être identifiés et classés comme étant les paramètres correspondants aux plus grands indices de Sobol.

Une fois les indices de Sobol définis, la question de leur estimation reste ouverte. Dans la pratique, il faut estimer (au sens statistique) ces indices à l'aide d'un échantillon fini (de taille généralement de l'ordre de centaines de milliers) d'évaluations de la sortie du modèle [116]. De nombreuses approches Monte-Carlo ou quasi Monte-Carlo ont été développées par les communautés des sciences expérimentales et de l'ingénieur. En particulier, une procédure d'estimation appelée *Pick-Freeze* a été proposée dans [235, 234]. Dans cette méthodologie, l'indice de Sobol est considéré comme le coefficient de régression entre la sortie du modèle et sa réplication Pick-Freeze (pick pour choisie et freeze pour gelée). Cette réplication est obtenue en fixant la valeur de la variable d'intérêt (variable choisie et gelée) et en échantillonnant les autres variables de manière indépendante. Les réplications échantillonnées sont ensuite combinées pour produire un estimateur de l'indice de Sobol. Cette méthode requiert par conséquent que la connaissance des distributions d'entrée ou qu'*a minima*, on puisse facilement générer les variables d'entrée. Nous supposerons ici que le choix des distributions des variables d'entrée est guidé par une connaissance approfondie du système apportée par des experts dans le domaine. Dans cette étude, nous ne discuterons pas des effets de ces choix sur l'analyse de sensibilité qui peuvent être importants en pratique. En d'autres termes, la robustesse de l'analyse de sensibilité quant à ces choix ne sera pas abordée.

Plan du chapitre

La définition rigoureuse des indices de Sobol dans le cadre d'une sortie scalaire est donnée en Section 4.2. Nous présentons ensuite le schéma d'estimation Pick-Freeze. Les indices de Sobol et leur estimation par le schéma Pick-Freeze ont été introduits depuis longtemps déjà mais jusqu'à maintenant leurs propriétés (asymptotiques et précises) n'avaient pas été étudiées théoriquement. C'est ce que nous nous proposons de faire dans cette section. Dans la Section 4.3, nous proposons une procédure d'analyse de sensibilité lorsque le modèle n'est pas accessible et que nous utilisons un métamodèle. Nous montrons en Section 4.4 qu'il est possible de faire mieux que les indices de Sobol dans des cas particuliers tel que le modèle de régression linéaire en utilisant des U-statistiques. La Section 4.5 est consacrée à la généralisation des indices de Sobol pour sorties scalaires aux sorties vectorielles, voire fonctionnelles. Dans la Section 4.6, nous définissons des indices basés sur la distribution de la sortie et pas seulement sur les moments d'ordre 2 comme les indices de Sobol. Enfin, en Section 4.7, la dernière généralisation aboutit à la définition d'indices de sensibilités sur des espaces métriques généraux.

L'objectif de ce travail est d'une part de proposer des indices adaptés et faciles à interpréter pour réaliser une analyse de sensibilité globale ainsi que des estimateurs faciles à mettre en œuvre. D'autre part, il s'agit de justifier théoriquement l'usage pratique de ces indices et de leurs estimateurs. Cette étude a été testée comme nous le verrons plus loin sur des exemples numériques jouets mais reste relativement théorique dans le sens où elle n'a pas été confrontée à des cas tests industriels.

Dans tout ce chapitre, nous supposons que les variables d'entrée $X_1, \dots, X_p$ sont indépendantes. En outre, nous supposerons que les distributions de ces variables sont connues ou qu'*a minima* nous sommes capables de les simuler.

4.2 Les indices de Sobol pour sortie scalaire

Cette section correspond aux résultats des articles [J10] et [J5].

4.2.1 La décomposition de Hoeffding de la variance

Pour tout entier p, notons l'entrée aléatoire par $X = (X_1, \dots, X_p)$, définie sur un espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$ et à valeurs dans un espace mesurable $E = E_1 \times \ldots \times E_p$. La sortie du code Y est définie par la relation suivante :

$$Y = f(X_1, \dots, X_p), \tag{4.1}$$

où $f : E \to \mathbb{R}$ est une fonction déterministe mesurable inconnue.

Le modèle de régression non paramétrique classique rentre dans ce cadre général. Il suffit en effet de considérer l'une des variables d'entrée comme le terme d'erreur (par exemple, $X_p = \varepsilon$) et de prendre $Y = g(X_1, \dots, X_{p-1}) + \varepsilon = f(X_1, \dots, X_p) = f(X)$.

Supposons que Y est de carré intégrable ($\mathbb{E}\left[Y^2\right] < \infty$) et n'est pas déterministe ($\mathrm{Var}(Y) \neq 0$). Soient maintenant $\mathbf{u}$ un sous-ensemble de $I_p = \{1, \dots, p\}$ et $\sim\mathbf{u}$ son complémentaire dans I_p. Par la suite, nous noterons $X_\mathbf{u} = (X_i, i \in \mathbf{u})$ et $E_\mathbf{u} = \prod_{i \in \mathbf{u}} E_i$.
Puisque les entrées aléatoires $X_1, \dots, X_p$ sont indépendantes, nous pouvons écrire la décomposition de Hoeffding de f (voir [249]) :

$$Y = f(X) = c + f_\mathbf{u}(X_\mathbf{u}) + f_{\sim\mathbf{u}}(X_{\sim\mathbf{u}}) + f_{\mathbf{u},\sim\mathbf{u}}(X_\mathbf{u}, X_{\sim\mathbf{u}}), \tag{4.2}$$

où $c \in \mathbb{R}$, $f_{\mathbf{u}} : E_{\mathbf{u}} \to \mathbb{R}$, $f_{\sim\mathbf{u}} : E_{\sim\mathbf{u}} \to \mathbb{R}$ et $f_{\mathbf{u},\sim\mathbf{u}} : E \to \mathbb{R}$ sont donnés par

$$c = \mathbb{E}[Y],\ f_{\mathbf{u}} = \mathbb{E}[Y|X_{\mathbf{u}}] - c,\ f_{\sim\mathbf{u}} = \mathbb{E}[Y|X_{\sim\mathbf{u}}] - c,\ f_{u,\sim\mathbf{u}} = Y - f_{\mathbf{u}} - f_{\sim\mathbf{u}} - c.$$

Il suffit ensuite de prendre la variance de chaque côté de (4.2) et d'utiliser l'orthogonalité dans L^2 pour obtenir

$$\mathrm{Var}(Y) = \mathrm{Var}(f(X)) = \mathrm{Var}(\mathbb{E}[Y|X_{\mathbf{u}}]) + \mathrm{Var}(\mathbb{E}[Y|X_{\sim\mathbf{u}}]) + \mathrm{Var}(Y - f_{\mathbf{u}} - f_{\sim\mathbf{u}}). \tag{4.3}$$

Enfin, il reste à renormaliser chaque terme de (4.3) par la variance de Y :

$$1 = \frac{\mathrm{Var}(\mathbb{E}[Y|X_{\mathbf{u}}])}{\mathrm{Var}(Y)} + \frac{\mathrm{Var}(\mathbb{E}[Y|X_{\sim\mathbf{u}}])}{\mathrm{Var}(Y)} + \frac{\mathrm{Var}(Y - f_{\mathbf{u}} - f_{\sim\mathbf{u}})}{\mathrm{Var}(Y)}.$$

L'indice de Sobol associé aux variables d'entrée $X_{\mathbf{u}}$, noté $S^{X_{\mathbf{u}}}$ ou plus simplement $S^{\mathbf{u}}$, est défini de la manière suivante :

$$S^{\mathbf{u}} = \frac{\mathrm{Var}\left(\mathbb{E}[Y|X_{\mathbf{u}}]\right)}{\mathrm{Var}(Y)} \in [0,1]. \tag{4.4}$$

Exemple 4.1 (Un exemple pour comprendre)**.** *Considérons le modèle suivant :*

$$Y = f(X) = X_1 + X_1X_2$$

où le vecteur d'entrée $X = (X_1, X_2, X_3)$ a des coordonnées indépendantes distribuées selon la loi gaussienne standard. Intuivement, il est clair que X_1 doit avoir plus d'influence que X_2 puisqu'elle apparaît deux fois : une fois seule dans le terme X_1 et une fois dans le terme d'intéraction X_1X_2. Les indices de Sobol sont ici :

$$\left(S^{\mathbf{1}}, S^{\mathbf{2}}, S^{\mathbf{3}}, S^{\mathbf{1,2}}\right) = (1/2, 0, 0, 1/2).$$

Il apparaît bien que X_1 est impliqué à l'ordre 1 ce qui se traduit par $S^{\mathbf{1}} \neq 0$, contrairement à X_2 et X_3 pour lesquels $S^{\mathbf{2}} = S^{\mathbf{3}} = 0$. Nous retrouvons aussi que X_1 apparaît à l'ordre 2 dans le terme X_1X_2 puisque $S^{\mathbf{1,2}} \neq 0$.

Interprétation et compléments

L'indice $S^{\mathbf{u}}$ quantifie l'influence à l'ordre 1 de la variable d'entrée $X_{\mathbf{u}}$ sur la sortie. Il est possible de définir les indices de Sobol d'ordre supérieur et totaux. La construction des premiers est directe en prenant pour $\mathbf{u}$ un ensemble d'indices quelconques. Par exemple, lorsque nous nous intéressons à l'influence des deux premières entrées X_1 et X_2, il suffit de prendre $\mathbf{u} = \{1, 2\}$. L'indice $S^{\mathbf{1,2}}$ quantifie l'influence des entrées X_1 et X_2 sur Y tandis $S^{\mathbf{1,2}} - S^{\mathbf{1}} - S^{\mathbf{2}}$ quantifie l'influence de l'intéraction entre X_1 et X_2 sur Y.

En outre, l'indice $S^{\mathbf{u}}$ est parfois appelés *closed index*. Il est à distinguer de l'indice défini de la façon suivante

$$\sum_{\mathbf{v}\subseteq\mathbf{u}} (-1)^{|\mathbf{u}|-|\mathbf{v}|} \frac{\mathrm{Var}(\mathbb{E}[Y|X_{\mathbf{u}}])}{\mathrm{Var}(Y)}$$

où $|\mathbf{u}|$ désigne le cardinal du sous-ensemble $\mathbf{u}$, qui quantifie la sensibilité de la sortie par rapport à $X_{\mathbf{u}}$ pris dans sa globalité, contrairement à $S^{\mathbf{u}}$ qui quantifie la sensibilité de la sortie par rapport à toutes les variables dont l'indice est dans $\mathbf{u}$.

Enfin, l'indice de Sobol total $S^{\mathbf{u},\mathrm{Tot}}$ par rapport à $\mathbf{u}$ est défini par

$$S^{\mathbf{u},\mathrm{Tot}} := 1 - S^{\sim\mathbf{u}} = 1 - \frac{\mathrm{Var}\left(\mathbb{E}[Y|X_{\sim\mathbf{u}}]\right)}{\mathrm{Var}(Y)}$$

et quantifie l'influence (totale) de $X_{\mathbf{u}}$ sur Y (à tous les ordres).

4.2.2 L'estimation Pick-Freeze des indices de Sobol

Pour tout X et tout sous-ensemble $\mathbf{u}$ de I_p, introduisons $X^{\mathbf{u}}$ le vecteur défini par $X_i^{\mathbf{u}} = X_i$ si $i \in \mathbf{u}$ et $X_i^{\mathbf{u}} = X_i'$ si $i \notin \mathbf{u}$ où X_i' est une copie indépendante de X_i. Posons alors

$$Y^{\mathbf{u}} := f(X^{\mathbf{u}}). \tag{4.5}$$

Remarquons d'une part que

$$S^{\mathbf{u}} = \frac{\mathrm{Cov}\,(Y, Y^{\mathbf{u}})}{\mathrm{Var}(Y)} \tag{4.6}$$

et d'autre part, en utilisant un résultat classique de régression, que nous avons la formulation variationnelle suivante :

$$S^{\mathbf{u}} = \underset{a\in\mathbb{R}}{\mathrm{argmin}} \left\{ \mathbb{E}\left[(Y^{\mathbf{u}} - \mathbb{E}[Y^{\mathbf{u}}]) - a(Y - \mathbb{E}[Y]) \right]^2 \right\}. \tag{4.7}$$

L'équation (4.6) conduit à une procédure d'esimation naturelle facile à mettre en œuvre. En effet, considérons le plan d'expérience de taille N suivant :

1) $(X_j)_{j=1,\dots,N}$ sont N copies indépendantes de X.

2) $(X_j^{\mathbf{u}})_{j=1,\dots,N}$ sont telles que $X_{i,j}^{\mathbf{u}} = X_i$ si $i \in \mathbf{u}$ et $X_{i,j}^{\mathbf{u}} = X_i'$ si $i \notin \mathbf{u}$ où les X_i' sont des copies indépendantes de X_i. [1]

Pour tout $j = 1, \dots, N$, on calcule ensuite

$$Y_j = f(X_{1,j}, \dots, X_{p,j}), \quad Y_j^{\mathbf{u}} = f(X_j^{\mathbf{u}}),$$

Une première estimation de $S^{\mathbf{u}}$. Au vu de (4.6), nous proposons le premier estimateur suivant :

$$\widehat{S}^{\mathbf{u}} = \frac{\frac{1}{N}\sum_{j=1}^N Y_j Y_j^X - \left(\frac{1}{N}\sum_{j=1}^N Y_j\right)\left(\frac{1}{N}\sum_{j=1}^N Y_j^X\right)}{\frac{1}{N}\sum_{j=1}^N Y_j^2 - \left(\frac{1}{N}\sum_{j=1}^N Y_j\right)^2}, . \tag{4.8}$$

Cet estimateur a déjà été étudié dans [122]. Il est montré qu'il est en pratique très utilisé.

Une seconde estimation de $S^{\mathbf{u}}$. De façon à tenir compte de toute l'information contenue dans l'échantillon, nous utilisons aussi l'échantillon $(X_j^{\mathbf{u}})_{j=1,\dots,N}$ pour estimer $\mathbb{E}[Y]$ et $\mathrm{Var}(Y)$ et nous proposons donc l'estimateur suivant :

$$\widehat{T}^{\mathbf{u}} = \frac{\frac{1}{N}\sum_{j=1}^N Y_j Y_j^{\mathbf{u}} - \left(\frac{1}{N}\sum_{j=1}^N \left(\frac{Y_j + Y_j^{\mathbf{u}}}{2}\right)\right)^2}{\frac{1}{N}\sum_{j=1}^N \left(\frac{(Y_j)^2 + (Y_j^{\mathbf{u}})^2}{2}\right) - \left(\frac{1}{N}\sum_{j=1}^N \left(\frac{Y_j + Y_j^{\mathbf{u}}}{2}\right)\right)^2}. \tag{4.9}$$

Cet estimateur a été introduit dans [182]. Dans [191, 194], l'auteur introduit de nouveaux estimateurs des indices de Sobol et compare numériquement leurs performances.

1. Ceci requiert donc la connaissance des distributions des variables d'entrée ou *a minima* que l'on sache générer ces variables. Rappelons en outre que les variables d'entrée sont supposées indépendantes.

4.2.3 Propriétés asymptotiques

Nous étudions maintenant les propriétés assymptotiques des estimateurs présentés ci-dessus.

Théorème 4.2 (Consistance et normalité asymptotique de $\widehat{S}^{\mathbf{u}}$ et $\widehat{T}^{\mathbf{u}}$)**.** *Les estimateurs $\widehat{S}^{\mathbf{u}}$ et $\widehat{T}^{\mathbf{u}}$ définis par* (4.8) *et* (4.9) *sont consistants pour estimer $S^{\mathbf{u}}$:*

$$\widehat{S}^{\mathbf{u}} \underset{N\to\infty}{\overset{p.s.}{\longrightarrow}} S^{\mathbf{u}} \quad et \quad \widehat{T}^{\mathbf{u}} \underset{N\to\infty}{\overset{p.s.}{\longrightarrow}} S^{\mathbf{u}}.$$

Supposons en outre que $\mathbb{E}[Y^4] < \infty$. Alors ils sont aussi asymptotiquement gaussiens :

$$\sqrt{N}\left(\widehat{S}^{\mathbf{u}} - S^{\mathbf{u}}\right) \underset{N\to\infty}{\overset{\mathcal{L}}{\longrightarrow}} \mathcal{N}_1\left(0, \sigma_S^2\right) \tag{4.10}$$

et

$$\sqrt{N}\left(\widehat{T}^{\mathbf{u}} - S^{\mathbf{u}}\right) \underset{N\to\infty}{\overset{\mathcal{L}}{\longrightarrow}} \mathcal{N}_1\left(0, \sigma_T^2\right) \tag{4.11}$$

où

$$\sigma_S^2 = \frac{\mathrm{Var}\left([Y - \mathbb{E}[Y]]\left[(Y^{\mathbf{u}} - \mathbb{E}[Y]) - S^{\mathbf{u}}(Y - \mathbb{E}[Y])\right]\right)}{\left(\mathrm{Var}(Y)\right)^2},$$
$$\sigma_T^2 = \frac{\mathrm{Var}\left([Y - \mathbb{E}[Y]][Y^{\mathbf{u}} - \mathbb{E}[Y]] - S^{\mathbf{u}}/2\left((Y - \mathbb{E}[Y])^2 + (Y^{\mathbf{u}} - \mathbb{E}[Y])^2\right)\right)}{(\mathrm{Var}(Y))^2}.$$

La proposition suivante reposant sur l'échangeabilité des variables Y et $Y^{\mathbf{u}}$, nous permet de comparer les deux estimateurs.

Proposition 4.3. *La variance asymptotique de $\widehat{T}^{\mathbf{u}}$ est toujours inférieure à celle de $\widehat{S}^{\mathbf{u}}$, avec cas d'égalité ssi $S^{\mathbf{u}} = 0$ ou $S^{\mathbf{u}} = 1$.*

Dans ce contexte, le meilleur estimateur sera sans doute celui conduisant à l'intervalle de confiance asymptotique le moins étendu, *i.e.* celui ayant la plus petite variance asymptotique. Nous justifions l'introduction de l'estimateur $\widehat{T}^{\mathbf{u}}$ par le fait que sa variance asymptotique est toujours inférieure ou égale à celle de $\widehat{S}^{\mathbf{u}}$, les seuls cas d'égalité étant les cas dégénérés où $S^{\mathbf{u}} = 0$ ou $S^{\mathbf{u}} = 1$. Dans ce qui suit, nous montrons l'efficacité asymptotique de $\widehat{T}^{\mathbf{u}}$. Cette notion est une propriété naturelle qui généralise celle d'estimateur sans biais de variance minimale de borne de Cramér-Rao au contexte semi-paramétrique et permet de définir un critère d'optimalité pour les estimateurs. Voir [249, Chapters 8 et 25] et [125] pour plus de détails. Nous montrons en fait plus généralement que la variance asymptotique de $\widehat{T}^{\mathbf{u}}$ est inférieure ou égale à la variance de tout estimateur régulier de $S^{\mathbf{u}}$ construit à partir des observations de $(Y, Y^{\mathbf{u}})$.
Soit $\mathcal{P}$ l'ensemble des FR des vecteurs aléatoires échangeables dans $L^2(\mathbb{R}^2)$. Il est clair que la FR Q d'un vecteur aléatoire dans $L^2(\mathbb{R}^2)$ est dans $\mathcal{P}$ si, et seulement si, Q est symétrique :

$$Q(a,b) = Q(b,a) \quad \forall (a,b) \in \mathbb{R}^2.$$

Proposition 4.4 (Efficacité asymptotique de $\widehat{T}^{\mathbf{u}}$)**.** *Soit maintenant P la FR de $(Y, Y^{\mathbf{u}})$. La suite d'estimateurs $(\widehat{T}^{\mathbf{u}})_N$ est asymptotiquenemt efficace pour estimer $S^{\mathbf{u}}$ pour $P \in \mathcal{P}$ parmi les suites d'estimateurs réguliers s'écrivant comme des fonctions échangeables de la paire $(Y, Y^{\mathbf{u}})$.*

Nous avons bien que $P \in \mathcal{P}$ puisque les variables Y et $Y^{\mathbf{u}}$ sont échangeables. Pour prouver la proposition précédente, le lemme suivant est nécessaire. Nous le mentionnons pour son intérêt en tant que tel.

Lemme 4.5 (Efficacité asymptotique dans $\mathcal{P}$)**.**

(i) Soit $\Phi_1 : \mathbb{R} \to \mathbb{R}$ une fonction de $L^2(P)$. La suite d'estimateurs $\left(\Phi_N^1\right)_N$ donnée par :

$$\Phi_N^1 = \frac{1}{N} \sum \frac{\Phi_1(Y_i) + \Phi_1(Y_i^{\mathbf{u}})}{2}$$

est asymptotiquenemt efficace pour estimer $\mathbb{E}[\Phi_1(Y)]$ pour $P \in \mathcal{P}$.

(ii) Soit $\Phi_2 : \mathbb{R}^2 \to \mathbb{R}$ une fonction symétrique de $L^2(P)$. La suite d'estimateurs $\left(\Phi_N^2\right)_N$ donnée par :

$$\Phi_N^2 = \frac{1}{N} \sum \Phi_2\left(Y_i, Y_i^X\right)$$

est asymptotiquenemt efficace pour estimer $\mathbb{E}[\Phi_2(Y, Y^X)]$ pour $P \in \mathcal{P}$.

4.2.4 Propriétés non asymptotiques : inégalités de concentration

En pratique, comme expliqué précédemment, le nombre d'appels au code est limité . La taille N du plan d'expérience peut parfois être petite et les distributions asymptotiques inatteignables. Dans la pratique pour garantir la validité et la précision de nos estimateurs, il convient alors d'avoir des résultats non asymptotiques (inégalités de concentration, borne de Berry-Esseen,...). Nous proposons dans cette section des inégalités de concentration pour les indices de Sobol pour une sortie scalaire. Classiquement, ces bornes exponentielles peuvent ensuite être utilisées pour construire des IC non asymptotiques avec une probabilité donnée.

Théorème 4.6 (Inégalités de concentration pour $\widehat{S}^{\mathbf{u}}$)**.** *Soient $b > 0$ et $t > 0$. Supposons que la sortie du code est dans $[-b, b]$. Introduisons les variables aléatoires suivantes :*

$$U_j^{\pm} = Y_j Y_j^{\mathbf{u}} - (S^{\mathbf{u}} \pm y)(Y_j)^2 \text{ et } J_j^{\pm} = (S^{\mathbf{u}} \pm y) Y_j - Y_j^{\mathbf{u}}$$

et notons V_U^+ (respectivement V_U^-, V_J^+ et V_J^-) le moment d'ordre 2 des v.a. i.i.d. U_j^+ (resp. U_j^-, J_j^+ et J_j^-). Alors

$$\mathbb{P}\left(\widehat{S}^{\mathbf{u}} - S^{\mathbf{u}} \geqslant t\right) \leqslant M_1 + 2M_2 + 2M_3, \tag{4.12}$$

$$\mathbb{P}\left(\widehat{S}^{\mathbf{u}} - S^{\mathbf{u}} \leqslant -t\right) \leqslant M_4 + 2M_2 + 2M_5, \tag{4.13}$$

où

$$M_1 = \exp\left\{-\frac{NV_U^+}{b_U^2} h\left(\frac{b_U}{V_U^+}\frac{tV}{2}\right)\right\} \qquad M_3 = \exp\left\{-\frac{NV_J^+ b^2}{b_U^2} h\left(\frac{b_U}{bV_J^+}\sqrt{\frac{tV}{2}}\right)\right\}$$

$$M_2 = \exp\left\{-\frac{NV}{b^2} h\left(\frac{b}{V}\sqrt{\frac{tV}{2}}\right)\right\} \qquad M_4 = \exp\left\{-\frac{NV_U^-}{b_U^2} h\left(\frac{b_U}{V_U^-}\frac{tV}{2}\right)\right\}$$

$$M_5 = \exp\left\{-\frac{NV_J^- b^2}{b_U^2} h\left(\frac{b_U}{bV_J^-}\sqrt{\frac{tV}{2}}\right)\right\},$$

$b_U = b^2(1 + S_{\mathrm{Cl}}^{\mathbf{u}} + t)$ et la fonction h est définie par $h(x) = (1+x)\ln(1+x) - x$ pour tout $x > -1$.

Nous pouvons aussi obtenir des inégalités de concentration pour $\widehat{S}^{\mathbf{u}}$ en appliquant le Corollaire 1.17 de [161]. De façon à être complet, nous rappelons ce corollaire.

Corollaire 4.7 (Corollaire 1.17 dans [161])**.** *Soit $P = \mu_1 \otimes \ldots \otimes \mu_n$ une mesure de probabilté produit définie sur le produit cartésien $X = X_1 \times \ldots \times X_n$, où (X_i, d_i) est un espace métrique de diamètre fini D_i, $i = 1, \ldots, n$, muni de la métrique l^1 : $d = \sum_{i=1}^n d_i$. Soit F une fonction 1-Lipschitz sur (X, d). Alors,*

pour tout $r \geqslant 0$,

$$\mathbb{P}\left(F \geqslant \mathbb{E}_P(F) + r\right) \leqslant e^{-r^2/2D^2}$$

où $D^2 = \sum_{i=1}^n D_i^2$.

Cela conduit au résultat suivant : si Y est borné p.s., alors pour tout $t \geqslant 0$, nous avons

$$\mathbb{P}\left(\widehat{S}^{\mathbf{u}} - S^{\mathbf{u}} \geqslant t\right) \leqslant \exp\left\{-\frac{N}{2}\left(\frac{\left(1-\frac{1}{N}\right)tV}{8(1+2(S^{\mathbf{u}}_{\mathrm{Cl}}+t))}\right)^2\right\},$$

et, pour tout $t \geqslant 0$, nous avons

$$\mathbb{P}\left(\widehat{S}^{\mathbf{u}} - S^{\mathbf{u}} \leqslant -t\right) \leqslant \exp\left\{-\frac{N}{2}\left(\frac{\left(1-\frac{1}{N}\right)tV}{8(1+2(S^{\mathbf{u}}_{\mathrm{Cl}}+t))}\right)^2\right\}.$$

De façon à s'affranchir de la quantité inconnue $S^{\mathbf{u}}$ des bornes précédentes, on majore brutalement $S^{\mathbf{u}}$ par 1 pour obtenir :

$$\forall t \geqslant 0,\ \ \mathbb{P}\left(\widehat{S}^{\mathbf{u}} - S^{\mathbf{u}} \geqslant t\right) \leqslant \exp\left\{-\frac{NV^2}{128}\left(1-\frac{1}{N}\right)^2\left(\frac{t}{3+2t}\right)^2\right\},$$
$$\forall t \geqslant 0,\ \ \mathbb{P}\left(\widehat{S}^{\mathbf{u}} - S^{\mathbf{u}} \leqslant -t\right) \leqslant \exp\left\{-\frac{NV^2}{128}\left(1-\frac{1}{N}\right)^2\left(\frac{t}{3+2t}\right)^2\right\}.$$

En procédant de même, nous obtenons des inégalités de concentration pour $\widehat{T}^{\mathbf{u}}$. Cf. [J10] pour les résultats complets.

Les bornes présentées dans le Théorème 4.6 sont plus fines que celles obtenues en utilisant le Corollaire 1.17 de [161] (Corollaire 4.7). Ceci peut être expliqué de la manière suivante : le résultat de [161] est très général et valable pour n'importe quelle fonctionnelle Lipschitz. Par conséquent, il tient compte du pire cas possible. Dans notre approche, nous utilisons la forme particulière de notre estimateur, ce qui nous permet de fournir des bornes plus précises.

4.2.5 Propriétés non asymptotiques : résultats de type Berry-Esseen

Une autre manière de quantifier les bonnes propriétés des estimateurs à N fixé est d'établir des bornes de type Berry-Esseen. Cette section est consacré à ce genre de résultats pour $\widehat{S}^{\mathbf{u}}$.

Le théorème de Pinelis

Nous commençons par rappeler un résultat général de type Berry-Esseen établi dans [202]. Soit $(V_i)_{i\geqslant 1}$ une suite de v.a. centrées i.i.d. dans $\mathbb{R}^p$, avec $p \in \mathbb{N}^*$. Soit f une fonction mesurable réelle définie sur $\mathbb{R}^p$ s'annulant en zéro et telle que :

$$\exists\, \varepsilon > 0,\ \exists\, M_\varepsilon > 0 \text{ t.q. } |f(x) - L(x)| \leqslant \frac{M_\varepsilon}{2}||x||^2 \tag{4.14}$$

où $L = Df(0)$ est la dérivée de Fréchet de f au point 0. La condition (4.14) est satisfaite dès que f est deux fois continuement différentiable au voisinage de 0.

Théorème 4.8 (Corollaire 3.7 dans [202])**.** *Soit* $q \in]2,3]$. *Supposons que* (4.14) *est satisfaite,*

$$\sigma = \sqrt{\mathbb{E}\left[L(V)^2\right]} > 0\,,$$

et $\mathbb{E}\left[\|V\|^q\right]^{1/q} < \infty$ *où* $\|\cdot\|$ *représente la norme euclidienne sur* $\mathbb{R}^p$*. Alors pour tout* $t \in \mathbb{R}$

$$\left|\mathbb{P}\left(\frac{f(\overline{V}_n)}{\sigma/\sqrt{n}} \leqslant t\right) - \Phi(t)\right| \leqslant \frac{\kappa}{n^{p/2-1}}\,, \tag{4.15}$$

où κ *est une constante générique qui ne dépend que de* q *et* Φ *est la FR de la gaussienne standard.*

Un résultat théorique dans le cas général

Pour toute variable aléatoire Z, notons Z^c sa version centrée $Z^c = Z - \mathbb{E}[Z]$.

Théorème 4.9 (Borne de Berry-Esseen pour $\widehat{S}^{\mathbf{u}}$)**.** *Supposons que* Y *a un moment fini d'ordre* 6*. Alors, pour tout* $t \in \mathbb{R}$*,*

$$\left|\mathbb{P}\left(\frac{\sqrt{N}}{\sigma}\left(\widehat{S}^{\mathbf{u}} - S^{\mathbf{u}}\right) \leqslant t\right) - \Phi(t)\right| \leqslant \frac{\kappa}{\sqrt{N}}\,. \tag{4.16}$$

Ici

$$\sigma^2 = \mathrm{Var}\left(\frac{1}{V}\left(Y^c(Y^{\mathbf{u}})^c - S^{\mathbf{u}}(Y^c)^2\right)\right)$$

est la variance asymptotique de $\sqrt{N}\widehat{S}^{\mathbf{u}}$*.*

Nous avons ainsi établi un résultat de type Berry-Esseen pour $\widehat{S}^{\mathbf{u}}$ quelle que soit la valeur de $\mathbb{E}[Y]$. Cependant, la constante dans la borne est difficile voire impossible à exprimer explicitement. Dans le paragraphe suivant, nous établissons un autre résultat de type Berry-Essen avec des bornes explicites dans le cas centré mais pour un estimateur de $S^{\mathbf{u}}$ légèrement différent de $\widehat{S}^{\mathbf{u}}$.

Résultat pratique dans le cas centré

Nous nous intéressons à l'estimateur suivant

$$\widehat{S}_c^{\mathbf{u}} = \frac{\frac{1}{N}\sum_{j=1}^N Y_j Y_j^{\mathbf{u}}}{\frac{1}{N}\sum_{j=1}^N Y_j^2}$$

dans le cas où $\mathbb{E}[Y] = 0$. Soit $\kappa \approx 0.42$ la meilleure constante connue dans le théorème de Berry-Esseen classique [150].

Théorème 4.10 (Borne de Berry-Esseen pour $\widehat{S}_c^{\mathbf{u}}$)**.** *Supposons que* Y *a un moment fini d'ordre* 6*. Alors, pour tout* $t \in \mathbb{R}$*,*

$$\left|\mathbb{P}\left(\frac{\sqrt{N}}{\sigma}(\widetilde{S}_c^{\mathbf{u}} - S^{\mathbf{u}}) \leqslant t\right) - \Phi(t)\right| \leqslant \frac{\kappa\mu_{3,N}}{\sqrt{N}} + \left|\Phi(t) - \Phi\left(\frac{t}{\sqrt{1 + \frac{t\nu_N}{\sigma\sqrt{N}V^2}}}\right)\right|. \tag{4.17}$$

Ici

$$\sigma^2 = \mathrm{Var}\left(\frac{1}{V}\left(YY^{\mathbf{u}} - S^{\mathbf{u}}Y^2\right)\right) \tag{4.18}$$

est la variance asymptotique de $\sqrt{N}\widetilde{S}^{\mathbf{u}}$ *et*

$$\begin{aligned}
\mu_{3,N} &= \mathbb{E}\left[\left|\frac{\Delta_n - \mathbb{E}[\Delta_n]}{\sqrt{\mathrm{Var}\Delta_n}}\right|^3\right], \\
\Delta_N &= \sigma^{-1}V\left[YY^{\mathbf{u}} - \left(S^{\mathbf{u}} + \frac{t\sigma}{\sqrt{N}}\right)Y^2\right], \\
\nu_N &= \left(\frac{t\sigma}{\sqrt{N}} + 2S^{\mathbf{u}}\right)\mathrm{Var}(Y^2) - 2\mathrm{Cov}(YY^{\mathbf{u}}, Y^2).
\end{aligned}$$

4.2.6 Estimation jointe des indices de Sobol

En général, afin de classer les variables d'entrée en fonction de leur importance, les utilisateurs vont estimer conjointement tous les indices de premier ordre (ainsi que les indices de Sobol totaux). Étant donné que ces différents estimateurs sont dépendants, les distributions marginales asymptotiques ne sont pas complètement informatives et il est utile la loi jointe de ces estimateurs.
En mettant à profit la Delta méthode [249], il est aisé de déduire du Théorème 4.2 un TCL vectoriel des estimateurs $\widehat{S}^{\mathbf{U}}$ et $\widehat{T}^{\mathbf{U}}$ (voir (4) et (5) dans [J10] pour leur définition précise) du vecteur des indices de Sobol défini par

$$S^{\mathbf{U}} := \left(\frac{\mathrm{Var}(\mathbb{E}[Y|X_{\mathbf{u}_1}])}{\mathrm{Var}(Y)}, \ldots, \frac{\mathrm{Var}(\mathbb{E}[Y|X_{\mathbf{u}_k}])}{\mathrm{Var}(Y)} \right), \tag{4.19}$$

où $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_k)$ est constitué de k sous-ensembles de I_p. Voir Théorème 3.1 dans [J10].

Quelques cas particuliers

1) Supposons que $k = p$, $\mathbf{U} = (\{1\}, \ldots, \{p\})$ et $\mathbb{E}[Y^4] < \infty$. Nous obtenons alors la normalité asymptotiques du vecteur de tous les indices de premier ordre.
2) Nous pouvons aussi obtenir directement un TCL pour tout indice d'ordre 2 $S^{i,i'}$ pour $(i, i') \in I_p^2$ avec $i \neq i'$. En effet, il suffit de prendre $k = 1$ et $\mathbf{u} = \{i, i'\}$.
3) Pour $\mathbf{v} \subset I_p$, il est possible d'établir un TCL vectoriel pour $\left(\widehat{S}^{\mathbf{u}}, \widehat{S}^{\mathbf{v}}, \widehat{S}^{\mathbf{u}\cup\mathbf{v}}\right)$ duquel on déduit un TCL pour $\widehat{S}^{\mathbf{u}\cup\mathbf{v}} - \widehat{S}^{\mathbf{u}} - \widehat{S}^{\mathbf{v}}$, estimateur naturel de $S^{\mathbf{u}\cup\mathbf{v}} - S^{\mathbf{u}} - S^{\mathbf{v}}$, qui quantifie l'influence (pour $u \cap v = \emptyset$) de l'intéraction entre les variables $X_{\mathbf{u}}$ et $X_{\mathbf{v}}$ sur la sortie Y.

Cette loi jointe permet, par exemple, d'effectuer des tests statistiques et des comparaisons entre différents indices, afin de classer rigoureusement les variables d'entrée en tenant compte des erreurs d'estimation des indices.

Application : test statistique sur un modèle réel
En aéronautique, la masse de fuel requise pour faire la liaison entre deux villes avec un avion commercial est généralement modélisée par la formule de Bréguet :

$$M_{fuel} = (M_{empty} + M_{pload}) \left(e^{\frac{SFC \cdot g \cdot Ra}{V \cdot F}\, 10^{-3}} - 1 \right). \tag{4.20}$$

Voir [209] pour la description du modèle et plus de détails. Les variables déterministes et fixées sont :

— M_{empty} : poids basique de l'avion (hors fuel et passagers) ;
— M_{pload} : capacité maximale de l'avion ;
— g : constante gravitationnelle ;
— Ra : distance parcourue par l'avion ;

tandis que les variables incertaines et leurs distributions sont listées dans la Table 4.1 (voir [209]).

Variable	Signification physique	Densité	Paramètres
V	Vitesse de croisière de l'avion	Uniforme	$(226, 234)$
F	Coefficient aérodynamique	Beta	$(7, 2, 18.7, 19.05)$
SFC	Cste caractéristique des moteurs	$\theta_2\, e^{-\theta_2(u-\theta_1)}\, \mathbb{1}_{[\theta_1,+\infty[}$	$\theta_1 = 17.23, \theta_2 = 3.45$

TABLE 4.1 – Modèle de consommation de fuel (4.20) en aéronautique. Modélisation alétoire des entrées

L'incertitude sur la vitesse de croisière V permet un écart absolu de 4 minutes sur l'heure prévue d'arrivée. Les constructeurs aéronautiques peuvent se demander s'ils doivent plutôt améliorer la performance SFC du moteur ou les propriétés aérodynamiques F de l'avion de façon à réduire la quantité de fuel nécessaire. Pour ce faire, nous menons donc une analyse de sensibilté de M_{fuel}par rapport à F et SFC et nous voulons tester $H_0 :\ S^{SFC} > S^F$ contre $H_1 :\ S^{SFC} \leqslant S^F$. En appliquant la procédure de test détaillée dans [J10], nous ne pouvons rejeter H_0.

4.3 Analyse de sensibilité sur un métamodèle

Comme expliqué dans l'introduction, en pratique, les utilisateurs se trouvent souvent dans une situation où la sortie exacte f est trop coûteuse pour être évaluée numériquement. Dans ce cas, Y et X ne sont pas des variables observables dans notre problème d'estimation et le code doit être remplacé par un métamodèle $\widetilde{f}$, plus rapide à évaluer et constituant une bonne approximation de f. Dans cette section, nous considérons cette approximation comme une perturbation du modèle exact par une fonction δ :

$$\widetilde{Y} = \widetilde{f}(X) = f(X) + \delta,$$

où la perturbation $\delta = \delta(X, \xi)$ est une fonction dépendant à la fois de l'entrée du code X et d'une autre variable aléatoire ξ independante de$X$. Les résultats de cette section se trouvent dans [J5].

4.3.1 Définition de l'indice de Sobol et de ses estimateurs

Supposons encore que $\widetilde{Y}$ n'est pas déterministe et admet un moment d'ordre 2. Nous pouvons alors considérer l'indice de Sobol par rapport à $X_{\mathbf{u}}$ et relatif au métamodèle :

$$S^{\mathbf{u}}_{\text{Méta}} = \frac{\text{Var}(\mathbb{E}[\widetilde{Y}|X_{\mathbf{u}}])}{\text{Var}(\widetilde{Y})}. \tag{4.21}$$

Comme précédemment, nous définissons

$$\widetilde{Y}^{\mathbf{u}} = \widetilde{f}(X^{\mathbf{u}})$$

où $\mathbf{u}$ est un sous ensemble de I_p. Les estimateurs de $S^{\mathbf{u}}_{\text{Méta}}$ sont alors donnés par

$$\widehat{S}^{\mathbf{u}}_{\text{Méta}} = \frac{\frac{1}{N}\sum \widetilde{Y}_j\widetilde{Y}^{\mathbf{u}}_j - \left(\frac{1}{N}\sum \widetilde{Y}_j\right)\left(\frac{1}{N}\sum \widetilde{Y}^{\mathbf{u}}_j\right)}{\frac{1}{N}\sum \widetilde{Y}_j^2 - \left(\frac{1}{N}\sum \widetilde{Y}_j\right)^2} \quad \text{et} \quad \widehat{T}^{\mathbf{u}}_{\text{Méta}} = \frac{\frac{1}{N}\sum \widetilde{Y}_j\widetilde{Y}^{\mathbf{u}}_j - \left(\frac{1}{N}\sum \left[\frac{\widetilde{Y}_j+\widetilde{Y}^{\mathbf{u}}_j}{2}\right]\right)^2}{\frac{1}{N}\sum \left[\frac{\widetilde{Y}_j^2+(\widetilde{Y}^{\mathbf{u}}_j)^2}{2}\right] - \left(\frac{1}{N}\sum \left[\frac{\widetilde{Y}_j+\widetilde{Y}^{\mathbf{u}}_j}{2}\right]\right)^2}. \tag{4.22}$$

Le but de cette section est de donner des conditions suffisantes sur la perturbation δ pour que $\widehat{S}^{\mathbf{u}}_{\text{Méta}}$ et $\widehat{T}^{\mathbf{u}}_{\text{Méta}}$ soient asymptotiquement gaussiens et que $\widehat{T}^{\mathbf{u}}_{\text{Méta}}$ soit asymptotiquement efficace dans l'estimation de l'indice de Sobol $S^{\mathbf{u}}$ du modèle exact.

4.3.2 Propriétés asymptotiques

Premier cas : δ ne dépend pas de N

Si $S^{\mathbf{u}}_{\text{Méta}} - S^{\mathbf{u}} \neq 0$, alors ni $\widehat{S}^{\mathbf{u}}_{\text{Méta}}$ ni $\widehat{T}^{\mathbf{u}}_{\text{Méta}}$ ne sont consistants pour estimer $S^{\mathbf{u}}$. En effet, nous avons la décomposition suivante

$$\widetilde{S}^{\mathbf{u}} - S^{\mathbf{u}} = (\widehat{S}^{\mathbf{u}}_{\text{Méta}} - S^{\mathbf{u}}_{\text{Méta}}) + (S^{\mathbf{u}}_{\text{Méta}} - S^{\mathbf{u}}).$$

Le premier terme converges p.s. vers 0. d'après le résultat de consistance des estimateurs des indices de Sobol appliqué à $S^{\mathbf{u}}_{\text{Méta}}$ (cf. Théorème 4.2). Cependant, le second terme est non nul par hypothèse. De même pour $\widehat{T}^{\mathbf{u}}_{\text{Méta}}$.

Cette remarque montre qu'une considération naïve de l'erreur du métamodèle (*i.e.* avec métamodèle fixe) n'est pas satisfaisante pour une justification asymptotique de l'utilisation d'un métamodèle. Plus précisément, il est impossible d'avoir la normalité asymptotique pour $\widehat{S}^{\mathbf{u}}_{\text{Méta}}$ et $\widehat{T}^{\mathbf{u}}_{\text{Méta}}$ dans n'importe quel cas non trivial si δ ne s'annule pas asymptotiquement. Cela justifie l'examen des cas où δ dépend de N, et c'est l'objet de ce qui suit.

Second cas : $\mathrm{Var}(\delta_N)$ converge vers 0 lorsque $N \to \infty$

Supposons maintenant que la perturbation δ est une fonction de la taille N de l'échantillon. Ceci entraîne que $\widetilde{f}, \widetilde{Y}, \widetilde{Y}^{\mathbf{u}}$ et $S^{\mathbf{u}}_{\text{Méta}}$ dépendent de N. Nous utiliserons donc plutôt les notations suivantes δ_N, $\widetilde{f}_N$, $\widetilde{Y}_N$ et $\widetilde{Y}^{\mathbf{u}}_N$ pour insister sur cette dépendance en N. Cependant, pour les estimateurs de $S^{\mathbf{u}}_{\text{Méta}}$, nous conservons les notations $\widehat{S}^{\mathbf{u}}_{\text{Méta}}$ et $\widehat{T}^{\mathbf{u}}_{\text{Méta}}$ définis par (4.22). Supposons que $\widetilde{f}_N - f = \delta_N \xrightarrow[N\to+\infty]{L^2} c$ pour une constante c donnée.

Proposition 4.11. *Supposons de plus qu'il existe $s > 0$ et $C > 0$ tels que*

$$\forall N, \ \mathbb{E}\left[\left|\widetilde{Y}_N\right|^{4+s}\right] < C. \tag{4.23}$$

Alors

$$\sqrt{N}\left(\widehat{S}^{\mathbf{u}}_{Méta} - S^{\mathbf{u}}_{Méta}\right) \xrightarrow[N\to\infty]{\mathcal{L}} \mathcal{N}_1\left(0, \sigma_S^2\right) \quad et \quad \sqrt{N}\left(\widetilde{T}^{\mathbf{u}} - S^{\mathbf{u}}_{Méta}\right) \xrightarrow[N\to\infty]{\mathcal{L}} \mathcal{N}_1\left(0, \sigma_T^2\right) \tag{4.24}$$

où σ_S^2 et σ_T^2 sont les variances asymptotiques de $\widehat{S}^{\mathbf{u}}$ et $\widehat{T}^{\mathbf{u}}$ données par (4.10) *et* (4.11).

En réalité, nous sommes intéressés par la distribution asymptotique de $\sqrt{N}\left(\widetilde{S}^{\mathbf{u}} - S^{\mathbf{u}}\right)$. Dans le reste de cette section, nous montrons que la convergence dépend de la vitesse de convergence vers 0 de $\mathrm{Var}(\delta_N)$.

Théorème 4.12 (Consistance et normalité asymptotique pour le métamodèle)**.** *Les estimateurs sont consistants : $\widehat{S}^{\mathbf{u}}_{Méta} \xrightarrow[N\to\infty]{p.s.} S^{\mathbf{u}}$ et $\widehat{T}^{\mathbf{u}}_{Méta} \xrightarrow[N\to\infty]{p.s.} S^{\mathbf{u}}$. De plus, posons*

$$C_{\delta_N,N,\boldsymbol{u}} = 2\mathrm{Var}(Y)^{1/2}\left[\mathrm{Cor}(Y,\delta_N^{\mathbf{u}}) - \mathrm{Cor}(Y,Y^{\mathbf{u}})\mathrm{Cor}(Y,\delta_N)\right] + \mathrm{Var}(\delta_N)^{1/2}\left[\mathrm{Cor}(\delta_N,\delta_N^{\mathbf{u}}) - \mathrm{Cor}(Y,Y^{\mathbf{u}})\right],$$

où $\delta_N^{\mathbf{u}} = \delta_N(X^{\mathbf{u}})$, et pour toutes v.a. A et B dans L^2 de variance non nulle, rappelons que la corrélation entre A et B est donnée par :

$$\mathrm{Cor}(A,B) = \frac{\mathrm{Cov}(A,B)}{\sqrt{\mathrm{Var}(A),\mathrm{Var}(B)}}.$$

Supossons que $C_{\delta_N,N,\boldsymbol{u}}$ ne tend pas vers 0.

1) *Si $\mathrm{Var}(\delta_N) = o\left(\frac{1}{N}\right)$, alors la normalité asymptotique de $\widetilde{S}^{\mathbf{u}}$ et celle de $\widetilde{T}^{\mathbf{u}}$ pour l'estimation de $S^{\mathbf{u}}$ sont vérifiées, i.e.*

$$\sqrt{N}(\widehat{S}^{\mathbf{u}}_{Méta} - S^{\mathbf{u}}) \xrightarrow[N\to\infty]{\mathcal{L}} \mathcal{N}_1(0,\sigma_S^2) \quad et \quad \sqrt{N}(\widetilde{T}^{\mathbf{u}} - S^{\mathbf{u}}) \xrightarrow[N\to\infty]{\mathcal{L}} \mathcal{N}_1(0,\sigma_T^2). \tag{4.25}$$

2) *Si $N\mathrm{Var}(\delta_N) \to \infty$, alors* (4.25) *est vérifiée.*

3) *Si $C_{\delta_N,N,\boldsymbol{u}}$ converge vers une constante C non nulle et que $\mathrm{Var}(\delta_N) = \frac{\gamma}{CN} + o\left(\frac{1}{N}\right)$ avec $\gamma \in \mathbb{R}$, alors*

$$\sqrt{N}(\widetilde{S}^{\mathbf{u}} - S^{\mathbf{u}}) \xrightarrow[N\to\infty]{\mathcal{L}} \mathcal{N}_1(\gamma,\sigma_S^2) \quad et \quad \sqrt{N}(\widehat{T}^{\mathbf{u}}_{Méta} - S^{\mathbf{u}}) \xrightarrow[N\to\infty]{\mathcal{L}} \mathcal{N}_1(\gamma,\sigma_T^2).$$

Bien sûr si $C_{\delta_N,N,\mathbf{u}}$ converge vers 0, alors la normalité asymptotique de $\widehat{S}^{\mathbf{u}}_{\text{Méta}}$ et $\widetilde{T}^{\mathbf{u}}$ est satisfaite sous des conditions plus faibles sur $\text{Var}(\delta_N)$.

Proposition 4.13 (Efficacité asymptotique pour le métamodèle)**.** *Supposons*

(i) $\exists s > 0, C > 0$ *t.q.* $\forall N$, $\mathbb{E}\left[\left|Y\right|^{4+s}\right] < C$ *et* $\mathbb{E}\left[\left|\widetilde{Y}\right|^{4+s}\right] < C$ *;*

(ii) $N\text{Var}(\delta_N) \to 0$ *;*

(iii) $\sqrt{N}\mathbb{E}[\delta_N] \to 0$.

Alors $\left(\widehat{T}^{\mathbf{u}}_{Méta}\right)_N$ *est asymptotiquement efficace pour estimer* $S^{\mathbf{u}}$.

D'après l'inégalité de Minkowski, la première hypothèse implique que $\mathbb{E}[\delta_N^{4+s}] < 2C^{\frac{1}{4+s}}$ et la normalité asymptotique d'après la Proposition 4.11 et le Théorème 4.12.

4.3.3 Applications numériques

Nous étudions la fonction Ishigami [127] définie par :

$$f(X_1, X_2, X_3) = \sin X_1 + 7\sin^2 X_2 + 0.1X_3^4 \sin X_1 \tag{4.26}$$

où $(X_i)_{i=1,2,3}$ sont des v.a. i.i.d. de loi uniforme sur $[-\pi, \pi]$. Dans le cas où le modèle exact n'est pas accessible mais que nous disposons uniquement d'un métamodèle $\widetilde{f}_N = f + \delta_N$, un intervalle de confiance peut malgré tout être estimé en utilisant les estimateurs $\widehat{S}^{\mathbf{u}}_{\text{Méta}}$ ou $\widehat{T}^{\mathbf{u}}_{\text{Méta}}$. D'après la Proposition 4.11, le niveau de confiance de l'intervalle de confiance obtenu devrait être proche de $1-\alpha$ pour une grande taille N d'échantillon si, et seulement si, $\text{Var}(\delta_N)$ décroît suffisamment rapidement en N. Dans les paragraphes suivants, nous travaillons avec l'estimateur $\widehat{S}^{\mathbf{u}}_{\text{Méta}}$ et présentons les estimations des niveaux de confiance des IC pour le modèle d'Ishigami (4.26) en considérant une perturbation du vrai modèle ainsi qu'un métamodèle de krigeage par RKHS. Nous avons aussi étudié un métamodèle de régression non paramétrique dans [J6] que nous ne présentons pas dans ce manuscrit.

Modèle perturbé par une variable de loi de Weibull

Nous considérons ici une autre perturbation du modèle initial :

$$\widetilde{f}_N(X) = f(X) + \frac{5WX_3^2}{N^{\beta/2}} \tag{4.27}$$

où W est une variable de loi de Weibull de paramètre d'échelle $\lambda = 1$ et de paramètre de forme $k = 1/2$. Ici, la perturbation dépend des entrées et, puisque pour toute variable d'entrée, $C_{\delta_N,N,\mathbf{u}}$ ne converge par vers zéro, le Théorème 4.12 établit en particulier que $\widehat{S}^{\mathbf{u}}_{\text{Méta}}$ est asymptotiquement gaussien pour $S^{\mathbf{u}}$ pour $\beta > 1$. Ce résultat est illustré pour $N = 50000$ dans la Figure 4.1. Nous voyons que les couvertures empiriques des IC de $S^{\mathbf{1}}$ et $S^{\mathbf{2}}$ "sautent" vers 0.95 près de $\beta = 1$, tandis que, pour $S^{\mathbf{3}}$, cette couverture est plus vite proche de 0.95.

Métamodèle RKHS

Dans ce paragraghe, nous prenons comme métamodèle $\widetilde{f}$ une interpolation RKHS (Reproducing Kernel Hilbert Space) [225, 227, 228]. De tels métamodèles, aussi connus sous le nom de métamodèles de krigeage ou processus gaussiens, ont largement été utilisés depuis quelques années pour réaliser l'analyse de sensibilité de codes de calculs boîtes noires coûteux [176]. Dans certains cas (par exemple, pour des entrées uniformes ou gaussiennes), il est possible d'établir une expression analytique des indices de Sobol et ainsi ne pas avoir recours à des schémas d'estimation de type Monte-Carlo [62]. Dans ce qui

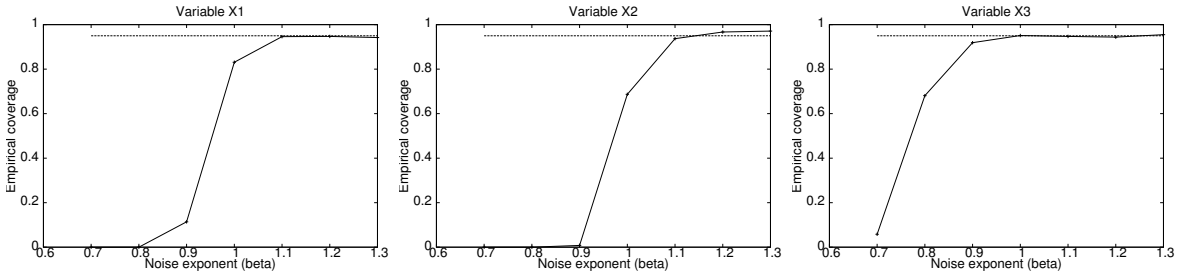


FIGURE 4.1 – Modèle d'Ishigami donné par (4.26) perturbé par une Weibull. Couvertures empiriques des IC asymptotiques de $S^{\mathbf{1}}$, $S^{\mathbf{2}}$ et $S^{\mathbf{3}}$, en fonction de β. La taille N de l'échantillon est fixée à 50000, le nombre R de réplications est de 1000 et le niveau de confiance est donné par $1-\alpha = 0.95$.

suit, nous choisissons de procéder à une estimation Monte-Carlo sur le métamodèle RKHS de façon à illustrer les résultats théoriques obtenus précédemment. En outre, l'approche Monte-Carlo est plus flexible et applicable pour des lois complexes des entrées. L'interpolateur dépend du plan d'échantillonnage $((d_1, f(d_1)), \ldots, (d_n, f(d_n)))$, où les points du plan d'échantillonnage initial $\mathcal{D} = (d_i)_{i=1,\ldots,n} \subset E$ sont généralement choisis selon une procédure de remplissage (space-filling), par exemple selon un échantillonage hypercube latin (LHS) [135] ou des tableaux orthogonaux (OA) [193]. Lorsque la taille n du plan d'échantillonnage initial augmente, le nombre d'évaluations du vrai modèle f augmente aussi (calcul des $(f(d_i))_{i=1,\ldots,n}$) mais conduira aussi à une meilleure qualité de l'interpolation et donc à une erreur de métamodèle plus faible.
On peut montrer (cf. [227, 173] et preuve dans [J5]) que

$$\mathrm{Var}(\delta) \leqslant Ce^{-kn^{1/p}}$$

pour des constantes C et k convenables.

Application numérique. Le plan d'échantillonnage initial a été construit par LHS. L'interpolation RKHS dépend également du choix d'un noyau, que nous choisissons gaussien. Toutes les simulations ont été réalisées avec le package `lhs` du logiciel `R` [208] pour l'échantillonnage du plan d'échantillonnage et le package `mlegp` pour le krigeage [102].
En utilisant une régression exponentielle (justifiée numériquement par la Figure 6 dans [J5]), nous obtenons que

$$\mathrm{Var}(\delta_N) \approx \widehat{C}e^{-\widehat{k}n^{1/3}} \tag{4.28}$$

où $\widehat{k} = 1.91$. Maintenant, si nous laissons la taille n de l'échantillon du plan d'échantillonnage initial dépendre de la taille N de l'échantillon Monte-Carlo intervenant dans l'estimation des indices par la relation : $n = (a \ln N)^3$ pour $a > 0$, le Théorème 4.12 suggère que les estimateurs des indices de Sobol pour le métamodèle sont asymptotiquement gaussiens si, et seulement si, $N^{-a\widehat{k}+1} \to 0$ lorsque $N \to +\infty$, *i.e.*,

$$a > \frac{1}{\widehat{k}} = 0.52, \tag{4.29}$$

puisque $\widehat{k} = 1.91$. Même s'il n'a pas été rigoureusement prouvé que cette condition est nécessaire et suffisante (en raison de l'estimation de k et du fait que (4.28) constitue sans doute une limite supérieure, éventuellement avec des constantes différentes), on devrait observer en pratique que le comportement des IC empiriques pour de grandes valeurs de N change une fois que cette valeur critique de a est franchie. La Table 4.2 donne les résultats obtenus pour différentes valeurs de a, certaines sous critiques pour lesquelles (4.29) n'est pas vérifiée et d'autres sur critiques pour lesquelles (4.29) est vérifiée. Cette table fournit une

a	N	n	Couv. pour S^1	Couv. pour S^2	Couv. pour S^3
0.4	3000	33	0.10	0.00	0.70
0.4	4000	37	0.08	0.00	0.78
0.4	6000	43	0.26	0.30	0.88
0.4	10000	51	0.28	0.18	0.78
0.4	20000	77	0.28	0.10	0.59
0.6	3000	111	0.79	0.37	0.90
0.6	4000	124	0.80	0.70	0.94
0.6	10000	169	0.92	0.82	0.94
0.6	20000	210	0.93	0.85	0.95
0.7	3000	177	0.93	0.88	0.93
0.7	4000	196	0.90	0.91	0.94
0.7	6000	226	0.94	0.93	0.97
0.8	4000	293	0.95	0.95	0.95

TABLE 4.2 – Métamodèle RKHS pour la fonction Ishigami donnée par (4.26). Estimation des couvertures asymptotiques. $R = 1000$ réplications sont considérées. La couverture théorique est de 0.95.

illustration claire de la conjecture précédente.

4.4 Un cas particulier avec entrées fonctionnelles

Dans cette section, nous proposons dans le cadre très particulier de la régression linéaire, un estimateur de l'indice de Sobol basé sur des U-statistiques. En raison de sa spécifité, ses performances sont meilleures que celles de l'estimateur Pick-Freeze comme l'on pouvait s'y attendre. L'étude menée dans cette section a fait l'objet de la publication [J6].

4.4.1 Contexte et notation

Nous considérons un espace de Hilbert séparable $\mathbb{H}$ muni du produit scalaire $\langle , \rangle$ et p processus stochastiques $X_1, \ldots, X_p$, centrés à valeurs dans $\mathbb{H}$. Le modèle boîte noire étudié est le modèle de régression linéaire suivant :

$$Y = \mu + \sum_{i=1}^{p} \langle \beta_i, X_i \rangle + \varepsilon. \tag{4.30}$$

où $\beta_i, 1 \leqslant i \leqslant p$ sont des éléments de $\mathbb{H}$, μ est un réel et ε est un bruit blanc centré indépendant de $X_1, \ldots, X_p$.

Dans le cas particulier où $p = 1$, de nombreuses procédures d'estimation de la fonction β ont été proposées. Par exemple, des estimateurs par splines ont été considérés dans [70], tandis que [50] a introduit des estimateurs par projection seuillés. Des procédures optimales pour de la prédiction point par point dans le modèle de régression linéaire fonctionnel ont été proposées par [48]. Les auteurs de [51] et [113] réalisent une analyse par composantes principales (ACP) du processus d'entrée X : β est alors estimé dans un espace de dimension finie engendré par les m premières fonctions propres de l'opérateur de covariance empirique de X. Nous renvoyons aussi à la revue [52] et aux références qu'elle contient.

Notre approche est basée sur la décomposition de Karhunen-Loève des processus X_i (voir [172, 44, 206, 24, 68]).Étant donnée cette décomposition, nous construisons des estimateurs naturels des indices de Sobol basés sur les U-statistiques dont nous prouvons qu'ils sont asymptotiquement gaussiens et efficaces. Cette approche est aussi celle adoptée par [51] pour prouver un TCL pour des estimateurs de β.

4.4.2 Modèle de régression linéaire simple

Supposons ici que $p = 1$ et considérons donc le modèle suivant :

$$Y = \mu + \langle \beta, X \rangle + \varepsilon. \tag{4.31}$$

Le résultat pour la régression multiple s'obtient aisément en appliquant la procédure qui suit à toutes les variables d'entrée. Supposons de plus que $\mathbb{E}\left[\|X\|^2\right] < \infty$. Ainsi l'opérateur de covariance de X, défini pour tout $f \in \mathbb{H}$ par $\Gamma(f) = \mathbb{E}\left[\langle X, f\rangle X\right]$, est Hilbert-Schmidt et donc diagonalisable via la décomposition de Karhunen-Loève dans une base orthonormée de fonctions propres $(\varphi_l)_{l\geqslant 1}$, dont les valeurs propres $(\lambda_l)_{l\geqslant 1}$ classées par ordre décroissant sont telles que $\sum_{l=1}^{\infty} \lambda_l < +\infty$. Voir, par exemple, [168, 169, 138]. Plus précisément,

$$X = \sum_{l=1}^{\infty} \sqrt{\lambda_l}\xi_l\varphi_l$$

et par conséquent, $\langle X, \varphi_l\rangle = \sqrt{\lambda_l}\xi_l$. Les variables $(\xi_l)_{l\geqslant 1}$ sont centrées, décorrélées et de variance 1.

Nous supposons ici que les valeurs propres et les fonctions propres de la décomposition sont connues. Cette hypothèse semble raisonnable puisque dans le contexte de l'AS, nous supposons généralement les distributions des paramètres d'entrée connues. La décomposition de Karhunen-Loève est connue pour certaines distributions classiques (quelques exemples sont donnés dans la Section 2 de [J6]). Par exemple, lorsque X est un processus gaussien, les variables $(\xi_l)_{l\geqslant 1}$ sont des variables i.i.d. gaussiennes standard.

Dans le cas contraire où les distributions des entrées sont inconnues, il conviendrait au préalable de procéder à l'estimation des valeurs propres et des fonctions propres la décomposition de Karhunen-Loève à partir d'observation des processus d'entrées. Cette étape supplémentaire ne fait pas l'objet de notre étude.

L'indice de Sobol $S^X_{\mathbf{Lin}}$ et son estimation par des U-statistiques d'ordre 2

Puisque $\mathbb{E}[Y|X] = \mu + \langle \beta, X\rangle$, il vient que l'indice de Sobol associé à X est donné par

$$S^X_{\mathrm{Lin}} = \frac{\mathrm{Var}(\langle \beta, X\rangle)}{\mathrm{Var}(Y)}.$$

En décomposant β dans la base des fonctions propres, nous obtenons $\beta = \sum_{l=1}^{\infty} \gamma_l\varphi_l$ et le numérateur N^X_{Lin} de l'indice de Sobol S^X_{Lin} est donné par

$$N^X_{\mathrm{Lin}} = \mathrm{Var}(\mathbb{E}[Y|X]) = \mathrm{Var}(\langle \beta, X\rangle) = \mathbb{E}[\langle \beta, X\rangle\langle \beta, X\rangle] = \langle \beta, \Gamma(\beta)\rangle = \sum_{l=1}^{\infty} \lambda_l\gamma_l^2.$$

Or, nous avons aussi

$$\mathbb{E}[YX] = \mathbb{E}[\langle X, \beta\rangle X] = \Gamma(\beta) = \sum_{l=1}^{\infty} \lambda_l\gamma_l\varphi_l.$$

et donc $\gamma_l = \langle \mathbb{E}[YX], \varphi_l\rangle/\lambda_l$ que l'on estimera par son estimateur empirique sans biais :

$$\widehat{\gamma}_l = \frac{1}{\lambda_l}\frac{1}{N}\sum_{j=1}^{N} \langle X_j, \varphi_l\rangle Y_j,$$

où $(X_j, Y_j)_{1 \leqslant j \leqslant N}$ est un N-échantillon i.i.d. distribué selon la loi de (X, Y), avec $X \in \mathbb{H}$ et $Y \in \mathbb{R}$, suivant le modèle (4.31). Pour tout $m \in \mathbb{N}^*$, nous introduisons alors la U-statistique d'ordre 2 :

$$\widehat{N}^X_{\mathrm{Lin},m} = \sum_{l=1}^{m} \frac{1}{\lambda_l} \frac{1}{N(N-1)} \sum_{1 \leqslant j \neq j' \leqslant N} \langle X_j, \varphi_l \rangle Y_j \langle X_{j'}, \varphi_l \rangle Y_{j'}. \tag{4.32}$$

Nous avons $\mathbb{E}[\widehat{N}^X_{\mathrm{Lin},m}] = \sum_{l=1}^{m} \lambda_l \gamma_l^2$, *i.e.* $\widehat{N}^X_{\mathrm{Lin},m}$ est un estimateur biaisé de N^X_{Lin}. En divisant $\widehat{N}^X_{\mathrm{Lin},m}$ par la variance empirique de Y, nous obtenons un estimateur de S^X_{Lin}, noté $\widehat{S}^X_{\mathrm{Lin},m}$.

Dans la suite de cette section, nous étudions ses propriétés asymptotiques et proposons une valeur m adaptée pour la troncature.

Propriétés asymptotiques de $\widehat{S}^X_{\mathbf{Lin},m}$

D'après la décomposition de Hoeffding [120] de la U-statistique $\widehat{N}^X_{\mathrm{Lin},m}$, nous obtenons le résultat suivant.

Théorème 4.14 (Normalité asymptotique de $\widehat{S}^X_{\mathrm{Lin},m}$)**.** *Supposons que $\mathbb{E}[\|X\|^4] < +\infty$, que $\mathbb{E}[\varepsilon^4] < +\infty$ et enfin que*

$$\sup_{l \geqslant 1} \mathbb{E}[\xi_l^4] < +\infty. \tag{4.33}$$

Prenons $m = m(N) = \sqrt{N} h(N)$, où $h(N)$ satisfait : $h(N) \to 0$ et $\forall \alpha > 0$, $N^\alpha h(N) \to +\infty$ lorsque $n \to +\infty$. Supposons qu'il existe $C > 0$ et $\delta > 1$ tels que

$$\forall l \geqslant 1, \quad \lambda_l \leqslant C l^{-\delta}.$$

Alors

$$\sqrt{N}\left(\widehat{S}^X_{Lin,m} - S^X\right) \xrightarrow[N \to \infty]{\mathcal{L}} \mathcal{N}_1\left(0, \frac{\mathrm{Var}(U)}{(\mathrm{Var}(Y))^2}\right)$$

où $U = 2Y\langle X, \beta \rangle - S^X(Y - \mathbb{E}[Y])^2$.

Notons que la condition (4.33) est satisfaite lorque X est un processus gaussien, puisque dans ce cas, les variables $(\xi_l)_{l \geqslant 1}$ sont des gaussiennes standard i.i.d. Dans le cas où la variance de Y est connue, il n'est pas nécessaire de l'estimer par la variance empirique. Il suffit de normaliser l'estimation du numérateur par $\mathrm{Var}(Y)$.

Proposition 4.15 (Efficacité asymptotique de $\widehat{S}^X_{\mathrm{Lin},m}$)**.** *Sous les hypothèses du Théorème 4.14, la suite d'estimateurs $\left(\widehat{S}^X_{Lin,m}\right)_N$ est asymptotiquement efficace pour estimer S^X_{Lin}.*

4.4.3 Comparaison avec les estimateurs Pick-Freeze

Supposons que le couple (X, Y) suit le modèle (4.30), avec $X = (X_1, \ldots, X_p)$. Soit $X' = (X'_1, \ldots, X'_p)$ une copie i.i.d. de X. Pour tout $i \in I_p$, soit Y^{X_i}, notée simplement Y^i, la version Pick-Freeze relative à X_i définie par

$$Y^i = Y^{X_i} = \mu + \langle X_i, \beta_i \rangle + \sum_{i'=1, i' \neq i}^{p} \langle X'_{i'}, \beta_{i'} \rangle + \varepsilon',$$

où ε' est une copie i.i.d. de ε. En vue de la construction de l'estimateur Pick-Freeze, nous considérons le plan d'expérience donné par $(Y_1, \ldots, Y_N)$ où les Y_j sont des copies i.i.d. de Y et pour tout $i \in I_p$, soient $(Y_1^i, \ldots, Y_N^i)$ N copies i.i.d. de Y^i. L'estimateur Pick-Freeze $\widehat{T}^{\mathbf{U}}_{\mathrm{PF}}$ de $S^{\mathbf{U}}$ s'obtient en suivant la démarche de la Section 4.2.6 et nous avons le résultat suivant

$$\sqrt{N}\left(\widehat{T}^{\mathbf{U}}_{\mathrm{PF}} - S^{\mathbf{U}}\right) \xrightarrow[N \to \infty]{\mathcal{L}} \mathcal{N}_p\left(0, \Gamma_{\mathrm{PF}}\right).$$

où $\Gamma_{\mathrm{PF}} = \left(\frac{\mathrm{Cov}(V^i, V^{i'})}{(\mathrm{Var}(Y))^2} \right)_{i,i'=1\dots p}$ et $V^i = YY^i - S^{X_i}\left[Y^2 + \sum_{i"=1}^{p} (Y^{i"})^2\right]/(p+1)$.
Notre but est de comparer la variance limite Γ_{Lin} apparaissant dans le TCL pour l'estimateur basé sur les U-statistiques à $\widehat{T}^{\mathbf{U}}_{\mathrm{PF}}$ dont on obtient assez facilement des expressions explicites.
Le plan d'expérience de Sobol requiert $(p+1)N$ observations pour estimer les p indices. De façon à avoir une comparaison juste des deux méthodes, nous considérons que nous avons $(p+1)N$ observations i.i.d. issues du modèle (4.30) pour estimer les indices de Sobol dans chacune des deux méthodes et nous sommes conduits à étudier

$$D = (p+1)\Gamma_{\mathrm{PF}} - \Gamma_{\mathrm{Lin}}.$$

Pour ce faire, nous calculons les valeurs propres de D et déterminons si cette dernière est définie positive. Si elle ne l'est pas, nous étudions uniquement le signe de ses termes diagonaux qui correspondent à la différence des variances asymptotiques obtenue par les deux méthodes.

Exemple 4.16. *Supposons que $\mu = 0$ et que pour $i = 1, \dots, p$ avec $p \geqslant 2$,*

$$\langle X_i, \beta_i \rangle \sim \mathcal{N}_1(0,1).$$

Pour $p \geqslant 2$, la matrice D n'est ni définie positive ni définie négative. En outre, nous montrons que $D_{i,i} > 0$ pour tout $i \in I_p$ et que la variance de chaque estimateur de Sobol par la méthode Pick-Freeze est plus grande que celle obtenue par l'estimateur basé sur les U-statistiques.

4.5 Les indices de Sobol pour sorties vectorielles et fonctionnelles

Afin de motiver le travail de cette section et la nécessité de définir proprement des indices de Sobol lorsque la sortie est vectorielle voire fonctionnelle, nous commençons par étudier un exemple.

Exemple 4.17. *Considérons le modèle non linéaire suivant :*

$$Y = f^{a,b}(X_1, X_2) = \begin{pmatrix} f_1^{a,b}(X_1, X_2) \\ f_2^{a,b}(X_1, X_2) \end{pmatrix} = \begin{pmatrix} X_1 + X_1X_2 + X_2 \\ aX_1 + bX_1X_2 + X_2 \end{pmatrix} \tag{4.34}$$

où X_1 et X_2 sont des v.a. gaussiennes standard indépendantes.
Calculons d'abord les indices de Sobol tels que nous les avons définis pour les sorties scalaires en Section 4.2 : $S^{\mathbf{i}}(f_k^{a,b})$ de $f_k^{a,b}$ par rapport à X_i pour $i, k = 1, 2$. Nous obtenons

$$(S^{\mathbf{1}}(f_1^{a,b}), S^{\mathbf{1}}(f_2^{a,b})) = (1/3, a^2/(1+a^2+b^2))$$
$$(S^{\mathbf{2}}(f_1^{a,b}), S^{\mathbf{2}}(f_2^{a,b})) = (1/3, 1/(1+a^2+b^2)).$$

de telle sorte que les ratios $S^{\mathbf{1}}(f_k^{a,b})/S^{\mathbf{2}}(f_k^{a,b})$ pour $k = 1, 2$ ne dépendent pas de b. De plus, pour $|a| > 1$, puisque ce ratio est supérieur ou égal à 1, X_1 semble avoir plus d'influence sur la sortie que X_2.
Maintenant réalisons une analyse de sensibilté sur la norme 2 de la sortie. Nous obtenons facilement :

$$S^{\mathbf{1}}(\|Y\|^2) \geqslant S^{\mathbf{2}}(\|Y\|^2) \iff (a-1)(a^3 + a^2 + 5a + 5 - 4b) \geqslant 0.$$

Ainsi pour $\|Y\|^2$, la région où X_1 est la variable la plus influente dépend de la valeur de b. Cette region n'est pas intuitive. Elle est représentée dans la Figure 1 de [J7].

Ainsi l'étude de la norme 2 de la sortie n'est pas satisfaisante tandis que l'interprétation du vecteur des indices de Sobol scalaires n'est pas facile, comme nous venons de le voir. Une dernière motivation pour introduire de nouveaux indices de Sobol est liée au problème statistique de leur estimation. En effet, lorsque la dimension augmente, l'estimation du vecteur des indices Sobol scalaires devient de plus en plus coûteuse. Cela renforce le fait qu'il est nécessaire de définir des indices de Sobol résumant toute l'information contenue dans la collection des indices scalaires. C'est ce que nous nous proposons de faire dans cette section. Les résultats et preuves peuvent être consultés dans [N2] et [J7].

4.5.1 Généralisation de l'indice de Sobol

Dans cette section, le code est donné comme précédemment par (4.1) mais la fonction boîte noire est maintenant à valeurs dans $\mathbb{R}^k$. L'objectif de cette section reste le même : déterminer les variables les plus influentes. Nous supposons que Y est de carré intégrable. Sans perte de généralité, nous supposerons en outre que la matrice de variance-covariance de Y est définie positive.
De la même façon que dans la Section 4.2.1, Y peut être décomposé selon la décomposition de Hoeffding [249] :

$$\Sigma = C_{\mathbf{u}} + C_{\sim\mathbf{u}} + C_{\mathbf{u},\sim\mathbf{u}}. \tag{4.35}$$

Ici Σ, $C_{\mathbf{u}}$, $C_{\sim\mathbf{u}}$ et $C_{\mathbf{u},\sim\mathbf{u}}$ représentent les matrices de variance-covariance de Y, $f_{\mathbf{u}}(X_{\mathbf{u}})$, $f_{\sim\mathbf{u}}(X^{\sim\mathbf{u}})$ et $f_{\mathbf{u},\sim\mathbf{u}}(X^{\mathbf{u}}, X^{\sim\mathbf{u}})$ respectivement. Alors, (4.35) peut être ramenée dans $\mathbb{R}$ en prenant la trace :

$$\mathrm{Tr}(\Sigma) = \mathrm{Tr}(C^{\mathbf{u}}) + \mathrm{Tr}(C^{\sim\mathbf{u}}) + \mathrm{Tr}(C^{\mathbf{u},\sim\mathbf{u}}).$$

Ceci suggère de définir, du moment que $\mathrm{Tr}(\Sigma) \neq 0$, la mesure de sensibilité relative de Y par rapport à $X_{\mathbf{u}}$ comme

$$S^{\mathbf{u},k} = \frac{\mathrm{Tr}(C^{\mathbf{u}})}{\mathrm{Tr}(\Sigma)}.$$

Notons que nous obtenons les mêmes indices de sensibilité que ceux introduits dans [153] et basés sur une ACP. Remarquons que la condition $\mathrm{Tr}(\Sigma) \neq 0$ (nécessaire pour que les indices soient bien définis) est satisfaite dès que Y n'est pas constante. La proposition suivante est immédiate.

Proposition 4.18.

1) Les indices somment à 1 : $S^{\mathbf{u},k} + S^{\sim\mathbf{u},k} + S^{\mathbf{u},\sim\mathbf{u},k} = 1$.

2) $0 \leqslant S^{\mathbf{u},k} \leqslant 1$.

3) $S^{\mathbf{u},k}$ *est invariant par toute composition à gauche de* Y *par une isométrie de* $\mathbb{R}^k$ *i.e. pour toute matrice* O *de taille* k *t.q.* $O^\top O = Id_k$, $S^{\mathbf{u}}(OY) = S^{\mathbf{u},k}$;

4) $S^{\mathbf{u},k}$ *est invariant par toute composition à gauche de* f *par un changement d'échelle i.e. pour tout* $\lambda \in \mathbb{R}$, $S^{\mathbf{u}}(\lambda Y) = S^{\mathbf{u},k}$.

5) Pour $k = 1$ *et tout* $M \neq 0$*, nous retrouvons les indices de Sobol scalaires :* $S^{\mathbf{u},1} = S^{\mathbf{u}}$.

Remarquons que pour toute matrice carrée M de taille k telle que $\mathrm{Tr}(M\Sigma) \neq 0$, l'indice défini par

$$S^{\mathbf{u}}(M) = \frac{\mathrm{Tr}(MC^{\mathbf{u}})}{\mathrm{Tr}(M\Sigma)}$$

satisfait les propriétés 1), 2) et 5) de la proposition précédente. Cependant, seul le choix *canonique* $M = \lambda \mathrm{Id}_k$ ($\lambda \in \mathbb{R}^*$) remplit les propriétés d'invariance.

Exemple 4.17 (suite). *Reprenons maintenant le modèle défini par* (4.34) *afin d'illustrer ces nouveaux indices. Nous avons*

$$S^{\mathbf{1},2}(f^{a,b}) = \frac{1+a^2}{4+a^2+b^2} \quad et \quad S^{\mathbf{2},2}(f^{a,b}) = \frac{2}{4+a^2+b^2}$$

et de manière évidente

$$S^{\mathbf{1},2}(f^{a,b}) \geqslant S^{\mathbf{2},2}(f^{a,b}) \iff a^2 \geqslant 1. \tag{4.36}$$

Ce résultat a une interprétation naturelle : puisque X_1 est multiplié par a, il a plus d'infuence si on agrandit son support i.e. dès que $|a| > 1$.

Quid de l'unicité ?

Nous montrons maintenant qu'il est possible de construire d'autres indices ayant les même propriétés d'invariance que $S^{\mathbf{u},k}$. A cet effet, nous considérons (4.35) selon un autre angle de façon à avoir une définition naturelle d'un indice de Sobol matriciel par rapport aux variables $X_{\mathbf{u}}$:

$$BC_{\mathbf{u}}A \tag{4.37}$$

pour toutes matrices A et B telles que $AB = \Sigma^{-1}$. Tout d'abord, notons que cet indice est une matrice carrée de taille k. Ensuite, toute combinaison convexe d'indices de Sobol matriciels de la forme (4.37) est encore un bon candidat pour définir un indice de Sobol matriciel par rapport à $X_{\mathbf{u}}$.

Afin de garantir que l'indice de Sobol matriciel a une définition consistante, nous devons exiger un peu plus. En premier lieu, une condition raisonnable est que l'indice de Sobol matriciel est une matrice symétrique : l'influence de l'entrée X_i sur les coordonnées k et l de la sortie Y doit être la même que l'influence de l'entrée X_i sur les coordonnées l et k de Y. Deuxièmement, l'indice de Sobol matriciel doit partager les propriétés de l'indice scalaire $S^{\mathbf{u},1}$. Autrement dit, il doit être invariant par toute isométrie, changement d'échelle et translation. Cela conduit à l'indice de Sobol matriciel suivant

$$T^{\mathbf{u},k} := T^{\ \mathbf{u},k,\mu^\star} = \frac{1}{2}\left(\int_{\mathcal{H}_k} (OP)^t \left(\Sigma^{-1}C_{\mathbf{u}} + C_{\mathbf{u}}\Sigma^{-1}\right) OP\mu^\star(\mathrm{d}P)\right) = \frac{\mathrm{Tr}\left(\Sigma^{-1}C_{\mathbf{u}}\right)}{k} I_k \tag{4.38}$$

où $\mu^\star$ la loi uniforme sur un ensemble fini du groupe $\mathcal{H}_k$ des matrices de permutations signées de longueur k. Cf. [J7] pour les détails du raisonnement. Remarquons que cet indice dépend uniquement de la quantité $\mathrm{Tr}\left(\Sigma^{-1}C_{\mathbf{u}}\right)/\mathrm{Tr}(I_k)$, facile à interpréter.

Comparaison entre $S^{\mathbf{u},k}$ et $T^{\mathbf{u},k}$

La question naturelle qui se pose maintenant est la suivante. Quel indice doit être préféré entre $S^{\mathbf{u},k}$ et $T^{\mathbf{u},k}$? Il n'existe pas de réponse universelle *a priori*. Cependant, d'un point de vue statistique, $T^{\mathbf{u},k}$ présente un désavantage majeur : son estimation requiert celle de l'inverse d'une matrice de variance-covariance Σ^{-1} qui peut s'avérer délicate. A contrario, celle de $S^{\mathbf{u},k}$ nécessite seulement l'estimation de traces de matrices de variance-covariance. En outre, l'exemple suivant montre que $T^{\mathbf{u},k}$ peut même s'avérer inutile dans certains cas particuliers.

Exemple 4.17 (suite). *Considérons à nouveau le modèle défini par* (4.34). *Facilement, nous avons*

$$T^{\mathbf{1},2} = \frac{(b-a)^2+(a-1)^2}{4[(b-a)^2+(a-1)(b-1)]} I_2, \qquad T^{\mathbf{2},2} = \frac{(b-1)^2+(a-1)^2}{4[(b-a)^2+(a-1)(b-1)]} I_2.$$

Donc

$$T^{\mathbf{1},2} \geqslant T^{\mathbf{2},2} \iff (a-1)(a-2b+1) \geqslant 0$$

tandis que nous avions obtenu précédemment le résultat bien plus intuitif (4.36)*. De plus,* $T^{\mathbf{u},2}$ *n'est pas informatif puisque pour* $a = 1$*, les indices* $T^{\mathbf{1},2}$ *et* $T^{\mathbf{2},2}$ *satisfont*

$$T^{\mathbf{1},2} = T^{\mathbf{2},2} = \frac{1}{4} I_2$$

et ne dépendent pas de b*.*
Ainsi, il semble que $S^{\mathbf{u},k}$ constitue un indice de sensibilité plus pertinent. Dans la suite de cette section, nous nous focaliserons sur $S^{\mathbf{u},k}$ et son estimation.

4.5.2 L'estimateur Pick-Freeze de $S^{\mathbf{u},k}$ et ses propriétés

Dans cette section, nous proposons un estimateur Pick-Freeze pour le cas vectoriel qui généralise l'estimateur $\widehat{T}^{\mathbf{u}}$ défini par (4.9). Nous reprenons les notations précédentes et notons $Y_{l,j}$ (resp. $Y^{\mathbf{u}}_{l,j}$) la l-ème coordonnée de Y_j (resp. $Y^{\mathbf{u}}_j$). Nous définissons alors l'estimateur de $S^{\mathbf{u},k}$ par le ratio des traces des versions empiriques des matrices de covariance $C^{\mathbf{u}}$ et Σ :

$$\widehat{S}^{\mathbf{u},k} = \frac{\mathrm{Tr}\left(\widehat{C}^{\mathbf{u}}_N\right)}{\mathrm{Tr}\left(\widehat{\Sigma}_N\right)}. \tag{4.39}$$

Proposition 4.19 (Consistance et normalité asymptotique de $\widehat{S}^{\mathbf{u},k}$)**.** *Par la loi forte des grands nombres, l'estimateur* $\widehat{S}^{\mathbf{u},k}$ *converge p.s. vers* $S^{\mathbf{u},k}$ *quand* $N \to +\infty$*. Supposons en outre que* $\mathbb{E}[Y_l^4] < \infty$ *pour tout* $l = 1, \ldots, k$ *et posons*

$$U_l = (Y_{l,1} - \mathbb{E}[Y_l])(Y^{\mathbf{u}}_{l,1} - \mathbb{E}[Y_l]), \qquad V_l = (Y_{l,1} - \mathbb{E}[Y_l])^2 + (Y^{\mathbf{u}}_{l,1} - \mathbb{E}[Y_l])^2.$$

Alors

$$\sqrt{N}\left(\widehat{S}^{\mathbf{u},k} - S^{\mathbf{u},k}\right) \xrightarrow[N\to\infty]{\mathcal{L}} \mathcal{N}_1\left(0, \sigma^2\right) \tag{4.40}$$

où

$$\sigma^2 = a^2 \sum_{l,l' \in I_k} \mathrm{Cov}(U_l, U_{l'}) + b^2 \sum_{l,l' \in I_k} \mathrm{Cov}(V_l, V_{l'}) + 2ab \sum_{l,l' \in I_k} \mathrm{Cov}(U_l, V_{l'}), \tag{4.41}$$

avec $a = \left(\sum_{l=1}^k \mathrm{Var}(Y_l)\right)^{-1}$ *et* $b = -(a/2)S^{\mathbf{u},k}$*.*

Proposition 4.20 (Efficacité asymptotique de $\widehat{S}^{\mathbf{u},k}$)**.** *Supposons que* $\mathbb{E}[Y_l^4] < \infty$ *pour* $l = 1, \ldots, k$*. Alors* $\left(\widehat{S}^{\mathbf{u},k}\right)_N$ *est asymptotiquement efficace pour estimer* $S^{\mathbf{u},k}$*.*

De même que dans la Section 4.2.4, il est possible d'établir des inégalités de concentration pour $\widehat{S}_{\mathbf{u},k}$. Voir [J7].

4.5.3 Les indices de Sobol pour sorties fonctionnelles

En pratique, il arrive aussi que la sortie du code boîte noire soit une fonction. Il est donc intéressant d'étendre la définition des indices de Sobol aux sorties fonctionnelles. C'est l'objet de cette section.

Définition d'un indice pour sorties fonctionnelles

Soit $\mathbb{H}$ un espace de Hilbert séparable muni du produit sclaire $\langle \cdot, \cdot \rangle$ et de la norme $\|\cdot\|$. Soit f une

fonction à valeurs dans $\mathbb{H}$, *i.e.*, Y et $Y^{\mathbf{u}}$ sont des variables aléatoires à valeurs dnas $\mathbb{H}$. Supposons que $\mathbb{E}\left[\|Y\|^2\right] < \infty$. Rappelons que $\mathbb{E}[Y]$ est défini par dualité comme l'unqiue élément de $\mathbb{H}$ satisfaisant

$$\mathbb{E}[\langle h, Y\rangle] = \langle h, \mathbb{E}[Y]\rangle \quad \text{pour tout} \quad h \in \mathbb{H}.$$

Rappelons que l'opérateur de covariance associé à Y est l'endomorphisme Γ de $\mathbb{H}$ défini, pour tout $h \in \mathbb{H}$ par $\Gamma(h) = \mathbb{E}[\langle Y, h\rangle Y]$. Rappelons aussi que $\mathbb{E}\left[\|Y\|^2\right] < \infty$ implique que Γ est alors un opérateur de type trace et que sa trace est bien définie.

Nous généralisons la définition de $S^{\mathbf{u},k}$ introduite en Section 4.5.1 aux sorties fonctionnelles de la façon suivante :

$$S^{\mathbf{u},\infty} = \frac{\mathrm{Tr}(\Gamma_{\mathbf{u}})}{\mathrm{Tr}(\Gamma)},$$

où $\Gamma_{\mathbf{u}}$ est l'endomorphisme de $\mathbb{H}$ défini par $\Gamma_{\mathbf{u}}(h) = \mathbb{E}[\langle Y^{\mathbf{u}}, h\rangle Y]$ pour tout $h \in \mathbb{H}$.

Estimation de $S^{\mathbf{u},\infty}$

Pour mettre en place, la procédure d'estimation de l'indice $S^{\mathbf{u},\infty}$, nous utilisons la décomposition polaire des traces de Γ et $\Gamma_{\mathbf{u}}$:

$$\mathrm{Tr}(\Gamma) = \mathbb{E}\left[\|Y\|^2\right] - \|\mathbb{E}[Y]\|^2 \quad \text{et} \quad \mathrm{Tr}(\Gamma_{\mathbf{u}}) = \frac{1}{4}\left[\mathbb{E}\left[\|Y + Y^{\mathbf{u}}\|^2\right] - \mathbb{E}\left[\|Y - Y^{\mathbf{u}}\|^2\right] - 4\,\|\mathbb{E}[Y]\|^2\right].$$

Soit maintenant $(\varphi_l)_{1\leqslant l}$ une base orthonormale de $\mathbb{H}$. Alors

$$\|Y\|^2 = \sum_{i=1}^{\infty} \langle Y, \varphi_i\rangle^2.$$

Nous tronquons ensuite les sommes précédentes et posons

$$\|Y\|_m^2 := \sum_{i=1}^{m} \langle Y, \varphi_i\rangle^2 \quad \text{et} \quad \|Y^{\mathbf{u}}\|_m^2 := \sum_{i=1}^{m} \langle Y^{\mathbf{u}}, \varphi_i\rangle^2.$$

Cela revient à tronquer le développement de Y à un niveau donné m. Soit Y_m l'approximation associée de Y :

$$Y_m = \sum_{l=1}^{m} \langle Y, \varphi_i\rangle \varphi_l,$$

vue comme un vecteur de longueur m. Nous allons maintenant appliquer les résultats de la Section 4.5.2 à Y_m. Ainsi, l'estimateur de $S^{\mathbf{u},\infty}$ est défini de la manière suivante :

$$\widehat{S}_m^{\mathbf{u},\infty} = \frac{\frac{1}{4N}\sum_{i=1}^{N}\left(\|Y_i + Y_i^{\mathbf{u}}\|_m^2 - \|Y_i - Y_i^{\mathbf{u}}\|_m^2 - \left\|\overline{Y} + \overline{Y^{\mathbf{u}}}\right\|_m^2\right)}{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{\|Y_i\|_m^2 + \left\|Y_i^{\mathbf{u}}\right\|_m^2}{2} - \left\|\frac{\overline{Y}+\overline{Y^{\mathbf{u}}}}{2}\right\|_m^2\right)}.$$

Propriétés asymptotiques

Nous procédons ensuite comme dans [J6] en décomposant les termes précédents en termes de U-statistiques d'ordre 2 totalement dégénérée, de termes linéaires centrés et de restes déterministes. Cela conduit au résultat suivant.

Théorème 4.21. *Supposons qu'il existe $\delta > 1$ et $\delta' > 1$ tesl que*

$$\mathbb{E}\left[\langle T, \varphi_l\rangle^2\right] = O(l^{-(\delta+1)}) \quad \textit{et} \quad \mathbb{E}\left[\langle T, \phi_l\rangle^4\right] = O(l^{-\delta'}) \tag{4.42}$$

pour $T = Y$, $Y^{\mathbf{u}}$, $Y - Y^{\mathbf{u}}$ *et* $Y + Y^{\mathbf{u}}$. *Alors pour tout* $m = m(N)$ *tel que :* $m(N)/N^{\frac{1}{2\delta}} \to +\infty$ *et* $m(N)/\sqrt{N} \to 0$, *nous avons*

$$\sqrt{N}(\widehat{S}_m^{\mathbf{u},\infty} - S^{\boldsymbol{u},\infty}) \xrightarrow[N\to\infty]{\mathcal{L}} \mathcal{N}_1(0,\sigma^2) \tag{4.43}$$

où σ^2 *est explicitement donné dans [J7].*

4.6 Au-delà de la variance et de l'ordre 2

Les indices de Sobol et leur estimation Pick-Freeze Monte-Carlo sont des méthodes bâties sur des moments d'ordre 2 maximum puisqu'elles découlent de la décomposition de Hoeffding (e.g., (4.2), (4.35)). Il peut s'avérer judicieux de considérer des indices prenant en compte toutes les caractéristiques des lois. A titre d'illustration, considérons l'exemple suivant.

Exemple 4.22. *Considérons le modèle linéaire suivant :*

$$Y = \alpha X_1 + X_2, \quad \alpha > 0, \tag{4.44}$$

où X_1 *suit une loi de Bernoulli de paramètre* $0 < p < 1$, X_2 *est une variable continue de FR* F_2 *sur* $\mathbb{R}$ *telle que* $\mathbb{E}[X_2] = \alpha p$ *et* $\mathrm{Var}(X_2) = \alpha^2 p(1-p)$ *et* X_1, X_2 *sont indépendantes. On peut facilement voir que* αX_1 *et* X_2 *ont alors la même espérance et la même variance. Ainsi, elles ont aussi les mêmes indices de Sobol (=1/2). Cependant, n'ayant pas la même distribution,* X_1 *et* X_2 *ne doivent certainement pas avoir la même influence sur la sortie. Cette différence n'est pas lisible dans les indices de Sobol.*

Cela montre la nécessité d'introduire un indice de sensibilité prenant en compte toute la distribution et pas seulement le comportement au second ordre. Comme discuté précédemment, les indices de Sobol sont fondés sur une décomposition L^2. Ils sont donc bien adaptés pour mesurer la contribution d'une entrée du code à la déviation autour de la moyenne de Y. Néanmoins, il semble très intuitif par exemple que la sensibilité d'un quantile extrême de Y nesoit pas correctement expliquée en utilisant seulement les variances. Ainsi, chaque objectif devrait faire appel à un indice spécifique adapté.

Comme souligné dans [36, 38, 39, 192, 191], dans certains cas pratiques, des méthodes d'ordre supérieur conduisent à une analyse plus fine sur l'influence relative de chacune des entrées et mènent à des procédures de sélection des variables influentes plus précises. Dans cette veine, Owen et ses co-auteurs suggèrent d'utiliser des procédures basées sur des moments plus élevés (voir [192, 191]). L'objectif de cette section reste le même : déterminer les variables les plus influentes. Dans cette section, nous revisitons le travail de Owen [192, 191] en étudiant les propriétés asymptotiques de l'estimateur Pick-Freeze d'ordre supérieur. Puis, nous proposons un nouvel indice basé sur la distance de Cramér-von Mises entre la distribution de Y et sa distribution conditionnelle lorsqu'une entrée (ou un ensemble d'entrées) est fixé. L'étude menée dans cette section a fait l'objet de la publication [J14].

4.6.1 Une première piste vers la généralisation des indices de Sobol

Tout d'abord remarquons que le numérateur de l'indice de Sobol défini en (4.4) peut se réécrire de la façon suivante :

$$\mathrm{Var}\left(\mathbb{E}\left[Y|X_{\mathbf{u}}\right]\right) = \mathbb{E}\left[\left(\mathbb{E}[Y|X_{\mathbf{u}}] - \mathbb{E}[Y]\right)^2\right] = \mathrm{Var}(Y) - \mathbb{E}\left[\left(\mathbb{E}[Y] - \mathbb{E}\left[Y|X_{\mathbf{u}}\right]\right)^2\right]. \tag{4.45}$$

Comme expliqué précédemment, nous suivons [192, 191] et généralisons les indices de Sobol en considérant des moments q, pour $q \geqslant 2$:

$$S_q^{\mathbf{u}} := \frac{\mathbb{E}\left[(\mathbb{E}[Y|X_{\mathbf{u}}] - \mathbb{E}[Y])^q\right]}{\mathrm{Var}(Y)}, \quad \text{pour} \quad \mathbf{u} \subset I_p.$$

Evidemment, $S_q^{\mathbf{u}}$ est positif dès que q est pair, est invariant par toute translation de la sortie Y, de plus

$$\left|S_q^{\mathbf{u}}\right| \leqslant \frac{\mathbb{E}\left[|Y - \mathbb{E}[Y]|^q\right]}{\mathrm{Var}(Y)}.$$

Estimation de $S_q^{\mathbf{u}}$

En vue de l'estimation de $S_q^{\mathbf{u}}$, remarquons que

$$\mathbb{E}\left[(\mathbb{E}[Y|X_{\mathbf{u}}] - \mathbb{E}[Y])^q\right] = \mathbb{E}\left[\prod_{i=1}^{q}\left(Y^{\mathbf{u},i} - \mathbb{E}[Y]\right)\right] = \sum_{l=0}^{q}\binom{q}{l}(-1)^{q-l}\mathbb{E}\left[Y\right]^{q-l}\mathbb{E}\left[\prod_{i=1}^{l}Y^{\mathbf{u},i}\right]$$

avec la convention usuelle $\prod_{i=1}^{0} Y^{\mathbf{u},i} = 1$ et $\binom{q}{l} = q!/l!(q-l)!$. Ici, $Y^{\mathbf{u},1} = Y$ et $i = 2, \ldots, q$, $Y^{\mathbf{u},i}$ est une copie indépendante de Y^v, la version Pick-Freeze de Y, définie en (4.5).

Ensuite, nous utilisons une nouvelle fois un schéma Monte-Carlo et considérons le plan d'expérience Pick-Freeze suivant constitué d'un échantillon de taille N : $\left(Y_j^{\mathbf{u},i}\right)_{(i,j)\in I_q\times I_N}$ de $\left(Y^{\mathbf{u},1}, \ldots, Y^{\mathbf{u},q}\right)$. L'estimateur de $S_q^{\mathbf{u}}$ est alors donné par

$$\widehat{S}_q^{\mathbf{u}} = \frac{\sum_{l=0}^{q}\binom{q}{l}(-1)^{q-l}\left(\overline{P}_1^{\mathbf{u}}\right)^{q-l}\overline{P}_l^{\mathbf{u}}}{\frac{1}{N}\sum_{j=1}^{N}\frac{1}{q}\sum_{i=1}^{q}(Y_j^{\mathbf{u},i})^2 - \left(\frac{1}{N}\sum_{j=1}^{N}\frac{1}{q}\sum_{i=1}^{q}Y_j^{\mathbf{u},i}\right)^2}$$

où pour tout $N \in \mathbb{N}^*$, $j \in I_N$ et $l \in I_q$,

$$P_{l,j}^{\mathbf{u}} = \binom{q}{l}^{-1}\sum_{k_1<\ldots<k_l\in I_q}\left(\prod_{i=1}^{l}Y_j^{\mathbf{u},k_i}\right) \quad \text{et} \quad \overline{P}_l^{\mathbf{u}} = \frac{1}{N}\sum_{j=1}^{N}P_{l,j}^{\mathbf{u}}.$$

Ainsi nous avons généralisé la procédure d'estimation de [100] et mis à profit toute l'information contenue dans l'échantillon en faisant des moyennes sur les ensembles d'indices $k_1, \ldots, k_l \in I_d,\ k_n \neq k_m$.

Théorème 4.23 (Consistance et normalité asymptotique de $\widehat{S}_q^{\mathbf{u}}$)**.** *$\widehat{S}_q^{\mathbf{u}}$ est consistant pour estimer $S_q^{\mathbf{u}}$ et il est aussi asymptotiquement gaussien :*

$$\sqrt{N}\left(\widehat{S}_q^{\mathbf{u}} - S_q^{\mathbf{u}}\right) \xrightarrow[N\to\infty]{\mathcal{L}} \mathcal{N}_1\left(0, \sigma_q^2\right)$$

où σ_q^2 s'exprime de façon analytique (cf. [J14]).

Interprétation et commentaires

La collection de tous les indices $\left(S_q^{\mathbf{u}}\right)_q$ est beaucoup plus informative que l'indice de Sobol classique par rapport à $\mathbf{u}$. Néanmoins, elle a plusieurs inconvénients. Tout d'abord, ces indices sont basés sur les moments et deviennent instables lorsque l'ordre augmente. Ensuite, ils peuvent être négatifs quand q est impair. Pour pallier ce problème, on peut introduire $\mathbb{E}\left[|\mathbb{E}[Y|X_{\mathbf{u}}] - \mathbb{E}[Y]|^q\right]$ mais alors la procédure d'estimation Pick-Freeze ne pourrait être mise en œuvre. Enfin, la procédure d'estimation Pick-Freeze devient très coûteuse et peut être instable. En effet, elle nécessite un échantillon de taille $q \times N$ de la sortie Y. Afin d'avoir une bonne idée de l'influence d'une entrée (ou d'un ensemble d'entrées) sur la loi de

la sortie, nous devons estimer les $K-1$ premiers indices $S_q^{\mathbf{u}} : S_2^{\mathbf{u}}, \ldots, S_K^{\mathbf{u}}$. Par conséquent, nous devons évaluer le code $K \times N$ fois. En conclusion, ces indices ne sont pas attrayants d'un point de vue pratique. Dans la section suivante, nous introduisons un nouvel indice de sensibilité qui est basé sur la distribution conditionnelle de la sortie et nécessite seulement $3 \times N$ évaluations de la sortie.

4.6.2 Définition des indices de Cramér-von Mises

La sortie du code est désormais notée $Z = f(X_1, \ldots, X_p) \in \mathbb{R}^k$. Il est important de noter que nous pouvons considérer des sorties vectorielles contrairement au contexte de la Section 4.6.1 et [40]. Soit F la FR de Z :

$$F(t) = \mathbb{P}\left(Z \leqslant t\right) = \mathbb{E}\left[\mathbb{1}_{\{Z \leqslant t\}}\right], \text{ pour } t = (t_1, \ldots, t_k) \in \mathbb{R}^k$$

et $F^{\mathbf{u}}$ la FR de Z conditionnellement à $X_{\mathbf{u}}$:

$$F^{\mathbf{u}}(t) = \mathbb{P}\left(Z \leqslant t | X_{\mathbf{u}}\right) = \mathbb{E}\left[\mathbb{1}_{\{Z \leqslant t\}} | X_{\mathbf{u}}\right], \text{ pour } t = (t_1, \ldots, t_k) \in \mathbb{R}^k.$$

Ici $\leqslant$ désigne l'ordre lexicographique sur $\mathbb{R}^d$ et $\{Z \leqslant t\}$ signifie donc $\{Z_1 \leqslant t_1, \ldots, Z_k \leqslant t_k\}$. Puisque pour tout $t \in \mathbb{R}^k$, $Y(t) = \mathbb{1}_{\{Z \leqslant t\}}$ est une v.a. scalaire, nous pouvons appliquer la procédure présentée en Section 4.2 et réaliser la décomposition de Hoeffding de $Y(t)$:

$$\mathrm{Var}(Y(t)) = F(t)(1-F(t)) = \mathbb{E}\left[\left(F^{v}(t) - F(t)\right)^2\right] + \mathbb{E}\left[\left(F^{\sim v}(t) - F(t)\right)^2\right] + \mathrm{Var}(R(t,v)) \tag{4.46}$$

où $R(t,v)$ désigne le reste. Il reste à intégrer chaque terme de (4.46) en $t \in \mathbb{R}^k$ par rapport à la loi de Z et à normaliser pour obtenir :

$$S_{2,CVM}^{\mathbf{u}} := \frac{\int_{\mathbb{R}^k} \mathbb{E}\left[\left(F(t) - F^{\mathbf{u}}(t)\right)^2\right] \mathrm{d}F(t)}{\int_{\mathbb{R}^k} F(t)(1-F(t))\mathrm{d}F(t)} \quad \text{et} \quad S_{2,CVM}^{\sim\mathbf{u}} := \frac{\int_{\mathbb{R}^k} \mathbb{E}\left[\left(F(t) - F^{\sim\mathbf{u}}(t)\right)^2\right] \mathrm{d}F(t)}{\int_{\mathbb{R}^k} F(t)(1-F(t))\mathrm{d}F(t)}.$$

Ces indices sont naturellement adaptés aux sorties vectorielles et satisfont les mêmes propriétés que les indices de Sobol. A savoir,

1) les différentes contributions somment à 1.

2) ils sont invariants par translation, isométrie et par tout changement d'échelle non dégénéré des composantes de Y.

4.6.3 Estimation de $S_{2,CVM}^{\mathbf{u}}$ et propriétés asymptotiques

Le procédure d'estimation se fait selon deux schémas Monte-Carlo en considérant le plan d'expériences suivant.

1) Un échantillon classique Pick-Freeze de taille N de Z : $(Z_j^{\mathbf{u},1}, Z_j^{\mathbf{u},2})$, $1 \leqslant j \leqslant N$;

2) Un troisième échantillon de taille N de Z indépendant de $(Z_j^{\mathbf{u},1}, Z_j^{\mathbf{u},2})_{1 \leqslant j \leqslant N}$: W_k, $1 \leqslant k \leqslant N$.

Le numérateur de $S_{2,CVM}^{\mathbf{u}}$ est estimé par

$$\frac{1}{N}\sum_{k=1}^{N}\left\{\frac{1}{N}\sum_{j=1}^{N}\mathbb{1}_{\{Z_j^{\mathbf{u},1} \leqslant W_k\}}\mathbb{1}_{\{Z_j^{\mathbf{u},2} \leqslant W_k\}} - \left[\frac{1}{2N}\sum_{j=1}^{N}\left(\mathbb{1}_{\{Z_j^{\mathbf{u},1} \leqslant W_k\}} + \mathbb{1}_{\{Z_j^{\mathbf{u},2} \leqslant W_k\}}\right)\right]^2\right\}, \tag{4.47}$$

tandis que son dénominateur est estimé par

$$\frac{1}{N}\sum_{k=1}^{N}\left\{\frac{1}{2N}\sum_{j=1}^{N}\left(\mathbb{1}_{\{Z_j^{\mathbf{u},1}\leqslant W_k\}}+\mathbb{1}_{\{Z_j^{\mathbf{u},2}\leqslant W_k\}}\right)-\left[\frac{1}{2N}\sum_{j=1}^{N}\left(\mathbb{1}_{\{Z_j^{\mathbf{u},1}\leqslant W_k\}}+\mathbb{1}_{\{Z_j^{\mathbf{u},2}\leqslant W_k\}}\right)\right]^2\right\}. \tag{4.48}$$

Théorème 4.24 (Consistance et normalité asymptotique de $\widehat{S}^{\mathbf{u}}_{2,CVM}$)**.** *La suite des estimateurs $\widehat{S}^{\boldsymbol{u}}_{2,CVM}$ est asymptotiquement gaussienne pour estimer $S^{\boldsymbol{u}}_{2,CVM}$ lorsque N tend vers l'infini : la variable aléatoire $\sqrt{N}\left(\widehat{S}^{\boldsymbol{u}}_{2,CVM}-S^{\boldsymbol{u}}_{2,CVM}\right)$ converge en loi vers une variable aléatoire centrée gaussienne dont la variance limite s'exprime explicitement. Cf. équation (12) dans [J14].*

Lorsque Z a des coordonnées indépendantes absolument continues par rapport à la mesure de Lebesgue, nous avons

$$\int_{\mathbb{R}^k} F(t)(1-F(t))\mathrm{d}F(t)=\mathbb{E}[F(Z)(1-F(Z)]=\frac{1}{2^k}-\frac{1}{3^k},$$

et le terme de normaliation est simplement égal à $\frac{1}{2^k}-\frac{1}{3^k}$. Dans ce cas particulier, il suffit donc d'établir un TCL pour l'estimateur du numérateur de l'indice $\widehat{S}^{\mathbf{u}}_{2,CVM}$.

4.6.4 Commentaires sur les indices de Cramér-von Mises et leur estimation

Indices de Cramér-von Mises et indices de Sobol

Les indices de Cramér-von Mises, de même que les indices de Sobol, sont basés sur la décomposition de Hoeffding et somment à 1. Comme nous l'avons déjà dit, les premiers sont basées sur la distribution de la sortie, en contraste avec les seconds qui ne prennent en compte que le moment d'ordre 2. Comme nous l'avons vu dans l'exemple introductif, deux variables peuvent avoir une influence différente sur la sortie mais les mêmes indices Sobol. Ce point représente une limite des indices Sobol et ne se produit pas avec les indices Cramér-von Mises comme on peut le voir dans l'Exemple 4.22.

En outre, un indice de Sobol nul ne signifie pas que l'entrée n'est pas importante tandis qu'un indice de Cramér-von Mises nul signifie que l'entrée n'est pas importante. Par définition, une grande valeur pour un indice de Cramér-von Mises signifie que la variable d'entrée concernée a une grande influence sur la sortie dans les régions portées par la distribution de la sortie.

Indices de Cramér-von Mises et indices indépendants des moments

Il existe déjà dans la littérature plusieurs indices indépendants des moments : certains d'entre eux ont été introduits par Borgonovo et ses co-auteurs (e.g., des indices basés sur la densité [38], des indices basés sur la fonction de répartition [41]). Voir aussi [37] pour d'autres indices et des références. Plus récemment, Da Veiga [75] montre que ces indices sont des cas particuliers d'une classe d'indices de sensibilité basés sur la f-divergence de Csizár. De nombreuses " distances" classiques entre les mesures de probabilité, comme, par exemple, la divergence de Kullback-Leibler, la distance de Hellinger et la distance en variation totale appartiennent à cette famille de divergences. D'autres mesures de dissimilarité existent pour comparer les distributions de probabilités : en particulier, les mesures de probabilité intégrales [185]. Ces indices sont estimés efficacement en faibles dimensions, mais comme l'affirme l'auteur de [75] : "it is well known that density estimation suffers from the curse of dimensionality". En revanche, comme nous l'avons vu, on peut facilement estimer les indices de Cramér-von Mises avec un faible coût de simulation qui ne dépend pas de la dimension de la sortie. L'échantillon requis pour leur estimation fournit également une estimation des indices de Sobol. En outre, ces estimateurs sont asymptotiquement normaux et convergent à la vitesse $\sqrt{N}$, ce qui permet en pratique de construire des IC.

Comparativement aux indices définis par l'équation (17) dans [41] pour lesquels l'intégration se fait par rapport à la mesure de Lebesgue, celle intervenant dans les indices de Cramér-von Mises se fait par rapport à la distribution de la sortie. En ce sens, notre méthode présente au moins deux avantages : (i) l'indice existe toujours quelle que soit la distribution de la sortie (ii) une telle intégration charge le support de la distribution de la sortie.

4.6.5 Applications numériques

Exemple 4.22 *(suite) Considérons à nouveau le modèle défini par* (4.44)*. Calculons maintenant les indices de Cramér-von Mises. La FR de* Y *est donnée par le mélange* $pF_2(\cdot-\alpha)+(1-p)F_2(\cdot)$*, ce qui conduit à*

$$S^1_{2,CVM} = 6p(1-p)\int_{\mathbb{R}} (F_2(t)-F_2(t-\alpha))^2\left[(1-p)\mathrm{d}F_2(t)+p\mathrm{d}F_2(t-\alpha)\right]$$

et

$$S^2_{2,CVM} = 1-6p(1-p)\left[\frac{1}{2}-\int_{\mathbb{R}} F_2(t-\alpha)\mathrm{d}F_2(t)\right]$$

(le terme de normalisation est simplement 1/6 comme expliqué précédemment).
Lorsque p *tend vers* 0 *(et* α *vers l'infini),* $(S^1_{2,CVM}, S^2_{2,CVM})$ *tend vers* $(0,1)$ *tandis que les indices de Sobol valent* $1/2$ *quelle que soit la valeur de* p*. Les indices de Cramér-von Mises traduisent bien le fait que, pour les petites valeurs de* p*,* X_2 *est plus influent sur* Y *que* X_1 *ce qui est conforme à l'intuition.*

Application à des données réelles : The Giant Cell Arthritis Problem
Nous considérons maintenant le problème du traitement de l'arthrite posé par Bunchbinder et Detsky [46]. Plus récemment, ce problème a également été étudié par Felli et Hazen [96] et Borgonovo *et al* dans [40]. Comme il est expliqué dans [46], l'arthrite à cellules géante (GCA) est une pathologie qui peut entraîner de graves complications (comme la perte d'accuité visuelle, de la fièvre, des maux de tête,...) Pire, une absence de traitement peut conduire à la cécité et l'occlusion des principaux vaisseaux. Les patients susceptibles d'avoir contracté la GCA reçoivent un traitement à base de Prednisone. Ainsi, face à un patient potentiellement atteint de cette pathologie, le médecin a le choix entre quatre stratégies introduites dans [46] :

A : Ne pas traiter ;

B : Réaliser une biopsie et traiter si le patient est positif ;

C : Réaliser une biopsie et traiter le patient quel que soit le résultat de sa biopsie ;

D : Traiter le patient.

Malheureusement, un traitement à fortes doses de Prednisone peut causer de graves complications. Aussi, lorsqu'un patient est susceptible d'avoir une GCA, le clinicien veut adopter la stratégie optimale. Les risques sont quantifiés à l'aide d'une fonction dite *fonction d'utilité*. Nous renvoyons le lecteur à [189] pour plus de détails sur les fonctions d'utilités. Ces utilités dépendent de 7 variables d'entrées. Dans [J14], nous avons quantifié l'importance des variables d'entrées à l'aide de l'indice basé sur la distance de Cramér-von Mises puis comparé nos résultats avec ceux obtenus dans [40]. Les classements obtenus sont comparables. En revanche, l'avantage de la méthodologie basée sur les indices de Cramér-von Mises par rapport à celle de Borgonovo *et al.* est que l'on peut utiliser le schéma d'estimation Pick-Freeze (4.47) qui fournit une estimation efficace et simple à mettre en œuvre.

Dans [40], les auteurs étudient un modèle légèrement différent ; ce qui explique les différences numériques entre les résultats donnés dans leur article et ceux de notre étude. En outre, ils effectuent une analyse de sensibilité sur la meilleure alternative ayant la plus grande moyenne au lieu de considérer la sortie multivariée.

4.7 Analyse de sensibilité sur des espaces métriques généraux

Dans cette section, nous généralisons les résultats précédents en considérant que la sortie du code Z est à valeurs dans un espace métrique général $\mathcal{X}$. L'objectif reste le même : déterminer les variables les plus influentes. La construction des nouveaux indices repose sur la constatation suivante : les indices définis précédemment peuvent se réécrire en faisant intervenir des collections de v.a. particulières ($\{\mathbb{1}_{\{Z\leqslant t\}},\, t\in\mathbb{R}^d\}$). En généralisant la procédure à des collections de v.a. quelconques, nous serons en mesure de définir de nouveaux indices englobant les indices définis jusqu'à présent. L'étude menée dans cette section a fait l'objet de la prépublication [P1].

4.7.1 Un nouvel indice

Considérons une collection de v.a. paramétrisées par $m\in\mathbb{N}^*$ élements de $\mathcal{X}$. Pour tout $a=(a_i)_{i=1,\dots,m}\in\mathcal{X}^m$, les v.a.

$$\begin{array}{rcl}\mathcal{X}^m\times\mathcal{X} & \to & \mathbb{R}\\ (a,x) & \mapsto & Y_a(x)\end{array}$$

sont supposées L^2 par rapport à la mesure produit $\mathbb{P}^{\otimes m}\otimes\mathbb{P}$ sur $\mathcal{X}^m\times\mathcal{X}$ où $\mathbb{P}$ est la distribution de Z. Définissons alors les indices de sensibilité sur les espaces métriques généraux par rapport à $\mathbf{u}\subset I_p$ par

$$S^{\mathbf{u}}_{2,GMS}:=\frac{\int_{\mathcal{X}^m}\mathbb{E}\left[\left(\mathbb{E}[Y_a(Z)]-\mathbb{E}[Y_a(Z)|X_{\mathbf{u}}]\right)^2\right]\mathrm{d}\mathbb{P}^{\otimes m}(a)}{\int_{\mathcal{X}^m}\mathrm{Var}(Y_a(Z))\mathrm{d}\mathbb{P}^{\otimes m}(a)}. \tag{4.49}$$

Cas particuliers

1) Pour $\mathcal{X}=\mathbb{R}$, $m=0$ et Y_a donné par $Y_a(x)=x$, on retrouve les indices de Sobol classiques (voir Section 4.2). Tandis que pour $\mathcal{X}=\mathbb{R}^k$ et $m=0$, nous retrouvons la généralisation des indices de Sobol pour les sorties vectorielles (voir Section 4.5) en étendant (4.49) de la façon suivante : la fonction Y_a peut prendre ses valeurs dans $\mathcal{X}=\mathbb{R}^k$ de telle sorte que $Y_a(x)=x$ et

$$S^{\mathbf{u}}_{2,GMS}=\frac{\int_{\mathcal{X}^m}\mathrm{Tr}\left(\mathrm{Var}(\mathbb{E}[Y_a(Z)|X_{\mathbf{u}}])\right)\mathrm{d}\mathbb{P}^{\otimes m}(a)}{\int_{\mathcal{X}^m}\mathrm{Tr}\left(\mathrm{Var}(Y_a(Z))\right)\mathrm{d}\mathbb{P}^{\otimes m}(a)}. \tag{4.50}$$

2) Pour $\mathcal{X}=\mathbb{R}^k$, $m=1$ et Y_a donné par $Y_a(x)=\mathbb{1}_{\{x\leqslant a\}}$, nous retrouvons les indices de Cramér-von Mises définis dans la Section 4.6.

3) Si $\mathcal{X}$ est une variété, $m=2$ et Y_a donné par $Y_a(x)=\mathbb{1}_{\{x\in B(a_1,a_2)\}}$, où $B(a_1,a_2)$ représente la boule de diamètre $\overline{a_1a_2}$, alors nous retrouvons alors les indices définis dans [98].

4.7.2 Procédure d'estimation via des U-statistiques

En vue du schéma Pick-Freeze, définissons $\mathbf{Z}=(Z,Z^{\mathbf{u}})^\top$ et considérons $(m+2)$ copies i.i.d. $(\mathbf{Z}_i,i=1,\dots,m+2)$ de $\mathbf{Z}$. Notons $\mathbb{P}_2^{\mathbf{u}}$ la distribution de $\mathbf{Z}=(Z,Z^{\mathbf{u}})^\top$. Alors le numérateur de (4.49) se réécrit

$$\mathbb{E}_{Z_1,\dots,Z_m}\left[\mathrm{Var}(\mathbb{E}[Y_a(Z_{m+1})|X_{\mathbf{u}}])\right]=\mathbb{E}_{Z_1,\dots,Z_m}\left[\mathrm{Cov}_{\mathbf{Z}_{m+1}}(Y_{Z_1,\dots,Z_m}(Z_{m+1}),Y_{Z_1,\dots,Z_m}(Z^{\mathbf{u}}_{m+1}))\right]. \tag{4.51}$$

La notation $\mathbb{E}_Z$ (resp. Cov_Z) représente l'espérance (resp. la covariance) par rapport à la distribution de Z. Maintenant pour tout $1 \leqslant i \leqslant m+2$, posons $\mathbf{z}_i = (z_i, z_i^{\mathbf{u}})$ et définissons

$$\begin{aligned}
\Phi_1(\mathbf{z}_1, \ldots, \mathbf{z}_{m+1}) &:= Y_{z_1,\ldots,z_m}(z_{m+1}) Y_{z_1,\ldots,z_m}(z_{m+1}^{\mathbf{u}}) \\
\Phi_2(\mathbf{z}_1, \ldots, \mathbf{z}_{m+2}) &:= Y_{z_1,\ldots,z_m}(z_{m+1}) Y_{z_1,\ldots,z_m}(z_{m+2}^{\mathbf{u}}) \\
\Phi_3(\mathbf{z}_1, \ldots, \mathbf{z}_{m+1}) &:= Y_{z_1,\ldots,z_m}(z_{m+1})^2 \\
\Phi_4(\mathbf{z}_1, \ldots, \mathbf{z}_{m+2}) &:= Y_{z_1,\ldots,z_m}(z_{m+1}) Y_{z_1,\ldots,z_m}(z_{m+2}).
\end{aligned}$$

Posons encore

$$m(1) = m(3) = m+1 \quad \text{et} \quad m(2) = m(4) = m+2 \tag{4.52}$$

et définissons pour tout $j = 1, \ldots, 4$,

$$I(\Phi_j) := \int_{\mathcal{X}^{m(j)}} \Phi_j(\mathbf{z}_1, \ldots, \mathbf{z}_{m(j)}) \mathrm{d}\mathbb{P}_2^{v,\otimes m(j)}(\mathbf{z}_1 \ldots, \mathbf{z}_{m(j)}). \tag{4.53}$$

Finalement, introduisons l'application Ψ définie par

$$\begin{array}{rccc} \Psi : & \mathbb{R}^4 & \to & \mathbb{R} \\ & (x,y,z,t) & \mapsto & \frac{x-y}{z-t}. \end{array} \tag{4.54}$$

Ainsi nous pouvons exprimer $S_{2,GMS}^{\mathbf{u}}$ de la façon suivante

$$S_{2,GMS}^{\mathbf{u}} = \Psi\left(I(\Phi_1), I(\Phi_2), I(\Phi_3), I(\Phi_4)\right). \tag{4.55}$$

Au vu de [120], nous remplaçons les fonctions Φ_1, Φ_2, Φ_3 et Φ_4 par leur version symétrisée Φ_1^s, Φ_2^s, Φ_3^s et Φ_4^s :

$$\Phi_j^s(\mathbf{z}_1, \ldots, \mathbf{z}_{m(j)}) = \frac{1}{(m(j))!} \sum_{\tau \in \mathcal{S}_{m(j)}} \Phi_j(\mathbf{z}_{\tau(1)}, \ldots, \mathbf{z}_{\tau(m(j))})$$

pour $j = 1, \ldots, 4$ où $\mathcal{S}_k$ est le groupe symétrique d'ordre k. Pour $j = 1, \ldots 4$, les intégrales $I(\Phi_j^s)$ sont naturellement estimées par les U-statistiques d'ordre $m(j)$. Plus précisément, considérons un N échantillon i.i.d. $(\mathbf{Z}_1, \ldots, \mathbf{Z}_N)$ de loi $\mathbb{P}_2^{\mathbf{u}}$ et pour tout $j = 1, \ldots, 4$, définissons

$$U_{j,N} := \binom{N}{m(j)}^{-1} \sum_{1 \leqslant i_1 < \cdots < i_{m(j)} \leqslant N} \Phi_j^s\left(\mathbf{Z}_{i_1}, \ldots, \mathbf{Z}_{i_{m(j)}}\right). \tag{4.56}$$

Le Théorème 7.1 dans [120] établit que $U_{j,N}$ converge en probabilité vers $I(\Phi_j)$ pour tout $j = 1, \ldots, 4$. En outre, nous pouvons aussi montrer la convergence p.s. en procédant de même que dans la preuve du Lemme 6.1 dans [J14]. Nous estimons alors $S_{2,GMS}^{\mathbf{u}}$ par

$$\widehat{S}_{2,GMS}^{\mathbf{u}} := \frac{U_{1,N} - U_{2,N}}{U_{3,N} - U_{4,N}} = \Psi(U_{1,N}, U_{2,N}, U_{3,N}, U_{4,N}). \tag{4.57}$$

Théorème 4.25. *Si pour $j = 1, \ldots, 4$, $\mathbb{E}\left[\Phi_j^s\left(\mathbf{Z}_1, \ldots, \mathbf{Z}_{m(j)}\right)^2\right] < \infty$ alors*

$$\sqrt{N}\left(\widehat{S}_{2,GMS}^{\boldsymbol{u}} - S_{2,GMS}^{\boldsymbol{u}}\right) \xrightarrow[N\to\infty]{\mathcal{L}} \mathcal{N}_1(0, \sigma^2) \tag{4.58}$$

où la variance asymptotique σ^2 est connue (cf. (22) dans [P1]).

Remarquons que dans la définition (4.55) de $S^{\mathbf{u}}_{2,GMS}$, nous considérons $(m+2)$ copies de $\mathbf{Z}$. Néanmoins, la procédure d'estimation ne nécessite qu'un N échantillon de $\mathbf{Z}$ (voir (4.57)), *i.e.* seulement $2N$ évaluations de la boîte noire ; ce qui constitue un avantage attrayant de cette méthode basée sur les U-statistiques. En outre, le nombre d'appels nécessaire au code est indépendant de la taille m de la collection de v.a. contrairement aux indices de Cramér-von Mises (Section 4.6) ou à ceux introduits dans [98] pour lesquels $(m+2)\times N$ appels au code sont nécessaires.

Commentaires

Pour chaque code de calcul, on peut considérer différents choix de la famille $(Y_a)_{a\in\mathcal{X}^m}$ de v.a. indexées par $a\in\mathcal{X}^m$ conduisant à des indices très différents. C'est à l'utilisateur de faire son choix en fonction de son objectif. Pour quantifier la sensibilité de la sortie autour de la moyenne, il faut considérer les indices Sobol classiques basés sur la variance et correspondant au cas particulier précédent 1). Si l'on s'intéresse à la la sensibilité de la distribution, il convient de prendre une famille de fonctions test qui caractérisent la distribution. Par exemple, dans le cas 3. précédent, les fonctions Y_a sont les fonctions indicatrices des demi-droites ou des quadrants en dimension plus grande et correspondent aux indices de Cramér-von Mises. De plus, puisque dans la procédure d'estimation le nombre d'appels au code est indépendant du choix de la famille $(Y_a)_{a\in\mathcal{X}^m}$, on peut considérer et estimer simultanément plusieurs indices sans coût additionnel. En fait, le seul enjeu repose sur notre capacité à évaluer les fonctions Φ aux points d'observation.

Cas particuliers

1) Pour $\mathcal{X}=\mathbb{R}$, $m=0$ et Y_a donné par $Y_a(x)=x$, l'estimateur défini dans (4.57) est basé sur les U-statistiques $U_{j,N}$ pour $j=1,\dots,4$ données par

$$U_{1,N}=\frac{1}{N}\sum_{i=1}^N Z_iZ_i^{\mathbf{u}},\qquad U_{2,N}=\frac{1}{N(N-1)}\left(\sum_{i=1}^N Z_i\sum_{i=1}^N Z_i^{\mathbf{u}}-\sum_{i=1}^N Z_iZ_i^{\mathbf{u}}\right)=:U^1_{2,N}-U^2_{2,N},$$
$$U_{3,N}=\frac{1}{N}\sum_{i=1}^N Z_i^2,\qquad U_{4,N}=\frac{1}{N(N-1)}\left(\left(\sum_{i=1}^N Z_i\right)^2-\sum_{i=1}^N Z_i^2\right)=:U^1_{4,N}-U^2_{4,N}$$

conduisant à

$$\widehat{S}^{\mathbf{u}}_{2,GMS}=\frac{U_{1,N}-U_{2,N}}{U_{3,N}-U_{4,N}}=\Psi(U_{1,N},U_{2,N},U_{3,N},U_{4,N})$$

tandis que dans la Section 4.2, l'estimateur $\widehat{S}^{\mathbf{u}}$ de $S^{\mathbf{u}}_{2,GMS}$ s'écrit

$$\widehat{S}^{\mathbf{u}}=\frac{U_{1,N}-U^1_{2,N}}{U_{3,N}-U^1_{4,N}}=\Psi(U_{1,N},U^1_{2,N},U_{3,N},U^1_{4,N}) \tag{4.59}$$

et prend en compte les termes diagonaux. Les deux procédures requièrent $2N$ évaluations du code et ont la même vitesse de convergence (seules les variances asymptotiques diffèrent). Bien sûr, lorsque le code est centré les procédures sont les mêmes. Remarquons que nous pouvons améliorer la procédure en prenant en compte toute l'information contenue dans l'échantillon et conduisant à l'analogue de $\widehat{T}^{\mathbf{u}}$ défini dans (4.9). Rappelons que la suite d'estimateurs $\widehat{T}^{\mathbf{u}}$ est asymptotiquement efficace (voir Proposition 4.4). Cependant, basés sur le même plan d'expérience que $\widehat{S}^{\mathbf{u}}$ et $\widehat{T}^{\mathbf{u}}$, ni $\widehat{S}^{\mathbf{u}}_{2,GMS}$ ni ce nouvel estimateur $\widehat{S}^{\mathbf{u}}_{2,GMS}$ ne pourrait donc être asymptotiquement efficace. En revanche, la procédure d'estimation basée sur les U-statistiques surpasse celle basée sur les indices

de Cramér-von Mises dès que $m \geqslant 1$. Pour $\mathcal{X} = \mathbb{R}^k$ et $m = 0$, la même analogie peut être faite.

2) Pour $\mathcal{X} = \mathbb{R}^k$, $m = 1$ et Y_a donné par $Y_a(x) = \mathbb{1}_{\{x \leqslant a\}}$, nous surpassons le Théorème 4.24. En effet, l'estimateur $\widehat{S}^{\mathbf{u}}_{2,GMS}$ nécessite $3N$ évaluations du code tandis que dans cette section $2N$ appels suffisent. Enfin, la preuve du Théorème 4.24 repose sur la puissante mais complexe Delta méthode fonctionnelle tandis que la preuve du Théorème 4.25 est une application élémentaire du Théorème 7.1 dans [120] combinée à la Delta méthode classique.

4.7.3 Codes stochastiques

Contexte général

Dans certaines applications, nous sommes face à des codes stochastiques en ce sens que deux évaluations du code pour la même entrée x mènent à des sorties différentes. L'utilisateur s'intéresse alors à la distribution μ_x de la sortie pour une entrée x fixée. Nous pouvons nous ramener à un code déterministe en considérant une entrée additionnelle qui n'est pas choisie par l'utilisateur mais qui est une variable latente générée au hasard par le code à chaque évaluation. Nous construisons toutes les variables aléatoires (celle choisie par le praticien et celle générée par le code) sur le même espace de probabilité et nous considérons l'application :

$$\begin{aligned} f_s \ : \ E \times G \ &\to \mathbb{R} \\ (x, v) \ &\mapsto f_s(x, v). \end{aligned} \tag{4.60}$$

Nous noterons la v.a. $f_s(x, \cdot)$ simplement par $f_s(x)$. Ainsi, nous définissons une autre code (deterministe) associé à f_s dont la sortie est une mesure de probabilité :

$$\begin{aligned} f \ : \ E \ &\to \mathcal{M}_2(\mathbb{R}) \\ x \ &\mapsto \mu_x \end{aligned} \tag{4.61}$$

où $\mathcal{M}_2(\mathbb{R})$ est l'ensemble des mesures de probabilités μ telles que $\int x^2 \mu(\mathrm{d}x) < +\infty$. Evidemment en pratique, nous n'avons pas directement accès à f mais il est possible d'obtenir une approximation naturelle de la mesure μ_x à partir de n évaluations de f_s au point x :

$$\mu_{x,n} := \frac{1}{n} \sum_{j=1}^{n} \delta_{f_s(x, v_j)}.$$

Concrètement, pour une entrée $X \in E$ dont la distribution est notée $\mathcal{L}$, nous évaluons n fois le code f_s défini par (4.60) de telle sorte que le code va générer n variables $V_1, \ldots, V_n$ et que nous observons

$$f_s(X, V_1), \ldots, f_s(X, V_n)$$

conduisant à la mesure aléatoire $\mu_{X,n} = \frac{1}{n} \sum_{j=1}^{n} \delta_{f_s(X, V_j)}$ qui approxime la distribution de $f_s(X)$. Notons que les v.a. $V_1, \ldots, V_n$ ne sont pas observées.

Analyse de sensibilité

Afin d'étudier la sensibilité de μ_x, nous pouvons utiliser la procédure de la Section 4.7.1. A cet effet, nous munissons $\mathcal{M}_2(\mathbb{R})$ de la distance de Wasserstein W_2 d'ordre 2 (bien sûr, nous pouvons aussi considérer

$\mathcal{M}_p(\mathbb{R})$ et W_p au lieu de $\mathcal{M}_2(\mathbb{R})$ et W_2). Alors (4.49) devient

$$S^{\mathbf{u}}_{2,GMS} = \frac{\int_{\mathcal{X}^m} \mathbb{E}\left[\left(\mathbb{E}[\mathbb{1}_{W_2(\mu_1,\mu_X)\leqslant W_2(\mu_1,\mu_2)}] - \mathbb{E}[\mathbb{1}_{W_2(\mu_1,\mu_X)\leqslant W_2(\mu_1,\mu_2)}|X_{\mathbf{u}}]\right)^2\right] \mathrm{d}\mathbb{P}^{\otimes 2}(\mu_1,\mu_2)}{\int_{\mathcal{X}^m} \mathrm{Var}(\mathbb{1}_{W_2(\mu_1,\mu_X)\leqslant W_2(\mu_1,\mu_2)})\mathrm{d}\mathbb{P}^{\otimes 2}(\mu_1,\mu_2)}$$

ou encore $S^{\mathbf{u}}_{2,GMS} = \Psi\left(I(\Phi_1), I(\Phi_2), I(\Phi_3), I(\Phi_4)\right)$ avec Ψ et I définis par (4.54) et (4.53) et

$$\begin{aligned}
\Phi_1(\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_3) &= \mathbb{1}_{W_2(\mu_1,\mu_3)\leqslant W_2(\mu_1,\mu_2)}\mathbb{1}_{W_2(\mu_1,\mu_3^{\mathbf{u}})\leqslant W_2(\mu_1,\mu_2)}\\
\Phi_2(\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_4) &= \mathbb{1}_{W_2(\mu_1,\mu_3)\leqslant W_2(\mu_1,\mu_2)}\mathbb{1}_{W_2(\mu_1,\mu_4^{\mathbf{u}})\leqslant W_2(\mu_1,\mu_2)}\\
\Phi_3(\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_3) &= \mathbb{1}_{W_2(\mu_1,\mu_3)\leqslant W_2(\mu_1,\mu_2)}\\
\Phi_4(\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_4) &= \mathbb{1}_{W_2(\mu_1,\mu_3)\leqslant W_2(\mu_1,\mu_2)}\mathbb{1}_{W_2(\mu_1,\mu_4)\leqslant W_2(\mu_1,\mu_2)}
\end{aligned}$$

où $\boldsymbol{\mu}_i = \boldsymbol{\mu}_{X_i}$ est la concaténation de la mesure μ_{X_i}et de sa version Pick-Freeze notée $\mu_{X_i^{\mathbf{u}}}$.

Estimation des indices

Dans un scénario idéal correspondant à (4.61), nous pouvons observer μ_x pour tout x. Alors grâce à la procédure de la Section 4.7.2, nous obtenons une estimation de l'indice $S^{\mathbf{u}}_{2,GMS}$ dont le comprtement asymptotique est donné par le Théorème 4.25.

Dans le contexte plus réaliste de (4.60), nous avons seulement accès à l'approximation $\mu_{x,n}$ de μ_x rendant plus complexe la procédure d'estimation et l'étude des propriétés asymptotiques. Dans ce cas, le plan d'expérience est le suivant :

$$\begin{aligned}
(X_1, V_{1,1},\ldots,V_{1,n}) &\to f_s(X_1,V_{1,1}),\ldots,f_s(X_1,V_{1,n})\\
(X_1^{\mathbf{u}}, V'_{1,1},\ldots,V'_{1,n}) &\to f_s(X_1^{\mathbf{u}},V'_{1,1}),\ldots,f_s(X_1^{\mathbf{u}},V'_{1,n})\\
&\vdots\\
(X_N, V_{N,1},\ldots,V_{N,n}) &\to f_s(X_N,V_{N,1}),\ldots,f_s(X_N,V_{N,n})\\
(X_N^{\mathbf{u}}, V'_{N,1},\ldots,V'_{N,n}) &\to f_s(X_N^{\mathbf{u}},V'_{N,1}),\ldots,f_s(X_N^{\mathbf{u}},V'_{N,n})
\end{aligned}$$

Le nombre total d'appels au code stochastique (4.60) est $2\times N\times n$. L'approximation empirique $\mu_{i,n} = \mu_{X_i,n}$ de μ_i est donnée par

$$\mu_{i,n} = \frac{1}{n}\sum_{j=1}^{n}\delta_{f_s(X_i,V_{i,j})},$$

pour tout $i = 1,\ldots,N$. Maintenant, pour $j = 1,\ldots,4$, posons

$$U_{j,N,n} := \binom{N}{m(j)}^{-1} \sum_{1\leqslant i_1<\cdots<i_{m(j)}\leqslant N} \Phi_j^{\mathbf{s}}\left(\boldsymbol{\mu}_{i_1,n},\ldots,\boldsymbol{\mu}_{i_{m(j)},n}\right) \tag{4.62}$$

où $\Phi_j^{\mathbf{s}}$ est la version symétrisée de Φ_j, pour $j = 1,\ldots,4$. Finalement, nous estimons $S^{\mathbf{u}}_{2,GMS}$ par

$$\widehat{S}^{\mathbf{u}}_{2,GMS,n} := \frac{U_{1,N,n}-U_{2,N,n}}{U_{3,N,n}-U_{4,N,n}} = \Psi(U_{1,N,n},U_{2,N,n},U_{3,N,n},U_{4,N,n}). \tag{4.63}$$

Remarque 4.26. *1) L'estimateur* (4.62) *est facile à calculer puisque pour deux mesures discrètes supportées par un même nombre de points données par*

$$\nu_1 = \frac{1}{n}\sum_{k=1}^{n}\delta_{x_k},\ \nu_2 = \frac{1}{n}\sum_{k=1}^{n}\delta_{y_k},$$

la distance de Wasserstein d'ordre 2 entre ν_1 et ν_2 s'écrit simplement

$$W_2^2(\nu_1, \nu_2) = \frac{1}{n}\sum_{j=1}^{n}(x_{(j)} - y_{(j)})^2,$$

où $x_{(j)}$ est la j-ème statistique d'ordre de x. De plus, cette remarque est encore vraie si on remplace W_2 par W_p (en remplaçant l'exposant 2 par p). Ainsi, il est possible de calculer simultanément plusieurs indices basés sur une collection finie de distances de Wasserstein.

2) *Dans [45], [159] et [184], les auteurs considèrent des codes stochastiques à sortie densité. En d'autres termes, ils définissent l'application suivante :*

$$\begin{aligned} f \ : \ E &\to \mathcal{F} \\ x &\mapsto f(x) \end{aligned} \tag{4.64}$$

où $\mathcal{F}$ est l'ensemble des densités de probabilité :

$$\mathcal{F} := \left\{ g \in L^1(\mathbb{R});\ g \geqslant 0,\ \int_{\mathbb{R}} g(x)\mathrm{d}x = 1 \right\}.$$

Proposition 4.27. *Considérons trois copies i.i.d. X_1, X_2 et X_3 de X. Soit $\delta(N)$ une suite tendant vers 0 lorsque N tend vers l'infini et telle que*

$$\mathbb{P}\left(|W_2(\mu_{X_1}, \mu_{X_3}) - W_2(\mu_{X_1}, \mu_{X_2})| \leqslant \delta(N)\right) = o\left(\frac{1}{\sqrt{N}}\right).$$

Soit n tel que $\mathbb{E}[W_2(\mu_X, \mu_{X,n})] = o(\delta(N)/\sqrt{N})$. Sous les hypothèses du Théorème 4.25, nous avons

$$\sqrt{N}\left(\widehat{S}^{\boldsymbol{u}}_{2,GMS,n} - S^{\boldsymbol{u}}_{2,GMS}\right) \xrightarrow[N\to\infty]{\mathcal{L}} \mathcal{N}_1(0, \sigma^2) \tag{4.65}$$

où la variance asymptotique σ^2 est connue (cf. (22) dans [1]).

Choix pratiques de $\delta(N)$. Dans des cas particuliers, il est possible de déterminer une valeur adaptée de $\delta(N)$. Nous considérons deux exemples dans la suite.

— Si l'inverse de la v.a. $W := |W_2(\mu_{X_1}, \mu_{X_3}) - W_2(\mu_{X_1}, \mu_{X_2})|$ admet un moment d'ordre 1, alors par l'inégalité de Markov,

$$\mathbb{P}\left(W \leqslant \delta(N)\right) = \mathbb{P}\left(W^{-1} \geqslant \delta(N)^{-1}\right) \leqslant \frac{1}{\delta(N)}\mathbb{E}\left[\frac{1}{W}\right]$$

et il suffit de prendre $\delta(N)$ tel que $\delta(N)^{-1} = o\left(N^{-1/2}\right)$ lorsque N tend vers l'infini.

— Supposons que X est uniformément distribué sur $[0, 1]$ et que μ_X est la distribution gaussienne centrée en X et réduite. Alors la distance de Wasserstein $W_2(\mu_{X_1}, \mu_{X_2})$ s'écrit $(X_1 - X_2)^2$ et la v.a. $W = |W_2(\mu_{X_1}, \mu_{X_3}) - W_2(\mu_{X_1}, \mu_{X_2})|$ est donc donnée par

$$\left|(X_1 - X_3)^2 - (X_1 - X_2)^2\right| = \left|(X_3 - X_2)(X_2 + X_3 - 2X_1)\right|.$$

Par conséquent,

$$\mathbb{P}(W \leqslant \delta(N)) \leqslant \mathbb{P}(|X_3 - X_2| \leqslant \sqrt{\delta(N)}) + \mathbb{P}(|X_2 + X_3 - 2X_1| \leqslant \sqrt{\delta(N)}).$$

Notons que $(X_2 + X_3)/2$ et X_1 sont deux v.a. indépendantes uniformes sur $[0, 1]$. Ainsi il reste à calculer $\mathbb{P}(|U_1 - U_2| \leqslant \alpha)$ pour U_1 et U_2 indépendantes et uniformes sur $[0, 1]$ pour $\alpha = \sqrt{\delta(N)}$ et $\alpha = \sqrt{\delta(N)}/2$. On montre que

$$\mathbb{P}(|U_1 - U_2| \leqslant \alpha) = \alpha(2 - \alpha)$$

conduisant à $\mathbb{P}(W \leqslant \delta(N)) = O\left(\sqrt{\delta(N)}\right)$. Aussi, un choix convenable est $\delta(N) = o(1/N)$.

Choix pratiques de n. De façon analogue, on peut facilement obtenir des choix appropriés de la valeur de n dans certains cas particuliers. Par exemple, nous renvoyons le lecteur à [35] pour obtenir des limites supérieures sur $\mathbb{E}[W_p(\mu_X, \mu_{X,n})]$ pour plusieurs valeurs de $p > 1$ et plusieurs hypothèses sur la distribution sur μ_X : cas général, distribution uniforme, distribution gaussienne, distribution beta, distribution log concave...

Exemple 4.28. *Considérons l'exemple précédent pour lequel X suit la loi uniforme sur $[0, 1]$ et μ_X est la distribution gaussienne centré en X et réduite. Alors par [35, Corollary 6.14], nous avons pour tout $n \geqslant 3$,*

$$\mathbb{E}[W_2(\mu_X, \mu_{X,n})^2] \leqslant \frac{(Const) \log\log n}{n}.$$

et pour tout $p > 2$ et tout $n \geqslant 3$,

$$\mathbb{E}[W_p(\mu_X, \mu_{X,n})^p] \leqslant \frac{C_p}{n(\log n)^{p/2}},$$

où C_p dépend de p uniquement. Puisque nous avons déjà choisi $\delta(N) = o(N^{-1})$, il suffit de prendre $\log\log n/n = o(N^{-2})$pour satisfaire la condition $\mathbb{E}[W_2(\mu_X, \mu_{X,n})] = o(\delta(N)/\sqrt{N})$.

Chapitre 5

Activités de recherche en lien avec l'industrie

Dans ce chapitre, je présente très succinctement quelques résultats obtenus lors de mon Post-Doctorat sur un contrat IMT-EDF (Section 5.1), d'un contrat Thales-CNES (Section 5.2), du projet Idex MetaDEB (Section 5.3) et enfin de l'ANR PEPITO (Plan d'Expérience Pour l'Industrie du Transport et l'Optimisation - Section 5.4).

5.1 Prédiction de courbes de charge

Dans le cadre d'un contrat avec EDF, j'ai travaillé pendant mon Post-Doctorat (encadré par Jean-Marc Azaïs (IMT) et Jean-Claude Fort (Paris V)) sur les bandes de confiance simultanées pour la prévision de courbes. Pour relever le défi de l'ouverture des marchés de l'énergie, mieux connaître les clients et leur façon de consommer l'électricité dans le temps est une nécessité pour EDF. Cette connaissance s'acquiert par l'analyse des courbes de charge qui donnent l'évolution de la puissance appelée par les clients. cependat, ces courbes ne sont disponibles que pour un nombre restreint de clients. Pour les autres, il s'agit de les prédire. Pour cela, nous disposions de courbes de charge d'entreprises ainsi que des variables explicatives pour chacune des entreprises. En déterminant le maximum de la valeur absolue de processus gaussiens, nous avons déduit des intervalles de confiance simultanés. Ce travail a donné lieu à la publication [J4].

5.2 Méthode d'événements rares pour l'intégrité et la continuité EGNOS - GALILEO

Dans le cadre d'un contrat avec le CNES, j'ai travaillé en partenariat avec Thalès sur les erreurs de positionnement des systèmes de navigation. Plus précisément, il s'agissait d'estimer le plus finement possible la probabilité de non-intégrité du système de navigation EGNOS. Cette non-intégrité, rare, est obtenue comme la conjonction d'événements redoutés qui peuvent être d'origines diverses et ont une certaine durée. Jusque là, la non-intégrité était calculée sans tenir compte de la concomitance éventuelle entre événements. Nous avons utilisé des outils classiques de modélisation statistique tels que la fiabilité et les arbres de défaillance ainsi que les réseaux de Pétri et l'outil MissRdP utilisés par Thalès combinés à des algorithmes de renforcement afin de pallier la rareté de ces événements et proposer des estimées fiables des probabilités de concomitances double voire triple. L'impact de ces concomitances s'est révélé

être non négligeable d'où la nécessité pour le CNES d'en tenir compte pour le calcul de la non-intégrité. Ce travail en collaboration avec Jean-Marc Azaïs (IMT), Sébastien Gadat (UT1) et Cécile Mercadier (Institut Camille Jordan Lyon 1) a donné lieu à un rapport interne [R1].

5.3 Méthodes statistiques en halieutique

Ce travail est le résultat d'une coopération entre chercheurs académiques et industriels (Sébastien da Veiga (Safran), Nicolas Bousquet (EDF), Bertrand Iooss (EDF)) et scientifique de la pêche (Emmanuel Chassot (IRD)). Ici l'utilisation d'outils statistiques (tels que l'analyse de sensibilité, les copules, la calibration, les statistiques bayésiennes, la classification ABC...) a conduit à des résultats pertinents et prometteurs. Nous nous sommes intéressés à une représentation probabiliste de la croissance des thons Yellowfin de l'Océan Indien basée sur les modèles de budget énergétique dynamique (Dynamic Energy Budget models - DEB models).

L'estimation des paramètres du modèle bioénergétique et de leur incertitude correspondante est cruciale : dans la pratique, beaucoup d'entre eux sont fixés à une valeur nominale par facilité ou par manque d'information. D'autre part, le fait de lier différents aspects de la croissance et de la reproduction dans un modèle qui explique comment l'énergie est utilisée et attribuée est un grand défi dans les études écologiques. Le but était de comprendre et d'expliquer les compromis entre la survie, la croissance et la reproduction qui caractérisent la condition physique d'une population et sa stratégie générale pour faire face à un environnement (par exemple, l'impact de l'exploitation humaine). En outre, cela fournit des informations essentielles pour établir des règles d'exploitation durable.

Pour ce faire, nous avons d'abord procédé à une analyse de sensibilité pour déterminer les paramètres d'entrée les plus influents sur la forme de la courbe de croissance et la vitesse de croissance puis avons éliminé les paramètres d'entrée non importants. Dans un deuxième temps, nous avons exploité les données observées pour calibrer la distribution des paramètres d'entrée (conditions biologiques et environnementales). Ensuite, une nouvelle analyse de sensibilité a été effectuée sur cette distribution conjointe a posteriori. Enfin, nous avons procédé à la classification pour définir les domaines de paramètres d'entrée conduisant à des courbes bivariées avec un nombre donné de paliers. Ce travail a donné lieu à la prépublication [P2].

5.4 Optimisation de ventilateurs pour l'industrie automobile

Ces travaux sont le fruit d'une coopération fructueuse entre des chercheurs universitaires de l'Institut de Mathématiques de Toulouse entre autres et e partenaire industriel Valeo. Ils ont été réalisés dans le cadre du projet PEPITO (Plan d'Expérience Pour l'Industrie du Transport et l'Optimisation) soutenu par l'Agence de recherche (ANR). Les objectifs étaient d'expérimenter une approche extrême basée sur des simulations intensives et multiphysiques, l'utilisation de géométries paramétrisées, la détermination des conceptions d'expériences avec un grand nombre de facteurs et la recherche d'optima dans les domaines de grande dimension.

Le contexte industriel est le suivant. Dans l'industrie automobile, les besoins des clients évoluent rapidement dans un contexte de compétitivité. En particulier, le ventilateur impliqué dans le module de refroidissement du moteur joue un rôle clef dans l'architecture du moteur et reste un objectif majeur d'ingénierie. Dans ces conditions, les ingénieurs sont alors contraints de proposer dans des délais très courts de nouveaux modèles de ventilateurs répondant aux exigences du client en termes d'efficacité, couple, acoustique, encombrement... Malheureusement, un tel objectif est long et coûteux à atteindre. Dans cette optique, les approches statistiques sont incontournables pour proposer des métamodèles, accélérer les

simulations, rechercher des optima dans un délai raisonnable.

Plus précisément, nous avons utilisé l'interpolation par krigeage et l'algorithme d'optimisation d'Expected Improvement (EI) pour déterminer de nouveaux modèles de ventilateurs à hautes performances. Initialement introduit en géostatistique par [151], le krigeage est une méthode stochastique d'interpolation. L'objectif est de prédire la valeur d'un phénomène naturel f (analytiquement inconnu) en tout point arbitraire à partir des observations mesurées aux points d'échantillonnage. L'ingrédient clé du krigeage repose sur l'hypothèse que f est la réalisation d'un processus gaussien. L'EI introduit par [137] est l'un des algorithmes d'optimisation bayésiens les plus largement utilisés. Le but est de maximiser la fonction objectif f et pour ce faire, de proposer séquentiellement des nouveaux points où évaluer la fonction en espérant se rapprocher de son maximum. La séquence des décisions (pour choisir le nouveau point x où évaluer f) est guidée par un critère [181] qui mesure l'amélioration apportée par le point x dans la maximisation de f. L'algorithme choisit soit d'explorer des zones prometteuses (amélioration), soit des zones méconnues (exploration). À notre connaissance, une telle utilisation des méthodes de krigeage et d'EI est innovante et donne des résultats très prometteurs. Cet axe de recherche développé en collaboration avec Thi Mong Ngoc Nguyen (Université des Sciences d'Ho Chi Minh, Vietnam) et Bruno Demory et Manuel Henner, ingénieurs Valeo, a donné lieu à la prépublication suivante [S2].

Dans un second temps, afin de ne pas être limité dans le nombre d'appels au code, nous avons mis à profit le métamodèle dénomé TurboConcept™ et développé par le Laboratoire de Mécanique des Fluides et d'Acoustique de Lyon (LMFA), autre acteur de l'ANR. Nous avons ensuite comparé différentes stratégies d'optimisation (toutes basées sur une étape préliminaire de krigeage) : l'optimisation directe, une optimisation séquentielle basée sur l'algorithme d'EI, une optimisation séquentielle basée sur l'algorithme d'Upper Bound Confidence (UCB) et une dernière, séquentielle aussi, combinaison des deux précédentes. L'algorithme d'optimisation UCB issu de la théorie des jeux [14] ne cherche pas une amélioration comme l'EI, mais plutôt une stratégie optimiste. Plus précisément, il maximise séquentiellement un quantile de krigeage bien choisi. Les résultats obtenus avec Martin Buisson (LMFA), Thi Mong Ngoc Nguyen (Université des Sciences d'Ho Chi Minh, Vietnam) et Mélina Ribaud (doctorante à l'Université Lyon 1) conduisent à des géométries de ventilateurs prometteuses et sont présentés dans la prépublication [P3].

Chapitre 6

Travaux en cours et perspectives

6.1 Processus gaussiens profonds

Comme nous l'avons vu précédemment, les processus gaussiens (PG) permettent de modéliser des fonctions par une approche probabiliste, flexible et non paramétrique. En outre, leurs bonnes propriétés rendent une étude analytique possible. Cependant, leur maniabilité se fait à un certain prix : ils ne peuvent représenter qu'une classe restreinte de fonctions. En effet, même si des définitions sophistiquées et des combinaisons de fonctions de covariance ont été introduites [210], il n'en reste pas moins que le modèle fait l'hypothèse que les données sont gaussiennes. Une première généralisation classique consiste à considérer des PG bruités de la forme $y = Y(x) + \varepsilon$ où ε est un bruit blanc gaussien. Ainsi la loi de y conditionnée à x est une gaussienne. Une généralisation moins classique consiste à intégrer le bruit directement dans le PG induisant le processus suivant : $y = Y(x+\varepsilon)$. La loi de y conditionnée à x est une fonction d'une gaussienne sans pour autant être une gaussienne. Cependant, il est souvent déraisonnable de supposer de la gaussianité dans les modèles. Par exemple, les observations peuvent être positives et varier sur plusieurs ordres de grandeur. Il n'est donc pas adapté de modéliser directement ces quantités par un PG. En revanche, dans ce cas, il est courant et judicieux en pratique de prendre le logarithme des observations. Il convient ensuite de modéliser les données transformées par un PG [233]. Le logarithme est seulement une transformation parmi d'autres qui peut être appliquée. Un espace latent est donc ainsi introduit dans lequel les observations sont bien modélisées par un GP. Ceci permet donc une généralisation des PG puisque dans l'espace d'observation le processus est non gaussien avec un bruit non gaussien et asymétrique en général.

Dans cette optique de généralisation, une piste prometteuse est de considérer un cas particulier de composition conduisant aux processus gaussiens imbriqués, appelés *processus gaussiens profonds* ("Deep Gaussien processes"). Le terme profond est hérité du lien entre ces mdèles et les réseaux de neurones profonds. Dans les PG profonds [78, 79], les données sont modélisées par un PG multivarié dont les entrées sont elles-mêmes gouvernées par un autre PG : $y = Y(Z(x))$ où Y et Z sont deux PG. En itérant cette procédure, nous formons ainsi plusieurs couches. Le processus résultant n'est plus un PG comme espéré mais son analyse devient bien plus complexe. Ce faisant, la classe de modèles se trouve enrichie tant dans les distributions disponibles que dans l'irrégularité des processus. En effet, par la formule de dérivation en chaîne, les dérivées des PG sous-jacents sont multipliées et la dérivée du processus global peut donc être très grande entrainant ainsi plus d'irrégularités pour le processus global.

Malgré leur essor notable et leur succès pratique dans les domaines de l'intelligence artificielle et de la science des données, les méthodes basées sur les PG profonds n'ont quasiment aucune garantie mathématique du type de celles obtenues dans le cadre des PG classiques (voir par exemple [173, 228] pour un

contrôle de l'erreur). Les développements récents tendent à rendre le pouvoir prédictif des PG similaire à celui des réseaux de neurones profonds, avec l'avantage que les PG reposent sur un modèle bayésien explicite qui permet une plus grande fiabilité et interprétabilité.
Avec François Bachoc (IMT), nous souhaitons obtenir de telles garanties et en particulier des résultats de convergence (consistance et vitesse de convergence) pour les prédictions effectuées par processus gaussiens profonds. Plus précisément, nous souhaitons établir la vitesse de convergence de l'espérance conditionnelle d'un PG profond vers la fonction inconnue boîte-noire sous-jacente. Nous étendrons ensuite cette analyse asymptotique à l'estimation de lois conditionnelles dépendant de variables d'entrée, toujours en s'appuyant sur les PG profonds. D'un point de vue numérique, nous souhaitons développer des codes afin d'illustrer les résultats théoriques obtenus sur des données réelles et simulées. Il est espéré que l'analyse théorique conduise à des suggestions d'amélioration pratique des méthodes existantes.

6.2 Théorèmes limites et cascades de Mandelbrot

Sur les théorèmes limites de sommes de v.a.
Tout d'abord, comme nous l'avons déjà dit, le catalogue présenté dans le Théorème 3.21 concernant le comportement asymptotique du déplacement total pour les tables creuses dans le cadre du hachage n'est pas encore complet. En effet, il reste à établir les résultats limites pour les cas intermédiaires $\alpha = 2/3$ et $\alpha = 2$. Le cas $\alpha = 2/3$ est l'équivalent pour les sommes conditionnées du cas intermédiaire et difficile (ii) dans le Théorème 3.7. Quant à $\alpha = 2$, il est facile de montrer que

$$\mathbb{P}(d_{m_n,n} - \mathbb{E}[d_{m_n,n}] \geqslant N_n^2 x) = \Theta\left(e^{-(Const)N_n \log N_n}\right)$$

et il reste à déterminer la constante.

Dans la suite de [43], une autre perspective est d'étendre les résultats établis dans le Théorème 3.7 aux queues plus générales de la forme

$$\log \mathbb{P}(X \geqslant x) \sim -\ell(x)x^{1-\varepsilon} \tag{6.1}$$

où ε est une constante arbitraire et ℓ est une fonction à variations lentes ; c'est ce que nous nous proposons de faire avec Franck Barthe (IMT) et Fabien Brosset le doctorant que nous co-encadrons. Cette généralisation semble ne pas poser de problème la plupart du temps excepté dans le cas intermédiaire, où les études de variations faites vont se trouver perturbées. Ces résultats devraient permettre de dégager des principes généraux sur la correspondance entre la loi de la v.a. X, la vitesse et le taux des PGD établis.
Une extension aux vecteurs aléatoires pourra être envisagée ainsi qu'une comparaison avec les résultats de concentration non-asymptotiques, obtenus pour les v.a. non exponentiellement intégrables établis par Barthe, Cattiaux et Roberto [25].
Nous pensons aussi utiliser une autre approche, beaucoup plus proche des preuves modernes du théorème de Cramér (Chebychev exponentiel pour la borne supérieure). Cette approche pourrait être mise en œuvre pour les lois de probabilité ayant des moments (polynomiaux) de tous ordres. Le point de départ serait d'adapter un outil introduit par Latała dans [156], avec lequel il parvient à donner une expression courte des normes L_p de $X_1 + \ldots + X_n$ qui est exacte à un facteur près. Cet outil est une transformée "de type moment" qui est sous-multiplicative et pourrait se substituer à la transformée de Laplace.
Enfin, le travail précédent, qui aborde des questions très fondamentales de la théorie des probabilités, pourrait être complété par des applications des techniques de grandes déviations. Ceci permettrait aussi

de balayer un champ thématique plus large. Parmi les nombreuses pistes possibles, on pourrait étudier des systèmes de spins avec des lois sous-jacentes à queues lourdes. Un autre modèle intéressant venant de la géométrie stochastique, consiste à étudier les formes asymptotitiques de l'enveloppe de points aléatoires du plan, conditionnés à être en position convexe. D'autres applications, dans la lignée des travaux présentés dans la Section 3.4 et concernant le hachage sont aussi possibles.

Sur les cascades de Mandelbrot

Enfin, avec Pierre Petit (IMT) et Thierry Klein (IMT-ENAC), nous avons commencé à travailler sur une question posée par Alain Rouault (Université Paris-Saclay). Précisons le cadre de ce travail.
Soit W une v.a. positive telle que $\mathbb{E}[W] = 1$. Soit $(W_{i_1,\dots,i_n})_{i_1,\dots,i_n\in\mathbb{N}^*}$ une famille de v.a. i.i.d. distribuées comme W indéxée par toutes les suites $(i_1,\dots,i_n)$, $n \geqslant 1$ d'entiers positifs. Nous souhaitons déterminer les grandes déviations de la suite définie pour tout $n \geqslant 1$ par

$$Z_r^n := \frac{1}{r^n} \sum_{1\leqslant i_1,\dots,i_n\leqslant r} W_{i_1} W_{i_1,i_2} \dots W_{i_1,\dots,i_n} .$$

Soit $\Lambda(t) = \log \mathbb{E}[e^{tW}] \in]-\infty,+\infty]$ la log-Laplace de W et soit

$$\Lambda^*(x) = \sup_{t\geqslant 0}\{tx - \Lambda(t)\}$$

la transformée de Fenchel-Legendre associée. Supposons qu'il existe une constante $c \in [0\,,+\infty]$ telle que

$$\frac{1}{x}\log \mathbb{P}(W \geqslant x) \to -c$$

lorsque $x \to +\infty$. Le cas $c = +\infty$ correspond au cas où $\Lambda < +\infty$ partout. Le cas $c \in]0\,,+\infty[$ inclut le cas de la loi exponentielle de moyenne 1, les lois gamma de moyenne 1...
Nous avons montré le résultat suivant.

Théorème 6.1. *Pour tout $n \geqslant 1$, la suite $(Z_r^n)_{r\geqslant 1}$ satisfait un PGD à la vitesse r de fonction de taux I_n définie par récurrence par $I_1 = \Lambda^*$ et, pour tout $a \in \mathbb{R}$,*

$$I_{n+1}(a) = \inf\left\{cw + I_n(z) + \Lambda^*(s)\ ;\ w \geqslant 0,\ z \geqslant 0,\ s \geqslant 0,\ wz + s = a\right\}.$$

En outre, pour tout $n \geqslant 1$ et $a \leqslant 1$, $I_n(a) = \Lambda^*(a)$.

Proposition 6.2.

(i) Pour tout $t > 0$, $\Lambda(t) < +\infty$ et alors pour tout $n \geqslant 1$, $I_n = \Lambda^*$.

(ii) Pour tout $t > 0$, $\Lambda(t) = +\infty$ et alors pour tout $n \geqslant 1$, $I_n = 0$.

Remarquons que le point (i) est un PGD à la vitesse r prouvé dans la Proposition 1.2 de [217]. Lorsque $\bar{w} = \operatorname{essup} W < +\infty$, $\Lambda^*(x) = +\infty$ pour tout $x > \bar{w}$: de sorte que lePGD à la vitesse r est non informatif sur $]\bar{w},+\infty[$. L'auteur de [217] donne le comportement exact sur cet intervalle.

Dans un premier temps, il est immédiat de montrer que $(I_n)_{n\geqslant 1}$ est une suite décroissante de fonctions. Nous souhaiterions ensuite étudier plus finement les propriétés de la suite I_n et montrer entre autre que

$$\lim_{n\to+\infty} I_n = I_\infty,$$

où I_∞ est la fonction de taux du PGD pour $Z_r := Z_r^\infty = \lim_{n\to+\infty} Z_r^n$ établi dans [167].

6.3 Quantification d'incertitude et réduction de dimension

Analyse de sensibilité et étude d'évenements rares

La thèse de Gabriel Sarazin, co-encadrée par Jérôme Morio (ONERA) sur un financement ONERA-Région Occitanie, s'inscrit dans la poursuite de deux axes de recherche précédemment explorés : l'analyse de sensibilité et l'étude d'événements rares. Dans cette thèse commencée en octobre 2017 et faisant suite à un stage de Master 2, la problématique concerne l'évaluation des méconnaissances des systèmes embarqués sur la retombée d'un étage de lanceur spatial. Plus précisément, après la phase propulsée, les différents étages de lanceurs spatiaux atteignent successivement leur altitude de séparation et peuvent retomber dans l'océan. Une telle phase est cruciale car la conséquence d'une erreur dans la prédiction de la zone de retombée peut être dramatique pour la sécurité et l'impact environnemental. L'évaluation de la retombée d'un lanceur est effectuée par simulations grâce à un code boîte-noire. La sortie du code fournit la position de retombée tandis que les variables d'entrée décrivent notamment les différents paramètres des systèmes embarqués (masse, position, vitesse, date du largage...). Ces entrées étant des mesures sont donc de nature incertaine. Ainsi la modélisation et la réduction d'incertitudes sur ces entrées apparaissent primordiales pour l'évaluation de la retombée ainsi que sur la probabilité que l'étage retombe en dehors de la zone prévue. A la difficulté induite par la méconnaissance de la boîte noire s'ajoute le fait que cette probabilité de défaillance est généralement très faible et inférieure à 10^{-5}.

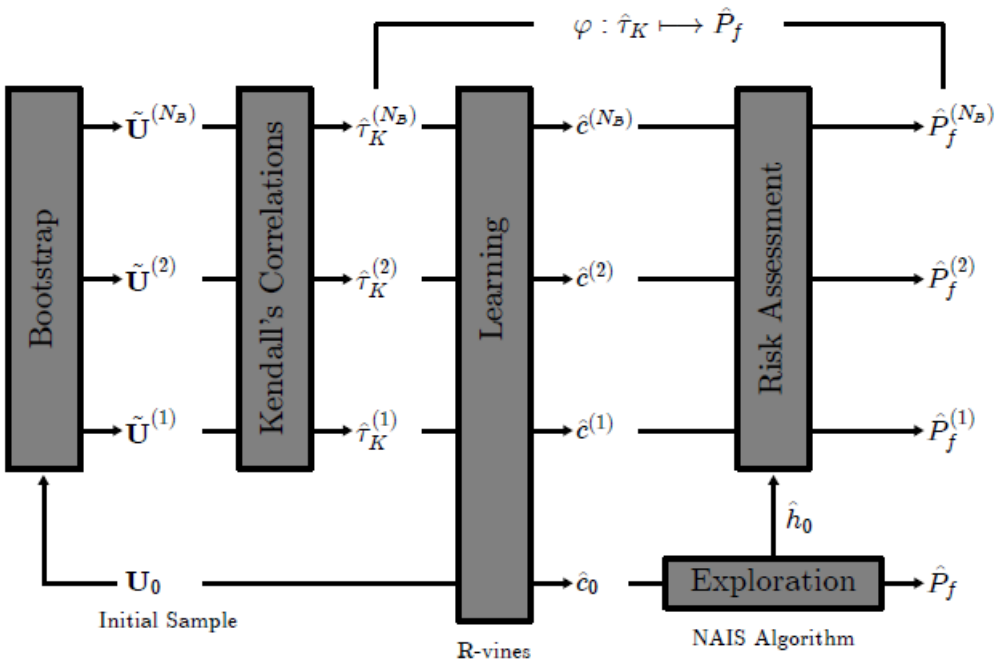
Le contexte de la thèse consiste donc en l'élaboration de méthodes de simulation pour l'estimation de la probabilité d'un événement rare avec prise en compte des méconnaissances de la distribution des entrées d'un code de calcul. Les méthodologies proposées dans le cadre des événements rares seront donc mises à profit et couplées à la modélisation de la dépendance de vecteurs aléatoires par des copules [136]. L'analyse de sensibilité s'appuira sur les indices de Borgonovo introduits dans le cadre d'entrées dépendantes [36].

Dans son stage, Gabriel a procédé à l'estimation des marginales en utilisant des modèles mixtes basés sur des méthodes par noyaux au centre et la théorie des distributions d'extremum généralisées pour la modélisation des queues de distribution.

Une étude numérique préliminaire a permis de proposer une procédure efficace à mettre en œuvre en pratique. Plus précisément, à partir d'un échantillon initial de données U_0 de taille relativement faible, nous ne pouvons obtenir qu'une seule estimation de la probabilité de défaillance
$P_f = \mathbb{P}(Y \in \mathcal{D}_f)$. Afin d'observer de la variabilité sur la probabilité de défaillance, l'idée est donc de réaliser des bootstrap conduisant à N_B jeux de données distribuées selon la loi sous-jacente f_U inconnue. A chaque jeu bootstrap de données $U_0^{(i)}$ est associée la matrice des taux de Kendall empriques $\hat{\tau}_K^{(i)}$ et la phase d'apprentssage conduit ensuite à l'estimation de la densité de la copule vigne $\hat{c}^{(i)}$. A la fin de cette phase d'apprentissage, nous procédons à l'estimation de la probabilité de défaillance P_f. Pour ce faire, nous mettons à profit les méthodes développés dans le cadre des événements rares et en particulier, nous utilisons l'algorithme NAIS (Nonparametric Adaptive Importance Sampling) dont le but est d'approcher la densité auxilliaire optimale de l'échantillonnage préférentiel sans modélisation *a priori* à l'aide d'estimateurs à noyaux pondérés [267, 183]. Dans [183], l'auteur suggère de ne considérer qu'une seule estimation $\hat{h}_0$ de h_{opt} dans l'algorithme NAIS. Cela conduit ainsi à l'estimation de la probabilité de défaillance suivante :

$$\hat{P}_f^i = \frac{1}{N_S} \sum_{k=1}^{N_S} \mathbb{1}_{\mathcal{D}_f}(U^{(k)}) \frac{\hat{f}^i(U^{(k)})}{\hat{h}_0(U^{(k)})},$$

pour $i = 1, \ldots, N_B$.



Enfin, nous avons réalisé une analyse de sensibilité sur les taux de Kendall afin de déterminer quelles sont les paires de copules les plus influentes sur l'estimation de P_f. Dans notre cadre de travail, les entrées du code constituées par les taux de Kendall estimés sont dépendantes et les méthodes proposées dans le Chapitre 4 s'avèrent donc inutilisables. L'idée est donc de mettre à profit les indices proposés par Borgonovo [36] :

$$\delta^{\mathbf{u}} = \frac{1}{2}\mathbb{E}\left[\left\|f_Y - f_Y^{X_{\mathbf{u}}}\right\|_{L^1(\mathbb{R})}\right] = \frac{1}{2}\mathbb{E}\left[\int \left|f_Y(y) - f_Y^{X_{\mathbf{u}}}(y)\right| \mathrm{d}y\right]$$

dont une procédure d'estimation a été proposée dans [89] basée sur la réécriture en terme de copule suivante :

$$\delta^{\mathbf{u}} = \frac{1}{2}\int_0^1 \int_0^1 |c(u,v) - 1| \,\mathrm{d}u\mathrm{d}v.$$

Il s'avère que ces indices sont, comme les indices de Cramér-von Mises, basés sur toute la distribution et non uniquement sur les moments d'ordre 2. Une fois la paire la plus influente identifiée, il sera alors ensuite possible de procéder à une phase d'enrichissement selon une procédure qui reste à développer. Un premier pas dans cette direction a fait l'objet de la prépublication suivante [S18].

L'objectif suivant sera de procéder à une phase d'analyse de sensibilité sur la probabilté de défaillance afin de déterminer qui des marginales ou de la copule a le plus d'influence sur l'estimation de la probabilité de défaillance P_f. Le contexte est donc atypique puisque n'ayant pas accès directement aux marginales et à la copule, nous devons procéder en premier lieu à une phase d'apprentissage. L'objectif est donc de développer une stratégie efficace dans ce cadre pour pouvoir réaliser une étude de sensibilité.

Sous espaces actifs

De nombreux problèmes issus de la quantification d'incertitude requièrent une réduction de dimension préalable. Pour ce faire, une première possibilité pour la réduction de la dimension de l'espace des entrées est de procéder à la décomposition de Karhunen-Loève de la fonction d'intérêt et de ne travailler qu'avec sa troncature [230]. D'autre part, comme nous l'avons expliqué précédemment, l'analyse de sensibilité permet d'ordonner les variables d'entrée par ordre d'importance relativement à la sortie du code. En pratique, ce classement peut ensuite être exploité pour mettre en lumière un sous-ensemble de variables

influentes. Numériquement, on fixe ensuite les variables non influentes à leur valeur nominale et on procède ainsi à une réduction de dimension de l'espace d'entrée. Néanmoins, en se limitant à des sous-ensembles de variables d'entrée, nous ne considérons pas des structures linéaires plus générales de dimension inférieures elles aussi. Pour pallier ce défaut, [66, 67] ont introduit les *sous-espaces actifs* (*active subspaces*) qui sont les espaces propres du produit scalaire moyen entre le gradient de la fonction et lui-même. L'idée naturelle sous-jacente est que ces espaces capturent les directions suivant lesquelles la fonction "varie" le plus. Plus précisément, on considère la fonction $f : \mathbb{R}^p \to \mathbb{R}$ et on diagonalise la matrice $p \times p$ suivante

$$C = \int (\nabla f(x))^\top \nabla f(x) \mu(dx) = W \Lambda W^\top,$$

où μ est la loi de probabilité du vecteur d'entrée X, W est la matrice orthogonale des vecteurs propres et Λ la matrice diagonale des valeurs propres rangées par ordre croissant. En considérant les k premières colonnes de W, on partitionne ainsi l'espace en deux sous-ensembles : le premier, noté W_1, correspondant à l'espace vectoriel engendré par les k premières colonnes de W et donnant le sous-espace actif de dimension k et le second engendré par les colonnes restantes de W. A partir d'un échantillon de taille N de X, on approche ensuite $f(x)$ par $g(\widehat{W}_1^\top x)$ où $g : \mathbb{R}^k \to \mathbb{R}$ et $\widehat{W}_1$ est défini à partir de

$$\widehat{C} = \frac{1}{N} \sum_{i=1}^{N} (\nabla f(X_i))^\top \nabla f(X_i) \mu(dx) = \widehat{W} \widehat{\Lambda} \widehat{W}^\top.$$

Une généralisation pour des fonctions multivariées (et non scalaires) a été proposée dans [263]. Les auteurs recherchent aussi une approximation de la fonction f de la forme $g(P_k)$ (où P_k est la projection sur le sous-espace actif de dimension k) qui exploite la structure et contrôlent l'erreur en utilisant des bornes de type Poincaré.

Une perspective qui nous paraît intéressante est de faire le lien avec les modèles d'indice monotone linéaire (linear monotone index) qui sont une généralisation du modèle linéaire et offrent une plus grande flexibilité. Ces modèles ont été étudiés par Hristache *et al.* [124], Chiou et Müller [63], Dalalyan *et al.* [77] (modèle d'indice multiple monotone) et Durot *et al.* [23] (modèles d'indice simple monotone) et d'envisager des extensions non linéaires. Dans les travaux précédents, le cadre considéré est le suivant :

$$Y = f(X) + \varepsilon = g(\alpha_1^\top X, \ldots, \alpha_k^\top X) + \varepsilon,$$

où on suppose que toute l'information sur $f : \mathbb{R}^p \to \mathbb{R}$ est contenue dans un sous-ensemble de dimension plus petite, $g : \mathbb{R}^k \to \mathbb{R}$ et $k < p$. L'objectif est d'estimer $\mathrm{Vect}\{\alpha_1^*, \ldots, \alpha_k^*\}$ où $(\alpha_1^*, \ldots, \alpha_k^*)$ donne le sous-espace effectif (minimal). Différentes stratégies ont été étudiées dans les travaux cités précédemment selon que $k = 1$ ou non, le gradient de f est connu ou non...

Avec Fabrice Gamboa (IMT), Alexandre Janon (Université Paris Sud) et Thierry Klein (IMT-ENAC), nous souhaitons généraliser le modèle d'indice monotone de la façon suivante : la variable réponse réelle $Y \in \mathbb{R}$ est une fonction du vecteur d'entrée $X \in U$, où U est un ouvert de $\mathbb{R}^p$, donnée par :

$$Y = f(\theta^\top \phi(X)) \tag{6.2}$$

où $\phi : U \to \mathbb{R}^k$ est une fonction différentiable, appelée le *dictionnaire*, $\theta \in \mathbb{R}^k$ est le vecteur des paramètres du modèle et $f : \mathbb{R} \to \mathbb{R}$ est la *fonction de lien.*
Nous supposons la fonction ϕ connue. L'objectif est alors d'estimer f et θ à partir de l'échantillon de taille N suivant $((X_1, Y(X_1)), \ldots, (X_N, Y(X_N)))$.
Afin d'assurer l'identifiabilité du modèle, nous supposons que

— la fonction de lien f est continue, (strictement) monotone, dérivable sur $\mathbb{R}$ (excepté éventuellement en un nombre fini de points) ;

— le jacobien du dictionnaire ϕ est non dégénéré : pour tout sous-ensemble fini S de $\mathbb{R}$, toute fonction $\lambda : \mathbb{R} \to \mathbb{R}$ et tous vecteurs $u, v \in \mathbb{R}^k$, nous avons

$$\big(\forall x \in \mathbb{R} \setminus S, (u - \lambda(x)v)^\top \phi'(x) = 0\big) \Longrightarrow (\forall x \in \mathbb{R}, u = \lambda(x)v)$$

où $\phi'(x)$ représente le jacobien de taille $k \times p$ de ϕ ;

— le paramètre vectoriel θ est normalisé : $\theta = (1, \theta_2, \ldots, \theta_k)$;

— le dictionnaire est de rang plein : $\{\theta^\top \phi(x), x \in \mathbb{R}\} = \mathbb{R}$.

Nous souhaitons proposer un algorithme itératif en deux étapes pour estimer f et θ conjointement.

Etape 1 : Estimation de f. Estimation linéaire par morceaux, voire plus régulière : par noyau par exemple en imposant la condition de monotonie.

Etape 2 : Estimation de θ. Estimation par moindres carrés avec éventuellement une condition de sparsité.

Il s'agira ensuite d'étudier les propriétés de l'algorithme ; notamment, de montrer que cet algorithme est convergeant en espérant qu'effectivement la minimisation alternée permet de réduire l'erreur totale. Nous pourrons nous interroger sur la nature du minimum atteint (local, global) et l'influence de l'initialisation. Ensuite, nous espérons montrer que les estimateurs ainsi obtenus sont consistants pour l'estimation de f et θ. Enfin, une analyse de l'erreur sera menée.

Liste des notations

En général, les quantités déterministes sont notées par des lettres minuscules tandis que les quantités aléatoires sont réprésentées avec des lettres capitales.
La notation $(Cste)$ représente une constante générique dont la valeur peut varier d'une ligne à l'autre.
Il est pratique d'avoir des expressions courtes pour des termes qui convergent en probabilité vers zéro. Ainsi la notation classique $o_{\mathbb{P}}(1)$ (resp. $O_{\mathbb{P}}(1)$) représente une suite de variables aléatoires qui convergent ves zéro en probabilité (resp. est bornée en probabilité) lorsque n ou N tend vers l'infini.

Abréviations

i.e.	*Id est*
v.a.	Variable aléatoire
i.i.d.	Indépendant et identiquement distribué
p.s.	Presque sûrement
FR	Fonction de répartition
IC	Intervalle de confiance
TCL	Théorème central limite
MB	Mouvement Brownien
PG	Processus Gaussien
LOO	Leave-one-out
MLE	Estimateur par maximum de vraisemblance
cMLE	Estimateur par maximum de vraisemblance sous contraintes d'inégalités
CLE	Estimateur par vraisemblance composite
ACP	Analyse en composantes principales

Symboles mathématiques

$\mathbb{1}$	Fonction indicatrice
$\mathbf{1}$	Vecteur ayant pour coordonnées 1
$\|v\|$	Norme du vecteur v
$\langle u, v\rangle$	Produit scalaire de u et v
$\|A\|$	Cardinal de l'ensemble A

Matrices

I_n	Matrice idendité de taille n
$\mathrm{Tr}(M)$	Trace de la matrice M
$\det M$	Déterminant de la matrice M
$\mathrm{Diag}(M)$	Pour une matrice M : $(\mathrm{Diag}(M))_{i,j} = M_{i,j}\mathbb{1}_{i=j}$

Théorie des probabilités

$\mathbb{P}(A)$	Probabilté de l'événement A
$\mathbb{E}[X]$	Espérance de X
$\mathrm{Var}(X)$	Variance de X
$\mathrm{Cov}(X, Y)$	Covariance entre X et Y
$\mathcal{N}_k(m, \Sigma)$	Distribution Gaussienne en dimension k de vecteur moyenne $m \in \mathbb{R}^k$ et de matrice de variance-covariance $\Sigma \in \mathcal{M}_{k\times k}$
$\mathrm{GP}(m, K)$	Gaussian process with mean function m and covariance function K
Φ_{m,σ^2}	La F.R. de la loi Gaussienne $\mathcal{N}(m, \sigma^2)$
Φ	$\Phi_{0,1}$

Evénements rares

A	Evénement cible $A = B_{M+1}$
B_i	Niveau intermédiaire pour $i = 1, \ldots, M+1$
P_i	Probabilité de transition entre le niveau B_{i-1} et le niveau B_i pour $i = 1, \ldots, M+1$
R_i	Nombre de retirages depuis le niveau B_i pour $i = 1, \ldots, M$
p_i	Probabilité d'atteindre le niveau B_i depuis O pour $i = 1, \ldots, M+1$ et $p_0 = 1$
r_i	Produit des nombres de réplications : $r_i = R_1 \ldots R_i$ pour $i = 1, \ldots, M$ et r_0=1
c	fonction de coût par particule

Processus Gaussiens

n	Nombre de points d'observation
d	Dimension de l'espace d'entrées du processus Gaussien
p	Nombre de paramètres de la famille dparamétrique de fonctions de covariance
σ^2	Variance du processus
α	Paramètre d'échelle $\alpha \in \mathbb{R}^p$
σ_0^2	Vraie variance du processus
α_0	Vrai paramètre d'échelle $\alpha_0 \in \mathbb{R}^p$
k_α	Fonction de covariance
$\mathcal{L}_n(\sigma^2, \alpha)$	Log vraisemblance
$\mathcal{L}_{c,n}(\sigma^2, \alpha)$	Log vraisemblance sous contraintes d'inégalités
$(\sigma^2_{ML}, \alpha_{ML})$	Estimateur de (σ^2, α) par maximum de vraisemblance
$(\sigma^2_{cML}, \alpha_{cML})$	Estimateur de (σ^2, α) par maximum de vraisemblance sous contraintes d'inégalités
$(\sigma^2_{CL}, \alpha_{CL})$	Estimateur de (σ^2, α) par vraisemblance composite
V	Variogramme du processus : $V(h) = \frac{1}{2}\mathbb{E}[(Y(t+h) - Y(t))^2]$
C	Constante intervenant dans le variogramme
a	Suite non nulle à support fini $a = (a_0, \ldots, a_{L(a)-1})$
$C_{a,n}$	Estimateur par variations quadratiques de C associé à a

Séquences biologiques

n	Longueurs des séquences
N	Taille des échantillons Monte-Carlo
M	Paramètre d'échelle

Temps discret

ε_i	Score ponctuel à la position i
S_n	Somme partielles des $(\varepsilon_i)_{i\leqslant i\leqslant n}$
U_n	Processus de Lindley
$\overline{U}_n$	Score local
θ_n	Longueur du score local
U_n^*	Score local pour les excursions complètes
θ_n^*	Longueur du score local pour les excursions complètes

Temps continu

$B(t)$	Mouvement Brownien standard issu de 0
$U(t)$	Processus de Lindley
$\overline{U}(t)$	Score local
$\theta(t)$	Longueur du score local
$U^*(t)$	Score local pour les excursions complètes
$\theta^*(t)$	Longueur du score local pour les excursions complètes
$R(t)$	Processus de Bessel de dimension 3 issu de 0

Principe de grandes déviations

Hachage avec essais linéaires

m	Nombre de places
n	Nombre de voitures
N	Nombre de blocs : $N = m - n$
$d_{m,n}$	Déplacement total
X	V.a. distribuée selon la loi de Borel de paramètre μ
Y	V.a. qui conditionnée à $\{X = l\}$ est distribuée comme $d_{l,l-1}$
S_n	Somme partielle des X_i pour $i = 1, \ldots, n$
T_n	Somme partielle des Y_i pour $i = 1, \ldots, n$

Analyse de sensibilité

N	Nombre de points d'observation dans le plan d'expérience initial
p	Dimension de l'espace d'entrée du code
k	Dimension de l'espace de sortie du code pour les codes vectoriels
m	Nombre de termes dans les sommes tronquées
q	Ordre des moments
$E = E_1 \times \ldots \times E_p$	Espace d'entrée
$\mathbb{H}$	Espace de Hilbert séparable
$\mathcal{X}$	Espace métrique séparable

I_p	L'ensemble des entiers de 1 à p
$\mathbf{u}$	Un sous-ensemble de I_p
$\mathbf{U}$	k sous-ensembles de I_p : $\mathbf{U} = \{\mathbf{u}_1, \dots, \mathbf{u}_k\}$
$X = (X_1, \dots, X_p)$	Entrée du code
$X_{\mathbf{u}}$	Coordonnées X_i de X telles que $i \in \mathbf{u}$
$X_{\sim \mathbf{u}}$	Coordonnées X_i de X telles que $i \notin \mathbf{u}$
$X^{\mathbf{u}}$	Version Pick-Freeze de X par rapport à $X_{\mathbf{u}}$
Y	Sortie du code
$Y^{\mathbf{u}}$	Version Pick-Freeze de Y par rapport à $X_{\mathbf{u}}$

Sortie/Modèle	Type d'indice	Indice premier ordre	Estimateur	Indice total
Scalaire $Y \in \mathbb{R}$	Sobol	$S^{\mathbf{u}}$	$\widehat{S}^{\mathbf{u}}$, $\widehat{T}^{\mathbf{u}}$ $\widehat{S}_c^{\mathbf{u}}$ (cas centré) $\widehat{S}^{\mathbf{u}}_{\text{Méta}}$ (métamodèle)	$S^{\mathbf{u},\text{Tot}}$
Vectorielle $Y \in \mathbb{R}^k$	Sobol Ordre q Cramér-von-Mises	$S^{\mathbf{u},k}$ $(T^{\mathbf{u},k})$ $H_q^{\mathbf{u}}$ $S^{\mathbf{u}}_{2,\text{CVM}}$	$\widehat{S}^{\mathbf{u},k}$ $\widehat{H}_q^{\mathbf{u}}$ $\widehat{S}^{\mathbf{u},\text{Tot}}_{2,\text{CVM}}$	
Fonctionnelle $Y \in \mathbb{H}$	Sobol	$S^{\mathbf{u},\infty}$	$\widehat{S}^{\mathbf{u},\infty}_m$	
Régression linéaire	Sobol	$S^{\mathbf{u}}_{\text{Lin}}$	$\widehat{S}^{\mathbf{u}}_{\text{Lin},m}$	
$Y \in \mathcal{X}$	Général	$S^{\mathbf{u}}_{q,\text{GMS}}$	$\widehat{S}^{\mathbf{u}}_{q,\text{GMS}}$	$S^{\mathbf{u},\text{Tot}}_{q,\text{GMS}}$

Bibliographie

[1] P. Abrahamsen. A review of Gaussian random fields and correlation functions. Technical report, Norwegian computing center, 1997.

[2] M. Abramowitz and I. A. Stegun. Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables. Dover, New York, ninth Dover printing, tenth GPO printing edition, 1964.

[3] L. V. Ahlfors. Lectures on quasiconformal mappings, volume 38 of University Lecture Series. American Mathematical Society, Providence, RI, second edition, 2006. With supplemental chapters by C. J. Earle, I. Kra, M. Shishikura and J. H. Hubbard.

[4] D. Aldous. Probability approximations via the Poisson clumping heuristic, volume 77 of Applied Mathematical Sciences. Springer-Verlag, New York, 1989.

[5] D. Aldous and U. V. Vazirani. "go with the winners" algorithms. IEEE Symposium on Fundations of Computer Science, (7) :492–501, 1994.

[6] S. Altschul, W. Gish, W. Miller, E. Myers, and D. Lipman. Basic Local Alignment Search Tool. JMB, 215 :403–410, 1990.

[7] E. Anderes. On the consistent separation of scale and variance for Gaussian random fields. The Annals of Statistics, 38 :870–893, 2010.

[8] R. Antonini. Sur le comportement asymptotique du processus de Ornstein-Uhlenbeck multidimensionnel. Ann. Sci. Univ. Clermont-Ferrand II Probab. Appl., 9 :33–44, 1991.

[9] R. Arratia and M. S. Waterman. The Erdos-Rényi strong law for pattern matching with a given proportion of mismatches. Ann. Probab., 17(3) :1152–1169, 1989.

[10] S. Asmussen. Applied probability and queues, volume 51 of Applications of Mathematics (New York). Springer-Verlag, New York, second edition, 2003. Stochastic Modelling and Applied Probability.

[11] S. Au and J. Beck. A new adaptive importance sampling scheme for reliability calculations. Structural Safety, 21(2) :135 – 158, 1999.

[12] S. Au, C. Papadimitriou, and J. Beck. Reliability of uncertain dynamical systems with multiple design points. Structural Safety, 21(2) :113 – 133, 1999.

[13] S.-K. Au and J. L. Beck. Estimation of small failure probabilities in high dimensions by subset simulation. Probabilistic Engineering Mechanics, 16 (4) :263–277, 2001.

[14] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. Machine Learning, 47(2) :235–256, May 2002.

[15] D. Bachmann and H. Dette. A note on the the Bickel-Rosenblatt test in autoregressive time series. Statist. Probab. Lett., 74 :221–234, 2005.

[16] F. Bachoc. Cross validation and maximum likelihood estimations of hyper-parameters of Gaussian processes with model mispecification. Computational Statistics and Data Analysis, 66 :55–69, 2013.

[17] F. Bachoc. Asymptotic analysis of the role of spatial sampling for covariance parameter estimation of Gaussian processes. Journal of Multivariate Analysis, 125 :1–35, 2014.

[18] F. Bachoc. Asymptotic analysis of the role of spatial sampling for covariance parameter estimation of Gaussian processes. Journal of Multivariate Analysis, 125 :1–35, 2014.

[19] F. Bachoc. Asymptotic analysis of covariance parameter estimation for Gaussian processes in the misspecified case. Bernoulli, 24(2) :1531–1575, 2018.

[20] F. Bachoc, K. Ammar, and J. Martinez. Improvement of code behavior in a design of experiments by metamodeling. Nuclear science and engineering, 183(3) :387–406, 1016.

[21] F. Bachoc, M. Bevilacqua, and D. Velandia. Composite likelihood estimation for a gaussian process under fixed domain asymptotics. arXiv :1807.08988, 2018.

[22] F. Bachoc, G. Bois, J. Garnier, and J. Martinez. Calibration and improved prediction of computer models by universal Kriging. Nuclear Science and Engineering, 176(1) :81–97, 2014.

[23] F. Balabdaoui, C. Durot, and H. Jankowski. Least squares estimation in the monotone single index model. arXiv e-prints, page arXiv :1610.06026, Oct 2016.

[24] M. Barczy and E. Iglói. Karhunen-Loève expansions of α-Wiener bridges. Cent. Eur. J. Math., 9(1) :65–84, 2011.

[25] F. Barthe, P. Cattiaux, and C. Roberto. Concentration for independent random variables with heavy tails. AMRX Appl. Math. Res. Express, (2) :39–60, 2005.

[26] M. S. Bartlett. The Characteristic Function of a Conditional Statistic. J. London Math. Soc., S1-13(1) :62.

[27] J. M. Bates and C. W. Granger. The combination of forecasts. Journal of the Operational Research Society, 20(4) :451–468, 1969.

[28] G. Baxter. A strong limit theorem for Gaussian processes. Proc. Amer. Math. Soc., 7 :522–527, 1956.

[29] J. Bertoin. Lévy processes, volume 121 of Cambridge Tracts in Mathematics. Cambridge University Press, Cambridge, 1996.

[30] M. Bevilacqua, T. Faouzi, R. Furrer, and E. Porcu. Estimation and prediction using generalized Wendland covariance functions under fixed domain asymptotics. The Annals of Statistics, 47(2) :828–856, 2019.

[31] P. Biane, J.-F. Le Gall, and M. Yor. Un processus qui ressemble au pont brownien. In Séminaire de Probabilités, XXI, volume 1247 of Lecture Notes in Math., pages 270–275. Springer, Berlin, 1987.

[32] P. Biane, J. Pitman, and M. Yor. Probability laws related to the Jacobi theta and Riemann zeta functions, and Brownian excursions. Bull. Amer. Math. Soc. (N.S.), 38(4) :435–465 (electronic), 2001.

[33] P. J. Bickel and M. Rosenblatt. On some global measures of the deviations of density function estimates. Ann. Statist., 1 :1071–1095, 1973.

[34] P. Billingsley. Convergence of probability measures. Wiley Series in Probability and Statistics : Probability and Statistics. John Wiley & Sons Inc., New York, second edition, 1999. A Wiley-Interscience Publication.

[35] S. Bobkov and M. Ledoux. One-dimensional empirical measures, order statistics, and kantorovich transport distances. Memoirs of the American Mathematical Society, To appear, 2019.

[36] E. Borgonovo. A new uncertainty importance measure. Reliability Engineering & System Safety, 92(6) :771–784, 2007.

[37] E. Borgonovo and M. Baucells. Invariant probabilistic sensitivity analysis. Management Science, 59(11) :2536–2549, 2013.

[38] E. Borgonovo, W. Castaings, and S. Tarantola. Moment independent importance measures : New results and analytical test cases. Risk Analysis, 31(3) :404–428, 2011.

[39] E. Borgonovo, W. Castaings, and S. Tarantola. Model emulation and moment-independent sensitivity analysis : An application to environmental modelling. Environmental Modelling & Software, 34 :105–115, 2012.

[40] E. Borgonovo, G. Hazen, and E. Plischke. Probabilistic Sensitivity Measures : Foundations and Estimation. Manuscript, pages 1–24, 2014.

[41] E. Borgonovo and B. Iooss. Moment Independent and Reliability-Based Importance Measures, pages 1–23. Springer International Publishing, Cham, 2016.

[42] A. A. Borovkov. Stochastic processes in queueing theory. Springer-Verlag, New York, 1976. Translated from the Russian by Kenneth Wickwire, Applications of Mathematics, No. 4.

[43] A. A. Borovkov and K. A. Borovkov. Probabilities of large deviations for random walks with regular distribution of jumps. Dokl. Akad. Nauk, 371(1) :14–16, 2000.

[44] J. C. Bronski. Small ball constants and tight eigenvalue asymptotics for fractional Brownian motions. J. Theoret. Probab., 16(1) :87–100, 2003.

[45] T. Browne, B. Iooss, L. Le Gratiet, J. Lonchampt, and E. Remy. Stochastic simulators based optimization by gaussian process metamodels - application to maintenance investments planning issues. Quality and Reliability Engineering International, 32 :2067–2080, 2016.

[46] R. Buchbinder and A. S. Detsky. Management of suspected giant cell arteritis : A decision analysis. J. Rheumatology, 19(9) :1220–1228, 1992.

[47] D. Cacuci. Sensitivity and Uncertainty Anlysis, volume Vol. I Theory. Chapman and Hall/CRC, 1994.

[48] T. T. Cai and P. Hall. Prediction in functional linear regression. Ann. Statist., 34(5) :2159–2179, 2006.

[49] Y. Cao and D. J. Fleet. Generalized product of experts for automatic and principled fusion of Gaussian process predictions. In Modern Nonparametrics 3 : Automating the Learning Pipeline workshop at NIPS, Montreal, 2014. arXiv preprint arXiv :1410.7827.

[50] H. Cardot and J. Johannes. Thresholding projection estimators in functional linear models. J. Multivariate Anal., 101(2) :395–408, 2010.

[51] H. Cardot, A. Mas, and P. Sarda. CLT in functional linear regression models. Probab. Theory Related Fields, 138(3-4) :325–361, 2007.

[52] H. Cardot and P. Sarda. Functional linear regression. In The Oxford handbook of functional data analysis, pages 21–46. Oxford Univ. Press, Oxford, 2011.

[53] F. Cérou, P. Del Moral, and A. Guyader. A nonasymptotic theorem for unnormalized Feynman-Kac particle models. Ann. Inst. Henri Poincaré Probab. Stat., 47(3) :629–649, 2011.

[54] F. Cérou, P. Del Moral, F. LeGland, and P. Lezaud. Genetic genealogical model in rare event analysis. Latin American Journal of Probability And Mathematical Statistics. 2006.

[55] F. Cérou and A. Guyader. Adaptive multilevel splitting for rare event analysis. Rapport de recherche de l'INRIA - Rennes , Equipe : ASPI. 2005.

[56] C.-H. Chang, H.-C. Huang, C.-K. Ing, et al. Mixed domain asymptotics for a stochastic process model with time trend and measurement error. Bernoulli, 23(1) :159–190, 2017.

[57] P. Chassaing and P. Flajolet. Hachage, arbres, chemins & graphes. Gaz. Math., (95) :29–49, 2003.

[58] P. Chassaing and S. Janson. A Vervaat-like path transformation for the reflected Brownian bridge conditioned on its local time at 0. Ann. Probab., 29(4) :1755–1779, 2001.

[59] P. Chassaing and G. Louchard. Phase transition for parking blocks, Brownian excursion and coalescence. Random Structures Algorithms, 21(1) :76–119, 2002.

[60] P. Chassaing and J.-F. Marckert. Parking functions, empirical processes, and the width of rooted labeled trees. Electron. J. Combin., 8(1) :Research Paper 14, 19, 2001.

[61] H.-S. Chen, D. Simpson, and Z. Ying. Infill asymptotics for a stochastic process model with measurement error. Statistica Sinica, 10 :141–156, 2000.

[62] W. Chen, R. Jin, and A. Sudjianto. Analytical variance-based global sensitivity analysis in simulation-based design under uncertainty. Transactions-American Society of Mechanical Engineers Journal of Mechanical Design, 127(5) :875, 2005.

[63] J.-M. Chiou and H.-G. Müller. Quasi-likelihood regression with multiple indices and smooth link and variance functions. Scand. J. Statist., 31(3) :367–386, 2004.

[64] F. Clarke. Functional analysis, calculus of variations and optimal control, volume 264 of Graduate Texts in Mathematics. Springer, London, 2013.

[65] J.-F. Coeurjolly. Estimating the parameters of a fractional Brownian motion by discrete variations of its sample paths. Stat. Inference Stoch. Process., 4(2) :199–227, 2001.

[66] P. G. Constantine. Active subspaces, volume 2 of SIAM Spotlights. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2015. Emerging ideas for dimension reduction in parameter studies.

[67] P. G. Constantine, E. Dow, and Q. Wang. Active subspace methods in theory and practice : applications to kriging surfaces. SIAM J. Sci. Comput., 36(4) :A1500–A1524, 2014.

[68] S. Corlay and G. Pagès. Functional quantization based stratified sampling methods. Preprint, http ://hal.inria.fr/hal-00464088, 2010.

[69] A. Cousin, H. Maatouk, and D. Rullière. Kriging of financial term-structures. European Journal of Operational Research, 255(2) :631–648, 2016.

[70] C. Crambes, A. Kneip, and P. Sarda. Smoothing splines estimators for functional linear regression. Ann. Statist., 37(1) :35–72, 2009.

[71] H. Cramér. Sur un nouveau théorème-limite de la théorie des probabilités. Actualités Sci. Ind., (736), 1938.

[72] N. Cressie. Statistics for spatial data. J. Wiley, 1993.

[73] N. Cressie and S. Lahiri. Asymptotics for REML estimation of spatial covariance parameters. Journal of Statistical Planning and Inference, 50 :327–341, 1996.

[74] I. Csiszar. Sanov property, generalized i-projection and a conditional limit theorem. The Annals of Probability, 12(3) :768–793, 08 1984.

[75] S. Da Veiga. Global sensitivity analysis with dependence measures. J. Stat. Comput. Simul., 85(7) :1283–1305, 2015.

[76] S. Da Veiga and A. Marrel. Gaussian process modeling with inequality constraints. Annales de la faculté des sciences de Toulouse Mathématiques, 21(3) :529–555, 4 2012.

[77] A. S. Dalalyan, A. Juditsky, and V. Spokoiny. A new algorithm for estimating the effective dimension-reduction subspace. J. Mach. Learn. Res., 9 :1648–1678, 2008.

[78] A. Damianou. Deep gaussian processes and variational propagation of uncertainty. PhD Thesis, University of Sheffield, 2015.

[79] A. Damianou and N. Lawrence. Deep Gaussian processes. In C. Carvalho and P. Ravikumar, editors, Proceedings of the Sixteenth International Workshop on Artificial Intelligence and Statistics (AISTATS), AISTATS '13, pages 207–215. JMLR W&CP 31, 2013.

[80] A. Datta, S. Banerjee, A. O. Finley, and A. E. Gelfand. Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. Journal of the American Statistical Association, 111(514) :800–812, 2016.

[81] J.-J. Daudin, M.-P. Etienne, and P. Vallois. Asymptotic behavior of the local score of independent and identically distributed random sequences. Stochastic Process. Appl., 107(1) :1–28, 2003.

[82] M. Deaconu and S. Herrmann. Hitting time for bessel processes-walk on moving spheres algorithm (woms). The Annals of Applied Probability, 23(6) :2259–2289, 12 2013.

[83] T. Dean and P. Dupuis. Splitting for rare event simulation : A large deviation approach to design and analysis. Stochastic Processes and their Applications, 119(2) :562 – 587, 2009.

[84] M. P. Deisenroth and J. W. Ng. Distributed Gaussian processes. Proceedings of the 32nd International Conference on Machine Learning, Lille, France. JMLR : W&CP volume 37, 2015.

[85] P. Del Moral. Feynman-Kac formulae. Probability and its Applications (New York). Springer-Verlag, New York, 2004. Genealogical and interacting particle systems with applications.

[86] A. Dembo and S. Karlin. Strong limit theorems of empirical distributions for large segmental exceedances of partial sums of markov variables. The Annals of Probability, 19(4) :1756–1767, 1991.

[87] A. Dembo and S. Karlin. Strong limit theorems of empirical functionals for large exceedances of partial sums of i.i.d. variables. The Annals of Probability, 19(4) :1737–1755, 1991.

[88] A. Dembo and O. Zeitouni. Large deviations techniques and applications, volume 38 of Applications of Mathematics (New York). Springer-Verlag, New York, second edition, 1998.

[89] P. Derennes, J. Morio, and F. Simatos. Estimation of moment independent importance measures using a copula and maximum entropy framework. In Proceedings of the 2018 Winter Simulation Conference, GOTEBORG, Sweden, Dec. 2018.

[90] O. Dubrule. Cross validation of Kriging in a unique neighborhood. Mathematical Geology, 15 :687–699, 1983.

[91] M. Duflo. Random iterative models. Applications of Mathematics (vol. 34), New York. Springer-Verlag, Berlin, 1997.

[92] M.-P. Etienne and P. Vallois. Approximation of the supremum of a centered random walk. application to the local score. Methodology and Computing in Applied Probability, 6 :255–275, 2004.

[93] Y. Fan. Testing the goodness of fit of a parametric density function by kernel method. Econometric Theory., 10 :316–356, 1994.

[94] W. Feller. Generalization of a probability limit theorem of Cramér. Trans. Amer. Math. Soc., 54 :361–372, 1943.

[95] W. Feller. An introduction to probability theory and its applications. Vol. I. Third edition. John Wiley & Sons Inc., New York, 1968.

[96] J. Felli and G. Hazen. Javelin diagrams : A graphical tool for probabilistic sensitivity analysis. Decision Analysis, 1(2) :93–107, 2004.

[97] P. Flajolet, P. Poblete, and A. Viola. On the analysis of linear probing hashing. Algorithmica, 22(4) :490–515, 1998. Average-case analysis of algorithms.

[98] R. Fraiman, F. Gamboa, and L. Moreno. Sensitivity indices for output on a Riemannian manifold. arXiv e-prints, page arXiv :1810.11591, Oct 2018.

[99] R. Furrer, M. G. Genton, and D. Nychka. Covariance tapering for interpolation of large spatial datasets. Journal of Computational and Graphical Statistics, 15(3) :502–523, 2006.

[100] F. Gamboa, A. Janon, T. Klein, A. Lagnoux, and C. Prieur. Statistical inference for Sobol pick-freeze Monte Carlo method. Statistics, 50(4) :881–902, 2016.

[101] F. Gamboa, T. Klein, and C. Prieur. Conditional large and moderate deviations for sums of discrete random variables. Combinatoric applications. Bernoulli, 18(4) :1341–1360, 2012.

[102] M. D. Garrett. mlegp : Maximum Likelihood Estimates of Gaussian Processes, 2011. R package version 3.1.2.

[103] M. J. Garvels. The splitting method in rare event simulation. PhD thesis, 2000. University of Twente.

[104] B. K. Ghosh and W. M. Huang. The power and optimal kernel of the Bickel-Rosenblatt test for goodness of fit. Ann. Statist., 19 :999–1009, 1991.

[105] E. G. Gladyšev. A new limit theorem for stochastic processes with Gaussian increments. Teor. Verojatnost. i Primenen, 6 :57–66, 1961.

[106] P. Glasserman, P. Heidelberger, P. Shahabuddin, and T. Zajic. A large deviations perspective on the efficiency of multilevel splitting. IEEE Trans. Automat. Control, 43(12) :1666–1679, 1998.

[107] P. Glasserman, P. Heidelberger, P. Shahabuddin, and T. Zajic. Multilevel splitting for estimating rare event probabilities. Oper. Res., 47(4) :585–600, 1999.

[108] S. Golchi, D. R. Bingham, H. Chipman, and D. A. Campbell. Monotone emulation of computer experiments. SIAM/ASA Journal on Uncertainty Quantification, 3(1) :370–392, 2015.

[109] R. B. Gramacy and D. W. Apley. Local Gaussian process approximation for large computer experiments. Journal of Computational and Graphical Statistics, 24(2) :561–578, 2015.

[110] R. B. Gramacy et al. laGP : large-scale spatial modeling via local approximate Gaussian processes in R. Journal of Statistical Software, 72(1) :1–46, 2016.

[111] U. Grenander. Abstract inference. John Wiley & Sons, Inc., New York, 1981. Wiley Series in Probability and Mathematical Statistics.

[112] X. Guyon and J. León. Convergence en loi des H-variations d'un processus gaussien stationnaire sur $\mathbf{R}$. Ann. Inst. H. Poincaré Probab. Statist., 25(3) :265–282, 1989.

[113] P. Hall and J. L. Horowitz. Methodology and convergence rates for functional linear regression. Ann. Statist., 35(1) :70–91, 2007.

[114] D. Hamby. A review of techniques for parameter sensitivity analysis of environmental models. Environmental Monitoring and Assessment, 32(2) :135–154, 1994.

[115] J. Heinonen. What is ... a quasiconformal mapping ? Notices Amer. Math. Soc., 53(11) :1334–1335, 2006.

[116] J. Helton, J. Johnson, C. Sallaberry, and C. Storlie. Survey of sampling-based methods for uncertainty and sensitivity analysis. Reliability Engineering & System Safety, 91(10-11) :1175–1209, 2006.

[117] J. Hensman and N. Fusi. Gaussian processes for big data. Uncertainty in Artificial Intelligence, pages 282–290, 2013.

[118] G. E. Hinton. Training products of experts by minimizing contrastive divergence. Neural computation, 14(8) :1771–1800, 2002.

[119] C. Hipp. Asymptotic expansions for conditional distributions : the lattice case. Probab. Math. Statist., 4(2) :207–219, 1984.

[120] W. Hoeffding. A class of statistics with asymptotically normal distribution. Ann. Math. Statistics, 19 :293–325, 1948.

[121] L. Holst. Two conditional limit theorems with applications. Ann. Statist., 7(3) :551–557, 1979.

[122] T. Homma and A. Saltelli. Importance measures in global sensitivity analysis of nonlinear models. Reliability Engineering & System Safety, 52(1) :1–17, 1996.

[123] L. Horváth and R. Zitikis. Asymptotics of the L_p-norms of density estimators in the first-order autoregressive models. Statist. Probab. Lett., 66 :91–103, 2004.

[124] M. Hristache, A. Juditsky, J. Polzehl, and V. Spokoiny. Structure adaptive approach for dimension reduction. Ann. Statist., 29(6) :1537–1566, 2001.

[125] I. Ibragimov and R. Has' Minskii. Statistical estimation–asymptotic theory. Springer-Verlag, New York, 1981. Applications of Mathematics, Vol. 16.

[126] I. Ibragimov and Y. Rozanov. Gaussian Random Processes. Springer-Verlag, New York, 1978.

[127] T. Ishigami and T. Homma. An importance quantification technique in uncertainty analysis for computer models. In First International Symposium on Uncertainty Modeling and Analysis Proceedings, 1990., pages 398–403. IEEE, 1990.

[128] J. Istas and G. Lang. Quadratic variations and estimation of the local Hölder index of a Gaussian process. Annales de l'Institut Henri Poincaré, 33 :407–436, 1997.

[129] J. Istas and G. Lang. Quadratic variations and estimation of the local Hölder index of a Gaussian process. Ann. Inst. H. Poincaré Probab. Statist., 33(4) :407–436, 1997.

[130] K. J. and D. R.F. A simple method for displaying the hydropathic character of a protein. J. Mol. Biol., 157(1) :105–132, 1982.

[131] S. Janson. Asymptotic distribution for the cost of linear probing hashing. Random Structures Algorithms, 19(3-4) :438–471, 2001. Analysis of algorithms (Krynica Morska, 2000).

[132] S. Janson. Moment convergence in conditional limit theorems. J. Appl. Probab., 38(2) :421–437, 2001.

[133] S. Janson. Individual displacements for linear probing hashing with different insertion policies. ACM Trans. Algorithms, 1(2) :177–213, 2005.

[134] S. Janson. Individual displacements in hashing with coalesced chains. Combin. Probab. Comput., 17(6) :799–814, 2008.

[135] R. Jin, W. Chen, and A. Sudjianto. An efficient algorithm for constructing optimal design of computer experiments. J. Statist. Plann. Inference, 134(1) :268–287, 2005.

[136] H. Joe. Dependence modeling with copulas, volume 134 of Monographs on Statistics and Applied Probability. CRC Press, Boca Raton, FL, 2015.

[137] D. Jones, M. Schonlau, and W. Welch. Efficient global optimization of expensive black box functions. Journal of Global Optimization, 13 :455–492, 1998.

[138] K. Karhunen. Über lineare Methoden in der Wahrscheinlichkeitsrechnung. Ann. Acad. Sci. Fennicae. Ser. A. I. Math.-Phys., 1947(37) :79, 1947.

[139] S. Karlin and S. Altschul. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. PNAS, 87 :2264–2268, 1990.

[140] S. Karlin and F. Ost. Maximal length of common words among random letter sequences. Ann. Probab., 16(2) :535–563, 1988.

[141] C. Kaufman and B. Shaby. The role of the range parameter for estimation and prediction in geostatistics. Biometrika, 100 :473–484, 2013.

[142] C. G. Kaufman, M. J. Schervish, and D. W. Nychka. Covariance tapering for likelihood-based estimation in large spatial data sets. Journal of the American Statistical Association, 103(484) :1545–1555, 2008.

[143] J. T. Kent and A. T. A. Wood. Estimating the fractal dimension of a locally self-similar Gaussian process by using increments. J. Roy. Statist. Soc. Ser. B, 59(3) :679–699, 1997.

[144] A. Kinchin. über einer neuen grenzwertsatz der wahrscheinlichkeitsrechnung. Math. Ann., 101 :745–752, 1929.

[145] A. D. Kiureghian and T. Dakessian. Multiple design points in first and second-order reliability. Structural Safety, 20(1) :37 – 49, 1998.

[146] D. E. Knuth. Computer science and its relation to mathematics. Amer. Math. Monthly, 81 :323–343, 1974.

[147] D. E. Knuth. The art of computer programming. Vol. 3. Addison-Wesley, Reading, MA, 1998. Sorting and searching, Second edition [of MR0445948].

[148] D. E. Knuth. Linear probing and graphs. Algorithmica, 22(4) :561–568, 1998. Average-case analysis of algorithms.

[149] V. F. Kolchin. Random mappings. Translation Series in Mathematics and Engineering. Optimization Software, Inc., Publications Division, New York, 1986. Translated from the Russian, With a foreword by S. R. S. Varadhan.

[150] V. Korolev and I. Shevtsova. An upper bound for the absolute constant in the Berry-Esseen inequality. Teor. Veroyatn. Primen., 54(4) :671–695, 2009.

[151] D. G. Krige. A statistical approach to some basic mine valuation problems on the witwatersrand. Journal of the Chemical, Metallurgical and Mining Society of South Africa, 52 :119–139, 1951.

[152] È. M. Kudlaev. Conditional limit distributions of sums of random variables. Teor. Veroyatnost. i Primenen., 29(4) :743–752, 1984.

[153] M. Lamboni, H. Monod, and D. Makowski. Multivariate sensitivity analysis to measure global contribution of input factors in dynamic models. Reliability Engineering & System Safety, 96(4) :450–459, 2011.

[154] G. Lang and F. Roueff. Semi-parametric estimation of the Hölder exponent of a stationary Gaussian process with minimax rates. Stat. Inference Stoch. Process., 4(3) :283–306, 2001.

[155] B. Lapeyre, É. Pardoux, and R. Sentis. Méthodes de Monte-Carlo pour les équations de transport et de diffusion. Mathématiques & Applications [Mathematics & Applications], 29. Springer-Verlag, Berlin, 1998.

[156] R. Latał a. Estimation of moments of sums of independent real random variables. Ann. Probab., 25(3) :1502–1513, 1997.

[157] F. Lavancier and P. Rochet. A general procedure to combine estimators. Computational Statistics & Data Analysis, 94 :175–192, 2016.

[158] J.-F. Le Gall. Random trees and applications. Probab. Surv., 2 :245–311, 2005.

[159] L. Le Gratiet. Asymptotic normality of a sobol index estimator in gaussian process regression framework. Preprint, 2013.

[160] P. L'Ecuyer, F. Le Gland, P. Lezaud, and B. Tuffin. Splitting techiques. In Rare event simulation using Monte Carlo methods, pages 39–61. Wiley, Chichester, 2009.

[161] M. Ledoux. The concentration of measure phenomenon, volume 89 of Mathematical Surveys and Monographs. American Mathematical Society, Providence, RI, 2001.

[162] S. Lee and S. Na. On the Bickel-Rosenblatt test for first-order autoregressive models. Statist. Probab. Lett., 56 :23–35, 2002.

[163] F. LeGland and N. Oudjane. A sequential particle algorithm that keeps the particle system alive. 2006. Lecture Notes in Control and Information Sciences 337.

[164] P. Lévy. Le mouvement brownien plan. Amer. J. Math., 62 :487–550, 1940.

[165] W. V. Li and W. Linde. Approximation, metric entropy and small ball estimates for gaussian measures. Ann. Probab., 27 :1556–1578, 1999.

[166] J. V. Linnik. On the probability of large deviations for the sums of independent variables. In Proc. 4th Berkeley Sympos. Math. Statist. and Prob., Vol. II, pages 289–306. Univ. California Press, Berkeley, Calif., 1961.

[167] Q. Liu and A. Rouault. Limit theorems for Mandelbrot's multiplicative cascades. Ann. Appl. Probab., pages 218–239, 2000.

[168] M. Loève. Fonctions aléatoires de second ordre. Revue Sci., 84 :195–206, 1946.

[169] M. Loève. Probability theory. I. Fourth edition. Springer-Verlag, New York, 1977. Graduate Texts in Mathematics, Vol. 45.

[170] A. F. López-Lopera. Lineqgpr : Gaussian process regression models with linear inequality constraints, 2017. R package version 0.0.1. This package will be freely available in June 2018.

[171] A. F. López-Lopera, F. Bachoc, N. Durrande, and O. Roustand. Finite-dimensional Gaussian approximation with linear inequality constraints. SIAM/ASA Journal on Uncertainty Quantification, forthcoming, 2018.

[172] H. Luschgy and G. Pagès. Functional quantization of Gaussian processes. J. Funct. Anal., 196(2) :486–531, 2002.

[173] W. Madych and S. Nelson. Bounds on multivariate polynomials and exponential error estimates for multiquadric interpolation. Journal of Approximation Theory, 70(1) :94–114, 1992.

[174] J.-F. Marckert. Parking with density. Random Structures Algorithms, 18(4) :364–380, 2001.

[175] K. Mardia and R. Marshall. Maximum likelihood estimation of models for residual covariance in spatial regression. Biometrika, 71 :135–146, 1984.

[176] A. Marrel, B. Iooss, B. Laurent, and O. Roustant. Calculations of sobol indices for the gaussian process metamodel. Reliability Engineering & System Safety, 94(3) :742–751, 2009.

[177] J. Mateu, E. Porcu, G. Christakos, and M. Bevilacqua. Fitting negative spatial covariances to geothermal field temperatures in Nea Kessani (Greece). Environmetrics : The official journal of the International Environmetrics Society, 18(7) :759–773, 2007.

[178] G. Matheron. La théorie des variables régionalisées et ses applications. Fasicule 5 in Les Cahiers du Centre de Morphologie Mathématique de Fontainebleau, page 212, 1970.

[179] S. Mercier, D. Cellier, and D. Charlot. An improved approximation for assessing the statistical significance of molecular sequence features. J. Appl. Probab., 40(2) :427–441, 2003.

[180] S. Mercier and J.-J. Daudin. Exact distribution for the local score of one i.i.d. random sequence. Jour. Comp. Biol, 8(4) :373–380, 2001.

[181] J. Mockus, V. Tiesis, and A. Zilinskas. The application of bayesian methods for seeking the extremum. Towards Global Optimization. In L. Dixon and Eds G. Szego., 2 :117–129, 1978.

[182] H. Monod, C. Naud, and D. Makowski. Uncertainty and sensitivity analysis for crop models. In D. Wallach, D. Makowski, and J. W. Jones, editors, Working with Dynamic Crop Models : Evaluation, Analysis, Parameterization, and Applications, chapter 4, pages 55–99. Elsevier, 2006.

[183] J. Morio. Non-parametric adaptive importance sampling for the probability estimation of a launcher impact position. Reliability Engineering & System Safety, 96(1) :178 – 183, 2011. Special Issue on Safecomp 2008.

[184] V. Moutoussamy, S. Nanty, and B. t. Pauwels. Emulators for stochastic simulation codes. In CEMRACS 2013—modelling and simulation of complex systems : stochastic and deterministic approaches, volume 48 of ESAIM Proc. Surveys, pages 116–155. EDP Sci., Les Ulis, 2015.

[185] A. Müller. Integral probability metrics and their generating classes of functions. Adv. in Appl. Probab., 29(2) :429–443, 1997.

[186] A. V. Nagaev. Integral limit theorems taking large deviations into account when cramér's condition does not hold. i. Theory of Probability and Its Applications, 14(1) :51–64, 1969.

[187] A. V. Nagaev. Integral limit theorems taking large deviations into account when cramér's condition does not hold. ii. Theory of Probability and Its Applications, 14(2) :193–208, 1969.

[188] S. V. Nagaev. An integral limit theorem for large deviations. Izv. Akad. Nauk UzSSR Ser. Fiz.-Mat. Nauk, 1962(6) :37–43, 1962.

[189] J. v. Neumann and O. Morgenstern. Theory of Games and Economic Behavior. Princeton, NJ. Princeton University Press, 1953.

[190] M. H. Neumann and E. Paparoditis. On bootstrapping L_2-type statistics in density testing. Statist. Probab. Lett., 50 :137–147, 2000.

[191] A. Owen. Variance components and generalized sobol' indices. SIAM/ASA Journal on Uncertainty Quantification, 1(1) :19–41, 2013.

[192] A. Owen, J. Dick, and S. Chen. Higher order sobol' indices. Information and Inference, 3(1) :59–81, 2014.

[193] A. B. Owen. Orthogonal arays for computer experiments, integration and visualization. Statistica Sinica, 2(2) :439–452, 1992.

[194] A. B. Owen. Better estimation of small sobol' sensitivity indices. ACM Trans. Model. Comput. Simul., 23(2) :11 :1–11 :17, May 2013.

[195] E. Pardo-Igúzquiza and P. A. Dowd. AMLE3D : A computer program for the inference of spatial covariance parameters by approximate maximum likelihood estimation. Computers & Geosciences, 23(7) :793–805, 1997.

[196] E. Parzen. On estimation of a probability density function and mode. Ann. Math. Statist., 33 :1065–1076, 1962.

[197] G. Patane, X. S. Li, and D. X. Gu. An introduction to ricci flow and volumetric approximation with applications to shape modeling. In SIGGRAPH Asia 2014 Courses, SA '14, pages 4 :1–4 :118, New York, NY, USA, 2014. ACM.

[198] R. Paulo, G. Garcia-Donato, and J. Palomo. Calibration of computer models with multivariate output. Computational Statistics and Data Analysis, 56 :3959–3974, 2012.

[199] Y. L. Pavlov. Limit theorems for the number of trees of a given size in a random forest. Mat. Sb. (N.S.), 103(145)(3) :392–403, 464, 1977.

[200] Y. L. Pavlov. Random forests. In Probabilistic methods in discrete mathematics (Petrozavodsk, 1996), pages 11–18. VSP, Utrecht, 1997.

[201] V. V. Petrov. Generalization of Cramér's limit theorem. Uspehi Matem. Nauk (N.S.), 9(4(62)) :195–202, 1954.

[202] I. Pinelis and R. Molzon. Berry-esseen bounds for general nonlinear statistics, with applications to pearson's and non-central student's and hotelling's. Arxiv preprint arXiv :0906.0177v3, 2012.

[203] J. Pitman and M. Yor. On the distribution of ranked heights of excursions of a Brownian bridge. Ann. Probab., 29(1) :361–384, 2001.

[204] D. Plachky and J. Steinebach. A theorem about probabilities of large deviations with an application to queuing theory. Period. Math. Hungar., 6(4) :343–345, 1975.

[205] H. Putter and G. A. Young. On the effect of covariance function estimation on the accuracy of Kriging predictors. Bernoulli, 7(3) :421–438, 2001.

[206] J.-R. Pycke. Multivariate extensions of the Anderson-Darling process. Statist. Probab. Lett., 63(4) :387–399, 2003.

[207] M. P. Quine and J. Robinson. A Berry-Esseen bound for an occupancy problem. Ann. Probab., 10(3) :663–671, 1982.

[208] R Development Core Team. R : A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2011. ISBN 3-900051-07-0.

[209] N. Rachdi, J.-C. Fort, and T. Klein. Stochastic inverse problem with noisy simulator-application to aeronautical model. Annales de la Faculté des Sciences de Toulouse, 6, 21 :593–622, 2012.

[210] C. Rasmussen and C. Williams. Gaussian Processes for Machine Learning. The MIT Press, Cambridge, 2006.

[211] G. Reinert and M. Waterman. On the length of the longest exact position match in a random sequence. IEEE/ACM Trans Comput Biol Bioinform, 4(1) :153–156, 2007.

[212] D. Revuz and M. Yor. Continuous martingales and Brownian motion, volume 293 of Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Springer-Verlag, Berlin, third edition, 1999.

[213] J. Riihimäki and A. Vehtari. Gaussian processes with monotonicity information. In Journal of Machine Learning Research : Workshop and Conference Proceedings, volume 9, pages 645–652, 2010.

[214] J. Robinson, T. Höglund, L. Holst, and M. P. Quine. On approximating probabilities for small and large deviations in $\mathbf{R}^d$. Ann. Probab., 18(2) :727–753, 1990.

[215] M. Rosenblatt. Remark on some nonparametric estimates of a density function. Ann. Math. Statist., 27 :832–837, 1956.

[216] M. Rosenblatt. A quadratic measure of deviation of two-dimensional density estimates and a test of independence. Ann. Statist., 3 :1–14, 1975.

[217] A. Rouault. Large deviations for cascades and cascades of large deviations. In Mathematics and computer science. III, Trends Math., pages 351–362. Birkhäuser, Basel, 2004.

[218] O. Roustant, D. Ginsbourger, and Y. Deville. DiceKriging, DiceOptim : Two R packages for the analysis of computer experiments by Kriging-based metamodeling and optimization. Journal of Statistical Software, 51(1), 2012.

[219] B. Roynette, P. Vallois, and M. Yor. Penalisations of Brownian motion with its maximum and minimum processes as weak forms of Skorokhod embedding. Theory Stoch. Process., 14(2) :116–138, 2008.

[220] H. Rue and L. Held. Gaussian Markov random fields, Theory and applications. Chapman & Hall, 2005.

[221] D. Rullière, N. Durrande, F. Bachoc, and C. Chevalier. Nested Kriging predictions for datasets with a large number of observations. Statistics and Computing, 28(4) :849–867, 2018.

[222] J. Sacks, W. Welch, T. Mitchell, and H. Wynn. Design and analysis of computer experiments. Statistical Science, 4 :409–423, 1989.

[223] J. S. Sadowsky. On Monte Carlo estimation of large deviations probabilities. Ann. Appl. Probab., 6(2) :399–422, 1996.

[224] T. Santner, B. Williams, and W. Notz. The Design and Analysis of Computer Experiments. Springer, New York, 2003.

[225] T. J. Santner, B. Williams, and W. Notz. The Design and Analysis of Computer Experiments. Springer-Verlag, 2003.

[226] T. J. Santner, B. J. Williams, and W. I. Notz. The design and analysis of computer experiments. Springer Series in Statistics. Springer-Verlag, New York, 2003.

[227] R. Schaback. Mathematical results concerning kernel techniques. In Prep. 13th IFAC Symposium on System Identification, Rotterdam, pages 1814–1819. Citeseer, 2003.

[228] M. Scheuerer, R. Schaback, and M. Schlather. Interpolation of spatial data - a stochastic or a deterministic problem ? Preprint, Universität Göttingen. http://num.math.uni-goettingen.de/schaback/research/papers/IoSD.pdf, 2011.

[229] G. Schuëller, H. Pradlwarter, and M. Pandey. Methods for reliability assessment of nonlinear systems under stochastic dynamic loading - a review. pages 751–9. EURODYN'93, Balkema, 1993.

[230] C. Schwab and R. A. Todor. Karhunen Loève approximation of random fields by generalized fast multipole methods. Journal of Computational Physics, 217 :100–122, Sept. 2006.

[231] B. A. Shaby and D. Ruppert. Tapered covariance : Bayesian estimation and asymptotics. Journal of Computational and Graphical Statistics, 21(2) :433–452, 2012.

[232] N. V. Smirnov. On the probabilities of large deviations. Mat. Sb., 40 :443–454, 1933.

[233] E. Snelson, Z. Ghahramani, and C. E. Rasmussen. Warped gaussian processes. In S. Thrun, L. K. Saul, and B. Schölkopf, editors, Advances in Neural Information Processing Systems 16, pages 337–344. MIT Press, 2004.

[234] I. Sobol. Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. Mathematics and Computers in Simulation, 55(1-3) :271–280, 2001.

[235] I. M. Sobol. Sensitivity estimates for nonlinear mathematical models. Math. Modeling Comput. Experiment, 1(4) :407–414, 1993.

[236] G. P. Steck. Limit theorems for conditional distributions. Univ. California Publ. Statist., 2 :237–284, 1957.

[237] M. Stein. Asymptotically efficient prediction of a random field with a misspecified covariance function. The Annals of Statistics, 16 :55–63, 1988.

[238] M. Stein. Bounds on the efficiency of linear predictions using an incorrect covariance function. The Annals of Statistics, 18 :1116–1138, 1990.

[239] M. Stein. A comparison of generalized cross validation and modified maximum likelihood for estimating the parameters of a stochastic process. The Annals of Statistics, 18 :1139–1157, 1990.

[240] M. Stein. Uniform asymptotic optimality of linear predictions of a random field using an incorrect second-order structure. The Annals of Statistics, 18 :850–872, 1990.

[241] M. Stein. Interpolation of Spatial Data : Some Theory for Kriging. Springer, New York, 1999.

[242] M. L. Stein. Limitations on low rank approximations for covariance matrices of spatial data. Spatial Statistics, 8 :1–19, 2014.

[243] M. L. Stein, Z. Chi, and L. J. Welty. Approximating likelihoods for large spatial data sets. Journal of the Royal Statistical Society : Series B (Statistical Methodology), 66(2) :275–296, 2004.

[244] H. Takahata and K. I. Yoshihara. Central limit theorems for integrated square error of nonparametric density estimators based on absolutely regular random sequences. Yokohama Math. J., 35 :95–111, 1987.

[245] M.-P. Étienne. Le score local : un outil pour l'analyse de séquences biologiques. PhD thesis, 2002. Thèse de doctorat dirigée par Vallois, Pierre Mathématiques appliquées Nancy 1 2002.

[246] V. Tresp. A Bayesian committee machine. Neural Computation, 12(11) :2719–2741, 2000.

[247] J. Valeinis and A. Locmelis. Bickel-Rosenblatt test for weakly dependent data. Math. Model. Anal., 17 :383–395, 2012.

[248] J. M. van Campenhout and T. M. Cover. Maximum entropy and conditional probability. IEEE Trans. Inform. Theory, 27(4) :483–489, 1981.

[249] A. W. van der Vaart. Asymptotic statistics, volume 3 of Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 1998.

[250] B. van Stein, H. Wang, W. Kowalczyk, T. Bäck, and M. Emmerich. Optimally weighted cluster kriging for big data regression. In E. Fromont, T. De Bie, and M. van Leeuwen, editors, Advances in Intelligent Data Analysis XIV, pages 310–321, Cham, 2015. Springer International Publishing.

[251] C. Varin, N. Reid, and D. Firth. An overview of composite likelihood methods. Statistica Sinica, 21 :5–42, 2011.

[252] A. V. Vecchia. Estimation and model identification for continuous spatial processes. Journal of the Royal Statistical Society : Series B (Methodological), 50(2) :297–312, 1988.

[253] M. Villén-Altamirano and J. Villén-Altamirano. Accelerated simulation of rare events using restart method with hysteresis. pages 675–686. Elsevier, Amsterdam, Netherlands, 1991.

[254] M. Villén-Altamirano and J. Villén-Altamirano. Restart : a method for accelerating rare event simulations. pages 71–76. North-Holland, 1991.

[255] D. Wang and W.-L. Loh. On fixed-domain asymptotics and covariance tapering in Gaussian random field models. Electronic Journal of Statistics, 5 :238–269, 2011.

[256] M. Waterman, L. Gordon, and R. Arratia. Phase transition in sequence matched and nucleic acid structure. PNAS, 84 :1239–1243, 1987.

[257] M. S. Waterman. Introduction to Computational Biology : Maps, Sequences and Genomes. Chapman & Hall, 1995.

[258] O. Weber and C. Gotsman. Controllable conformal maps for shape deformation and interpolation. ACM Trans. Graph., 29(4) :78 :1–78 :11, July 2010.

[259] J. G. Wendel. Left-continuous random walk and the Lagrange expansion. Amer. Math. Monthly, 82 :494–499, 1975.

[260] I. Wolfram Research. Mathematica, Version 10, 2015.

[261] L. M. Wu. Large deviations, moderate deviations and LIL for empirical processes. Ann. Probab., 22(1) :17–27, 1994.

[262] Z. Ying. Asymptotic properties of a maximum likelihood estimator with data from a Gaussian process. Journal of Multivariate Analysis, 36 :280–296, 1991.

[263] O. Zahm, P. Constantine, C. Prieur, and Y. Marzouk. Gradient-based dimension reduction of multivariate vector-valued functions. arXiv e-prints, page arXiv :1801.07922, Jan 2018.

[264] H. Zhang. Inconsistent estimation and asymptotically equivalent interpolations in model-based geostatistics. Journal of the American Statistical Association, 99 :250–261, 2004.

[265] H. Zhang and Y. Wang. Kriging and cross validation for massive spatial data. Environmetrics, 21 :290–304, 2010.

[266] H. Zhang and D. Zimmerman. Towards reconciling two asymptotic frameworks in spatial statistics. Biometrika, 92 :921–936, 2005.

[267] P. Zhang. Nonparametric importance sampling. J. Amer. Statist. Assoc., 91(435) :1245–1253, 1996.