Sur la décomposition ANOVA et l'estimation des indices de Sobol'. Application à un modèle d'écosystème marin.

Jean-Yves Tissot

UNIVERSITÉ DE GRENOBLE

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

Spécialité : **Mathématiques appliquées**

Arrêté ministériel : du 7 août 2006

Présentée par

Jean-Yves Tissot

Thèse dirigée par **Clémentine Prieur et Éric Blayo**

préparée au sein **du Laboratoire Jean Kuntzmann**
et de **l'école doctorale Mathématiques, Sciences et Technologies de l'Information, Informatique**

Sur la décomposition ANOVA et l'estimation des indices de Sobol'. Application à un modèle d'écosystème marin.

Thèse soutenue publiquement le 16 novembre 2012
devant le jury composé de :

Éric Blayo
Professeur, Université Joseph Fourier/LJK, Directeur de thèse
Pierre Brasseur
Directeur de recherche, CNRS/LEGI, Examinateur
Fabrice Gamboa
Professeur, Université Paul Sabatier/IMT, Président
Céline Helbert
Maître de conférence, École Centrale de Lyon, Examinatrice
Hervé Monod
Directeur de recherche, INRA, Rapporteur
Art B. Owen
Professeur, Stanford University, Rapporteur
Clémentine Prieur
Professeur, Université Joseph Fourier/LJK, Directrice de thèse
Bruno Sudret
Professeur, ETH Zürich, Examinateur

« You can't wait for inspiration, you have to go after it with a club. »
Jack London

Remerciements

D'abord, je remercie Clémentine et Éric de m'avoir accompagné et guidé tout au long de ma thèse. Merci à tous les deux pour tout ce que vous m'avez enseigné et pour la grande liberté que vous m'avez laissé dans mon travail. Clémentine, merci pour ton investissement dans mes travaux théoriques ; ta curiosité sans borne et ta rigueur mathématique ont été un stimulant essentiel durant ces trois années. Éric, merci d'avoir gardé la main sur les aspects applicatifs et de m'avoir constamment encouragé à ne pas négliger les questions concrètes.

Je tiens ensuite à saluer la contribution des deux rapporteurs de ma thèse, Hervé Monod et Art B. Owen. Merci à tous les deux pour vos précieuses remarques et suggestions qui m'ont permis d'améliorer mon manuscrit et de prendre plus de recul sur mon travail. Hervé, merci également pour toutes les lectures que tu m'as suggéré et tous les conseils que tu m'as prodigué en matière de plans factoriels. Je suis également reconnaissant à Pierre Brasseur, Fabrice Gamboa, Céline Helbert et Bruno Sudret d'avoir accepté de participer à mon jury de thèse en tant qu'examinateurs.

Plus généralement, je tiens à exprimer ma gratitude à toutes les personnes qui m'ont fait avancer dans ma recherche. Merci en particulier aux animateurs du GdR MASCOT-NUM, de l'ANR Costa Brava et du réseau MEXICO qui m'ont permis, comme à beaucoup d'autres, de travailler dans une certaine émulation en échangeant avec de nombreuses personnes.

Même si j'ai côtoyé dans cette équipe certains beaucoup plus que d'autres, je tiens à remercier l'ensemble des membres de l'équipe MOISE dans laquelle j'ai travaillé toutes ces années. Merci en particulier à celle qui finirait par me faire oublier les lourdeurs administratives : Anne, merci pour ta gentillesse, ta disponibilité et pour tout le temps que tu m'as fait gagner.

Enfin, je tiens à adresser mes remerciements les plus chaleureux à l'ensemble de mes proches, mes parents, mes frères et mes amis. Merci pour votre soutien durant ces trois ans. Et comme la vie ne se résume pas aux mathématiques, merci à tout ceux qui m'ont permis de m'évader dans les nuits grenobloises, parisiennes, londoniennes, madrilènes, entre Bamako et Ouagadougou, Montréal et Chicoutimi, Manzanarès et Salamanque, à la Dent de Crolles, à la Chamechaude, au Triglav, à la Dent de Morcles, au Mont de Grange ... Et peut-être un jour à Katmandou si le Grand Som ne s'en mêle pas.

Table des matières

Notations

À l'exception des ensembles notés $\mathfrak{u}$ dans l'indexation des indices de Sobol', tout ensemble est noté par une lettre majuscule. Tout vecteur noté sous forme développée $(x_1, \ldots, x_d)$ se voit implicitement attribuer une notation contractée en gras $\mathbf{x}$. De plus, pour tout sous-ensemble non-vide $\mathfrak{u} \subseteq \{1, \ldots, d\}$, la notation $\mathbf{x}_\mathfrak{u}$ désigne le vecteur dont les composantes sont les x_i, $i \in \mathfrak{u}$, et la notation $\mathbf{x}_\mathfrak{u} : \mathbf{z}_{\mathfrak{u}^c}$ désigne le vecteur dont les composantes sont les x_i pour $i \in \mathfrak{u}$ et les z_i pour $i \in \{1, \ldots, d\} \setminus \mathfrak{u}$. Pour désigner une famille de n vecteurs ou de n scalaires, on utilise la notation avec un exposant : $\mathbf{x}^j$, $\mathbf{x}_\mathfrak{u}^j$, et x_i^j, $j \in \{1, \ldots, n\}$, $\mathfrak{u} \subseteq \{1, \ldots, d\}$ et $i \in \{1, \ldots, d\}$. Sauf mention contraire, on adopte les notations suivantes

$\lvert E \rvert$	cardinal de l'ensemble E
$\lvert a \rvert$	valeur absolue du nombre a
$\mathcal{P}(E)$	ensemble des parties d'un ensemble E
$\sigma(X), \sigma(\mathcal{A})$	tribu engendrée par une variable aléatoire X, ou une tribu $\mathcal{A}$
$[1:d]$	ensemble des nombres entiers i tels que $1 \leq i \leq d$
$\mathbf{1}_E$	fonction indicatrice de l'ensemble E : $\mathbf{1}_E(x) = 1$ si $x \in E$, et 0 sinon
$E_1 \Delta E_2$	différence symétrique des deux ensembles E_1 et E_2. C'est l'ensemble des éléments de E_1 ou E_2 qui ne sont pas dans E_1 et E_2
$\mathbb{E}[Y]$	espérance de la variable aléatoire Y
$\mathrm{Var}[Y]$	variance de la variable aléatoire Y
$\mathrm{Cov}(X, Y)$	covariance entre les variables aléatoires X et Y
$M^\intercal$	transposée de la matrice M
$\lfloor x \rfloor$	partie entière du nombre réel x
$E \setminus A$	partie complémentaire de l'ensemble A dans l'ensemble E
A^c, $-A$	notations contractées de la notation précédente en l'absence de toute ambiguïté concernant l'ensemble E
$A \subseteq B$	signifie que l'ensemble A est inclus dans, ou égal à l'ensemble B
$A \subset B$	signifie que l'ensemble A est inclus dans, mais n'est pas égal à l'ensemble B
$\int_E f(\mathbf{x}) \mathrm{d}\mathbf{x}$	intégrale de f, au sens de Lebesgue, sur l'ensemble E
$\int f(\mathbf{x}) \mathrm{d}\mathbf{x}$	intégrale de f, au sens de Lebesgue, sur l'hypercube unité dont la dimension est le nombre de variables de f
$\mathbf{x} \cdot \mathbf{y}$	produit scalaire canonique des vecteurs $\mathbf{x}$ et $\mathbf{y}$ de $\mathbb{R}^k$, $k \in \mathbb{N}^*$
$f \circ g$	fonction composée des fonctions f et g

Les deux notations qui suivent peuvent induire une ambiguïté avec la notation ensembliste d'un singleton ; elles seront donc précisées à chaque utilisation par la suite :

$\{x\}$	partie fractionnaire du réel x
$\{\mathbf{x}\}$	vecteur dont les composantes sont les parties fractionnaires des composantes du vecteur $\mathbf{x}$
$\mathrm{supp}(X)$	support de la variable aléatoire X
$\mathbb{C}^*, \mathbb{R}^*, \mathbb{Q}^*, \mathbb{Z}^*, \mathbb{N}^*$	respectivement les ensembles des nombres complexes, réels, rationnels, entiers et entiers naturels, privés du singleton $\{0\}$

Introduction

Introduction

Contexte applicatif de la thèse

Du fait du développement continu des capacités de calcul en informatique, la modélisation numérique de systèmes complexes (physiques, mécaniques, biologiques, etc.) a connu une expansion ininterrompue durant les dernières décennies, et devient aujourd'hui une composante incontournable dans l'étude des phénomènes réels. Ainsi le développement de pneumatiques dans l'industrie automobile, la gestion des risques dans l'industrie nucléaire, la mise en œuvre de politiques de pêches pour la gestion des ressources halieutiques ou encore la prévision météorologique requièrent à l'heure actuelle l'utilisation de modèle numériques complexes — encore appelés *simulateurs* — et l'aide de moyens de calculs adéquats.

Construction d'un simulateur

La construction d'un simulateur se compose principalement d'une phase d'expertise consistant à modéliser le phénomène réel d'intérêt, et une phase technique se résumant successivement à discrétiser puis à implémenter le modèle élaboré dans la première phase. Le processus de modélisation est dévolu à l'expert de la branche d'intérêt (ingénieur en calcul de structures, biologiste spécialiste des écosystèmes marins, etc.) et consiste à identifier les *variables d'état* du système à modéliser, à déterminer leurs lois d'évolution (en temps et en espace), et enfin à établir les degrés de liberté dans ces lois, en introduisant un certain nombre de *paramètres* dont la valeur permettra à terme d'ajuster le modèle afin qu'il reproduise le plus fidèlement possible le phénomène réel d'intérêt. Dans la pratique, les lois d'évolution sont établies plus ou moins empiriquement suivant le domaine d'étude. Par exemple, les modèles développés en mécanique des fluides (météorologie, modélisation des circulations océaniques, etc.) sont issus d'un système d'équations aux dérivées partielles non-linéaires, connu sous le nom des *équations de Navier-Stokes*. Ils permettent de reproduire fidèlement de nombreux phénomènes à des échelles plus ou moins fines. Dans d'autres domaines, ces lois bénéficient d'une connaissance théorique moins étendue et doivent faire appel à l'expérience, et ainsi faire face à toute l'incertitude que cela engendre. Pour exemple, dans le modèle d'écosystème marin MODECOGeL [Lac98] étudié au Chapitre 8 de cette thèse, la concentration en nitrate est donnée par l'équation de diffusion

$$\frac{\partial \mathtt{no3}}{\partial t} = \frac{\partial}{\partial z}\left(\lambda \frac{\partial \mathtt{no3}}{\partial z}\right) + nitr_{\mathtt{nh4}} \times \mathtt{nh4} - \mu_{\mathtt{no3pp}} \times \mathtt{pp} - \mu_{\mathtt{no3np}} \times \mathtt{np} - \mu_{\mathtt{no3mp}} \times \mathtt{mp}$$

où `no3`, `nh4`, `pp`, `np` et `mp` désignent respectivement les variables d'état de concentration en nitrate, en ammonium, en pico-, nano, microphytoplancton — fonctions du temps t et de l'espace z (verticale) — et λ, $nitr_{\mathtt{nh4}}$, $\mu_{\mathtt{no3pp}}$, $\mu_{\mathtt{no3np}}$ et $\mu_{\mathtt{no3mp}}$ sont des coefficients dont les valeurs sont établies par des sous-modèles partiellement méconnus faisant intervenir parfois plus d'une dizaine de paramètres différents (voir Croissance du phytoplancton dans la Section 8.1.5 de cette thèse). De manière plus appliquée, parmi l'information globale que renferme un modèle, seules certaines quantités présentent un réel intérêt (usure d'un pneu, vitesse du vent, quantité de précipitations, etc.), chacune d'entre elles devenant une *sortie du modèle* sous la forme d'une fonction des variables d'état et des paramètres

$$y(t,z) = f\big(\mathbf{x}(t,z), \boldsymbol{\theta}\big)$$

où t est la variable temporelle, z est la variable spatiale — éventuellement vectorielle — $\mathbf{x}(t,z)$ est le vecteur formé des variables d'état, et $\boldsymbol{\theta}$ désigne le vecteur des paramètres. Remarquons qu'un même

simulateur peut permettre d'étudier plusieurs sorties d'intérêt simultanément. Par exemple, dans l'étude préliminaire que nous consacrons au modèle MODECOGeL, on s'intéresse principalement à la concentration de chlorophylle-a qui, de par sa définition, dépend directement des trois variables d'état phytoplanctoniques

$$\texttt{chla}(t,z) = 1.59 \times \big(\texttt{pp}(t,z) + \texttt{np}(t,z) + \texttt{mp}(t,z)\big).$$

En particulier, nous privilégions des sorties scalaires en considérant le maximum annuel de cette concentration à la surface de l'océan, sa moyenne annuelle de surface et son maximum annuel de la moyenne en profondeur (de -20m à -50m).

Enfin, le modèle est discrétisé suivant des schémas adaptés (différences finies, volumes finis, etc.) puis implémenté dans un langage bas niveau de préférence afin de privilégier les performances. Ainsi, lorsqu'un utilisateur prend en main un simulateur, il doit obligatoirement assumer le fait que les données qu'il simule forment une double approximation de la réalité, prenant en compte à la fois une erreur conceptuelle de modélisation et une erreur technique de discrétisation.

Niveau de complexité d'un simulateur

Lorsque l'erreur conceptuelle commise lors de la modélisation apparaît comme faible, la minimisation de l'erreur technique de discrétisation peut constituer le principal challenge dans la construction d'un simulateur. Néanmoins, lorsque la phase de modélisation est extrêmement difficile, et qu'aucun simulateur ne donne des résultats satisfaisants quant à la reproduction fidèle des phénomènes réels qu'il modélise, la recherche s'oriente principalement sur la minimisation de l'erreur conceptuelle. Cela a conduit les modélisateurs à développer des simulateurs de plus en plus complexes, prenant en compte un nombre toujours plus grand de variables d'état et de paramètres. Ce phénomène apparaît en particulier dans le domaine de la biochimie dans lequel les modèles décrivant les réactions chimiques sont difficiles à établir. Néanmoins la complexification croissante des simulateurs peut montrer très rapidement des limites lorsque, du fait d'une surparamétrisation, un simulateur complexe devient impossible à ajuster par rapport aux données réelles, et reproduit moins fidèlement la réalité qu'un simulateur plus simple. En effet, le processus de *calibration* — ou d'*ajustement* — des paramètres, qui a pour but d'ajuster les valeurs des différents paramètres afin de minimiser l'écart entre le phénomène simulé et le phénomène réel lui-même, consiste essentiellement en un problème de minimisation d'une fonction coût dépendant de tous les paramètres. En conséquence, plus le nombre de paramètres augmente, et plus il est difficile d'établir une fonction coût qui permette de contrôler les différents paramètres. Ainsi voit-on émerger depuis quelques années un mouvement inverse allant vers la simplification de modèles complexes existants (voir, e.g., [RSG06] dans le domaine des écosystèmes marins). Le but principal dans cette problématique est de converger vers un simulateur "optimal", suffisamment complexe pour être capable de reproduire correctement les processus principaux du phénomène réel considéré, et assez simple pour conserver une certaine maîtrise sur l'ensemble de ses paramètres et être capable de les ajuster.

Dans cette optique, l'étude des paramètres, et en particulier l'analyse qualitative et quantitative de leur impact sur les sorties du simulateur peut constituer une aide précieuse ; c'est l'objet principal de l'*analyse de sensibilité*.

Analyse de sensibilité : entre rigueur mathématique et méthodes approximatives

L'analyse de sensibilité telle qu'elle est présentée dans les ouvrages modernes (voir, e.g., [SCS00]) et telle qu'elle est appliquée actuellement aux simulateurs numériques complexes remonte approximativement au début des années 1970 (voir, e.g., [Ham94] pour ces considérations historiques). Elle possède néanmoins une préhistoire antérieure au développement de l'informatique moderne, qui se situe au début du XXème siècle avec notamment les travaux de Fisher [Fis25, Fis35] et de ses contemporains sur l'*analyse de la variance*. L'objet principal dans ce domaine d'étude est typiquement un système d'entrées-sortie $y = f(\mathbf{x})$ où $\mathbf{x}$ est le vecteur d'entrée composé d'un certain nombre de

paramètres, y la sortie scalaire observée et f un modèle déterministe qui permet de faire correspondre à toute entrée $\mathbf{x}$ une sortie y. Le modèle f s'exprime idéalement sous la forme d'une fonction mathématique explicite ; mais généralement elle se trouve sous une forme implicite, de sorte que la variation de la réponse y en fonction des changements des valeurs de paramètres d'entrée n'est pas transparente. Ainsi, le but premier de l'analyse de sensibilité est de comprendre comment les variations des paramètres d'entrée influent sur la réponse y dans un modèle implicite. Cette étude s'oriente à la fois sur des aspects qualitatifs :

L'effet d'un paramètre sur la réponse y est-il monotone, linéaire, additif, etc. ?

Les paramètres x_i et x_j ont-ils un effet conjoint sur la réponse y ?

et sur des aspects quantitatifs :

L'effet du paramètre x_i est-il plus important que celui du paramètre x_j ? Est-il négligeable ?

Pour répondre à ces questions, il existe un panel de méthodes établies mathématiquement de manière plus ou moins rigoureuse.

Méthode d'analyse locale

L'approche la plus élémentaire pour analyser l'influence des paramètres d'une fonction est de comparer les variations de sa sortie lorsque les paramètres d'entrée sont perturbés un à un par un incrément infinitésimal autour de leurs valeurs nominales. Cela revient essentiellement à considérer les dérivées partielles

$$d_i = \frac{\partial f}{\partial x_i}(\mathbf{x}^0)$$

puis à comparer entre elles les valeurs renormalisées $|d_i x_i^0|$, afin de prendre en compte les différentes échelles des paramètres. Plus généralement, il est possible de comparer les taux d'accroissements relatifs à chacun des paramètres en considérant non plus un incrément infinitésimal, mais un terme de l'ordre de 10%–20% de la valeur nominale de chacun des paramètres considérés. Une analyse rapide conduit alors à déduire que plus la quantité $|d_i x_i^0|$ est grande et plus le paramètre x_i est actif sur la sortie y étudiée. Une telle analyse quantitative reste néanmoins limitée car d'une part, elle ne permet pas de détecter les effets conjoints de plusieurs paramètres, et d'autre part elle n'intègre aucune connaissance relative aux paramètres mise à part la donnée absolue de leur valeur nominale respective. Ce dernier point a pour conséquence que la comparaison opérée ne prend pas en compte la nuance entre un paramètre connu de manière précise — i.e. par exemple avec une erreur relative inférieure à 5% — et un paramètre dont la valeur est extrêmement difficile à obtenir précisément — i.e. par exemple avec une mesure qui varie du simple au double ou au triple. Par suite, cette analyse de sensibilité locale peut facilement produire des résultats sans lien avec la réalité du modèle.

Pour pallier cette insuffisance, il est nécessaire de s'orienter vers une analyse qui prend en compte les plages de variations supposées des paramètres.

Méthode d'analyse globale

Pour chaque i dans $[1 : d]$, considérons D_i l'intervalle de variation du paramètre x_i ; alors l'objet mathématique d'intérêt est la fonction multivariée $f : \mathbf{D} = D_1 \times \cdots \times D_d \to \mathbb{R}$. L'idée principale de l'analyse de sensibilité globale est d'appréhender cette fonction comme une superposition de fonctions $f_\mathfrak{u}$ ne dépendant que des x_i, $i \in \mathfrak{u}$, où $\mathfrak{u}$ parcourt l'ensemble des sous-ensembles de $\{1, \dots d\}$. Plus précisément, il s'agit de considérer que la fonction f est la somme d'*effets*, chacun dû à un groupe de variables particulier :

$$f(\mathbf{x}) = f_\emptyset + f_1(x_1) + \cdots + f_d(x_d) + f_{12}(x_1, x_2) + \cdots + f_{1\dots d}(x_1, \dots, x_d). \tag{1}$$

Une telle décomposition permet alors de produire une analyse de sensibilité qualitative en explicitant les *effets principaux* $f_i(x_i)$ ainsi que les effets d'interactions $f_\mathfrak{u}(\mathbf{x}_\mathfrak{u})$, $|\mathfrak{u}| \geq 2$, et une analyse de sensibilité quantitative en évaluant l'importance de ces effets par le biais d'une norme.

Néanmoins, pour une fonction f quelconque, il existe une infinité de ces décompositions. On peut par exemple retrancher une constante à l'un des termes et la rajouter à un autre. Plus généralement, étant donnée une norme, il est possible de construire une décomposition de f dont les termes possèdent une norme arbitrairement petite ou grande. Le théorème de Pythagore fournit alors une solution pour obtenir l'unicité de cette décomposition moyennant une contrainte. En effet, si f est de carré intégrable sur $\mathbf{D}$, alors la décomposition (1) implique que

$$||f(\mathbf{x})||_2^2 \leq ||f_\emptyset||_2^2 + ||f_1(x_1)||_2^2 + \cdots + ||f_d(x_d)||_2^2 + ||f_{12}(x_1, x_2)||_2^2 + \cdots + ||f_{1\ldots d}(x_1, \ldots, x_d)||_2^2$$

où $||g||_2^2 = \int_{\mathbf{D}} g^2(\mathbf{x})\, \mathrm{d}\mathbf{x}$, et que l'égalité n'a lieu que si les termes de la décomposition sont orthogonaux deux à deux. Une telle décomposition sous contraintes d'orthogonalité existe effectivement et est unique ; elle se généralise en outre à la décomposition de fonctions de variables aléatoires, et est connue sous le nom de *décomposition ANOVA*. Même si elle n'est pas canonique au sens où elle dépend d'une norme particulière, on retient néanmoins qu'elle a du sens puisqu'en termes physiques, elle est la décomposition d'énergie minimale. Elle constitue la base de l'analyse de sensibilité globale et permet, comme nous l'avons décrit dans le paragraphe précédent, de mettre en œuvre une analyse de sensibilité à la fois qualitative et quantitative.

Dans la pratique, les termes de la décomposition ANOVA s'expriment tous de manière explicite en fonction de f et par conséquent, il est possible de connaître explicitement les fonctions f_{u}, ou tout du moins de les approcher numériquement. Néanmoins, on se contente la plupart du temps de ne calculer que leur norme, permettant de définir les *indices de Sobol'* qui quantifient de manière synthétique l'influence des différents groupes de paramètres. Le calcul de ces indices, qui se résume essentiellement à un calcul d'intégrales, peut être abordé sous plusieurs angles, certains mathématiquement rigoureux et d'autres plus approximatifs où la notion d'erreur d'estimation est difficile à appréhender.

Méthode d'analyse locale "globalisée"

Enfin, l'insuffisance de la méthode locale basée sur le calcul des dérivées partielles d'ordre 1 — ou des taux d'accroissements — au point constitué des valeurs nominales des paramètres, peut être comblé en mettant en œuvre un calcul de ces dérivées non plus en un seul point, mais sur un échantillon à l'intérieur de la plage de variation $\mathbf{D}$ des paramètres. Le traitement statistique de l'ensemble des dérivées — i.e. principalement le calcul de la moyenne et de la variance empiriques — permet alors de détecter les paramètres qui n'ont aucun influence significative sur la sortie y, ceux dont l'influence est linéaire et additive, et les autres. Cependant, même "globalisée", l'analyse locale fournit une information moins riche que l'analyse globale. En particulier, elle est incapable de différencier non-linéarités et interactions, et sa validité reste dépendante de conditions de régularité.

Dans ce contexte, nous nous concentrons principalement sur la décomposition ANOVA qui constitue un objet mathématiquement rigoureux et pertinent pour l'analyse de sensibilité globale, et l'objectif de cette thèse est de proposer une relecture du problème de l'estimation des indices de Sobol' au seul regard de la théorie de l'intégration numérique. Dans cette démarche, en prenant appui sur des théories classiques et des développements plus récents, notre travail vise à établir rigoureusement les bases de méthodes existantes, à les développer, ainsi qu'à proposer de nouvelles pistes de réflexion dans le but d'optimiser l'estimation des indices de Sobol'. Dans ce cadre, le but ultime est de construire un estimateur bénéficiant d'hypothèses d'application très faibles et d'un rapport coût/précision très élevé quelle que soit la dimension — i.e. le nombre de paramètres — du système étudié.

Plan de la thèse

Dans une première partie, nous nous attachons à présenter en détail les notions d'analyse de sensibilité évoquées précédemment. En particulier, nous consacrons le Chapitre 1 à la revue de la décomposition ANOVA et à l'introduction essentielle des définitions d'indices de Sobol' et de dimensions effectives. Nous donnons ensuite, dans le Chapitre 2, un aperçu global, mais non exhaustif, des méthodes d'estimation des indices de Sobol'. Dans ce chapitre, nous prenons soin de ne pas évoquer les méthodes indirectes basées sur la construction préalable d'un métamodèle ; seules les méthodes

directes d'intégration numérique sont introduites. C'est l'occasion d'entrevoir les différentes problématiques des méthodes d'estimation des indices de Sobol', principalement : champs d'hypothèses, coût, et vitesse de convergence. Enfin, dans un troisième chapitre, nous revenons brièvement sur les méthodes que nous avons décrites comme locale et locale "globalisée". On y décrit notamment les *indices de sensibilité basés sur les dérivées* dont l'introduction récente a été motivée par un soucis de minimiser le coût dans la recherche des paramètres inactifs.

Dans une deuxième partie, nous regroupons nos contributions (publication et prépublications) au problème d'estimation des indices de Sobol'. Le Chapitre 4 aborde la problématique du biais dans les méthodes Random Balanced Design (RBD) et Random Balanced Design-Fast Amplitude Sensitivity Test (RBD-FAST). Ces méthodes sont construites à partir de plans d'expériences stratifiés et bénéficient d'un post-traitement basé essentiellement sur l'analyse harmonique. En conséquence, les estimateurs obtenus possèdent un biais qui, bien qu'il soit asymptotiquement nul, peut s'avérer problématique lorsque sa vitesse de décroissance est d'un ordre inférieur ou égal à la vitesse de convergence de l'estimateur lui-même. Dans ce chapitre, nous proposons, de manière heuristique, une méthode de correction de ce biais. Dans le Chapitre 5, nous reprenons de manière rigoureuse l'étude des méthodes FAST et RBD. Dans un souci de mieux comprendre ces méthodes (champs d'hypothèses, erreur d'estimation, biais, etc.), nous proposons une nouvelle introduction de celles-ci en remontant à l'approximation originelle — i.e. l'application approximative du théorème ergodique de Weyl (voir Section 5.3.1) — sur laquelle elles sont construites. Par suite, nous définissons FAST et RBD de manière alternative, mais néanmoins rigoureusement équivalente, à l'aide de notions classiques d'intégration numérique. Cela nous permet en particulier de préciser et d'établir des résultats quant à la question de l'erreur d'estimation, et également d'introduire de nouvelles variantes de ces méthodes. Enfin, au Chapitre 6, nous introduisons une nouvelle manière d'aborder le problème d'optimisation combinatoire dans la méthode de Sobol' pour réduire son coût. Dans ce cadre, pour tout $k \leq d$, où d désigne la dimension du modèle considéré, le calcul de tous les indices de Sobol' d'ordre k ne requiert que deux échantillons distincts.

Enfin, dans une troisième partie nous faisons figurer un certain nombre d'applications de ces développements au calcul des indices de Sobol'. Dans le Chapitre 7, nous regroupons des tests comparatifs entre plusieurs méthodes effectués sur des modèles analytiques. Le Chapitre 8 est quant à lui consacré à l'étude d'un modèle d'écosystème marin de près de 90 paramètres. Le travail proposé dans ce chapitre est essentiellement introductif et a vocation à être développé en vue de proposer une aide au développement, et en particulier à la calibration, de ce type de modèles.

Première partie

Analyse de sensibilité globale : un état de l'art

Chapitre 1

La décomposition ANOVA

L'*analyse de la variance* (ANOVA) telle qu'elle a été théorisée au fil du XXème siècle remonte essentiellement aux travaux de Fisher [Fis25, Fis35], et est étroitement liée aux notions d'*expérience factorielle* et d'*interaction* entre *facteurs*. Dans ce cadre, le but de l'expérimentateur consiste à évaluer, de manière non ambiguë, quels facteurs sont prépondérants dans l'expérience qu'il mène i.e. quel facteur par sa variation propre, ou quels facteurs par leur variation conjointe produisent la plus grande variation sur l'observation. Pour cela, il doit définir un modèle mathématique sur les observations qu'il relève afin de mettre en œuvre des tests statistiques. Plus précisément, un modèle utilisé dans une expérience à deux facteurs A et B peut par exemple se présenter sous la forme

$$Y(i,j) = f(i,j) + Z(i,j), \quad i \in I,\ j \in J$$

où I et J sont des ensembles finis décrivant les valeurs — ou *niveaux* — possibles respectives de A et de B, et $Y(i,j)$, $f(i,j)$ et $Z(i,j)$ sont respectivement la donnée observée, le modèle déterministe théorique reliant les deux facteurs à la donnée observée et l'erreur d'observation, lorsque les deux facteurs prennent respectivement pour valeurs i et j. Les erreurs d'observation sont généralement modélisées par des variables aléatoires, et le modèle dépend des hypothèses faites sur celles-ci. L'ANOVA consiste dans un premier temps à décomposer le modèle déterministe théorique f de façon non-équivoque comme une somme d'*effets factoriels*

$$f(i,j) = f_0 + f_A(i) + f_B(j) + f_{AB}(i,j), \quad i \in I,\ j \in J. \tag{1.1}$$

Puis, après avoir estimé ces effets factoriels, l'ANOVA se résume à tester leur importance, i.e. à évaluer par des tests d'hypothèses si les effets f_A, f_B et f_{AB} sont significatifs ou s'ils peuvent être considérés comme négligeables. Plus précisément ces tests ont pour but d'évaluer si les variances de deux effets différents sont significativement égales ou pas. Notons que ces tests factoriels réalisés dans le cadre de l'ANOVA discrète — i.e. I et J finis — ont été étendus à des facteurs continus par Antoniadis [Ant84]. Dans ce cas, le modèle sous-jacent devient

$$X(s,t) = f(s,t) + X_0(s,t), \quad s \in S,\ t \in T$$

où S et T sont des espaces métriques compacts, $X_0(s,t)$ un processus gaussien de covariance K sur $(S \times T) \times (S \times T)$, et f doit se trouver dans l'espace de Hilbert à noyau reproduisant de noyau K.

L'intérêt porté à l'ANOVA par le domaine de l'analyse de sensibilité globale ne tient toutefois pas à ces tests d'hypothèses, mais uniquement à la décomposition du modèle f en effets factoriels et à la quantification de l'importance relative de ceux-ci par le calcul d'indices de Sobol' — voir Section 1.2.3. Cette décomposition, communément appelée *décomposition ANOVA*, *functional analysis of variance* (FANOVA) ou encore *high-dimensional model representation-ANOVA* (HDMR-ANOVA), est le sujet unique de ce chapitre. En prenant pour origine la décomposition due à Nelder en 1965 [Nel65]— dont la décomposition de Tjur [Tju84] énoncée dans le Théorème 1.1 est une généralisation essentielle — et pour fin temporaire la décomposition généralisée à des variables aléatoires continues et dépendantes due à Hooker [Hoo07] et précisée par Chastaing et al. [CGP12] en 2012, énoncée dans le Théorème 1.3 — nous proposons une revue des contributions apportées à cette décomposition pendant les cinquante dernières années.

Le chapitre se compose de deux sections principales et d'une section de notes. Dans chacune des deux sections principales, la décomposition ANOVA est introduite sous un angle singulier, d'abord par l'algèbre combinatoire et ensuite par l'algèbre tensorielle. Bien que la structure de l'exposé mette en avant la dissymétrie entre ces approches, on constate que la frontière entre les deux est extrêmement perméable. Et finalement, au delà des dénominations multiples, et de l'inflation d'abréviations, ne se trouve qu'un seul objet mathématique essentiellement combinatoire qui est *la* décomposition ANOVA.

1.1 Décomposition ANOVA par l'algèbre combinatoire

Cette section a pour but d'introduire la décomposition ANOVA telle qu'elle a été énoncée par Tjur [Tju84], laquelle généralise une précédente décomposition proposée par Nelder [Nel65]. Dans le cadre de ces articles, la décomposition ANOVA apparaît comme un objet essentiellement combinatoire. L'exposé qui suit s'inspire principalement de la synthèse effectuée par Bailey sur l'utilisation des partitions orthogonales dans les plans d'expérience [Bai96] et des articles de Tjur [Tju84, Tju91]. La section se décompose en cinq sous-sections. Dans les quatre premières, nous introduisons successivement les notions de *facteur*, *partition* et *structure en blocs* — en anglais, *block structure*, introduit par Nelder [Nel65] — puis nous énonçons la décomposition ANOVA pour un espace sous-jacent fini, avant d'évoquer la notion d'*effet* des facteurs et leur quantification. Dans la cinquième sous-section, nous abordons la généralisation à un espace probabilisé comme elle a été introduite par Helland [Hel98]. Enfin nous concluons en énonçant deux décompositions ANOVA classiques relatives à des variables aléatoires. Nous présentons d'abord la décomposition ANOVA pour des variables aléatoires indépendantes dans sa version donnée par Efron et Stein [ES81]— issue des travaux de Hoeffding sur les U-statistiques [Hoe48] — comme un cas particulier de la généralisation de Helland, et ensuite la décomposition généralisée pour des variables aléatoires dépendantes dans sa version très récente due à Hooker [Hoo07] et précisée par Chastaing et al. [CGP12].

1.1.1 Facteur et partition

Soit Ω un ensemble fini ; nous commençons par donner la définition de facteur telle qu'elle est introduite par Tjur [Tju91].

Définition 1.1. (facteur) *Un facteur F est une application*

$$F : \Omega \longrightarrow L_F$$

où L_F est un ensemble fini de cardinal n_F représentant les niveaux *du facteur.*

Lorsque Ω se décompose comme un produit cartésien $\Omega_1 \times \cdots \times \Omega_d$, on retrouve naturellement la notion de facteur — ou paramètre, ou variable etc. — communément utilisée dans l'étude d'une fonction de plusieurs variables, en considérant pour $i \in [1 : d]$, les applications

$$\begin{array}{rccc} L_i : & \Omega & \longrightarrow & \Omega_i \\ & \omega = (\omega_1, \ldots, \omega_d) & \longmapsto & \omega_i \end{array} .$$

Cette notion de facteur qui peut sembler centrale pour l'énoncé de la décomposition ANOVA n'est toutefois pas retenue dans la construction qui suit. Plus précisément, un facteur F peut être vu comme la donnée conjointe de ses niveaux $l \in L_F$, et des sous-ensembles $\left(F^{-1}(l)\right)_{l \in L_F}$ qui forment une partition de Ω. Par la suite, la notion de facteur est appauvrie de la donnée des niveaux afin de ne conserver que la notion purement algébrique de partition suivant une relation d'équivalence.

Définition 1.2. (partition) *La partition associée à un facteur F est définie par la donnée de l'application*

$$\begin{array}{rccc} C_F : & \Omega & \longrightarrow & \mathcal{P}(\Omega) \\ & \omega & \longmapsto & \{\alpha \in \Omega \mid F(\alpha) = F(\omega)\} \end{array}$$

associant à chaque élément de Ω sa classe d'équivalence suivant le facteur F.

Dans ce qui suit, la partition C_F associée à un facteur F sera simplement notée F, l'absence de notion de facteur par la suite levant toute ambiguïté. Les classes d'équivalence suivant F sont appelées F-classes.

Une partition est dite *uniforme* — ou équilibrée, ou régulière — dès lors que ses classes d'équivalence ont même cardinal. Si F est uniforme, on note k_F le cardinal commun de ses F-classes. Il existe deux partitions uniformes triviales quel que soit Ω ; la partition U dans laquelle la seule classe d'équivalence est Ω est définie par

$$U(\omega) = \Omega \quad \text{pour } \omega \text{ dans } \Omega,$$

et la partition E dans laquelle tout singleton est une classe d'équivalence est donnée par

$$E(\omega) = \{\omega\} \quad \text{pour } \omega \text{ dans } \Omega.$$

On définit en outre les notions duales d'*infimum* et de *supremum* de deux partitions en ayant au préalable introduit un ordre partiel sur l'ensemble des partitions [1].

Définition 1.3. (ordre partiel) *Soient F et G deux partitions de Ω. Si chacune des F-classes est contenue dans une G-classe, F est dite plus* fine *que G, ou G plus* grossière *que F. Si F est plus fine que G on notera $G \preceq F$, et si F est strictement plus fine que G — i.e. F est plus fine que G et G n'est pas plus fine que F — on notera $G \prec F$.*

Définition 1.4. (infimum, supremum) *Soient F et G deux partitions de Ω, on appelle infimum de F et G, et on note $F \wedge G$, la plus fine des partitions plus grossières que F et G. Le supremum de F et G, noté $F \vee G$, est quant à lui défini comme la plus grossière des partitions plus fines que F et G. Ces deux opérations sont clairement associatives et commutatives.*

1.1.2 Espace fonctionnel et sous-espaces associés à une partition

Notons $\mathbb{R}^\Omega$ l'espace vectoriel de dimension $|\Omega|$ des fonctions à valeurs réelles définies sur Ω. Muni du produit scalaire $\langle\ ,\ \rangle$ donné par

$$\langle f, g\rangle = \frac{1}{|\Omega|}\sum_{\omega\in\Omega} f(\omega)g(\omega)$$

l'espace $\mathbb{R}^\Omega$ possède une structure d'espace vectoriel euclidien. Par la suite, après avoir introduit la notion de sous-espace vectoriel (s.e.v.) de $\mathbb{R}^\Omega$ associé à une partition, on définit les projections orthogonales sur ces s.e.v.

Définition 1.5. (s.e.v. associé à une partition) *Soit F une partition de Ω ; on définit le s.e.v. V_F de $\mathbb{R}^\Omega$ associé à F par*

$$V_F = \{f \in \mathbb{R}^\Omega \mid \forall\omega,\ \omega' \in \Omega,\ F(\omega) = F(\omega')\ \Rightarrow\ f(\omega) = f(\omega')\}.$$

En notant trivialement que les fonctions caractéristiques des F-classes forment une base de V_F, on déduit que $\dim V_F = n_F$.

On note $W \leq V$ pour signifier que W est un s.e.v. de l'espace vectoriel V. L'ensemble des sous-espaces vectoriels de $\mathbb{R}^\Omega$ associés à des partitions de Ω peut être muni trivialement d'une relation d'ordre partiel induite par celle définie sur l'ensemble des partitions de Ω — voir Définition 1.3.

Lemme 1.1. *Si $G \preceq F$ alors $V_G \leq V_F$.*

En outre, on note $V_F^\perp$ le s.e.v. de $\mathbb{R}^\Omega$ orthogonal à V_F défini par

$$V_F^\perp = \{f \in \mathbb{R}^\Omega,\ \mid \forall g \in V_F,\ \langle f, g\rangle = 0\}$$

et P_F la projection orthogonale sur V_F qui à tout élément de $\mathbb{R}^\Omega$, avec $f = f_1 + f_2$, $f_1 \in V_F$ et $f_2 \in V_F^\perp$, associe f_1. Par la suite, la composée entre deux projections P_F et P_G sera simplement notée $P_F P_G$. On aboutit alors à la notion fondamentale d'orthogonalité entre partitions.

1. On prendra garde en manipulant ces notions car la définition de l'ordre partiel ainsi que les définitions de supremum et d'infimum sont parfois inversées comme on le constate dans [Tju91] et [Bai96]. Les définitions que l'on donne sont celles de Tjur [Tju91].

Définition 1.6. (partitions orthogonales) *Deux partitions F et G sont dites orthogonales si $V_F \cap V_{F\wedge G}^{\perp}$ et $V_G \cap V_{F\wedge G}^{\perp}$ sont des s.e.v. orthogonaux.*

En particulier, une partition est orthogonale à elle même, et si une partition F est strictement plus fine ou plus grossière que G alors F et G sont orthogonales. On termine par le résultat élémentaire suivant qui lie l'orthogonalité entre partitions à la commutativité entre projections associées.

Lemme 1.2. *Deux partitions F et G de Ω sont orthogonales si et seulement si*

$$P_F P_G = P_G P_F = P_{F\wedge G}. \tag{1.2}$$

Démonstration. Soit $v \in \mathbb{R}^\Omega$; on a $P_G P_F(v) = P_G(v_F)$ où on peut décomposer v_F par

$$v_F = P_{F\wedge G}(v_F) + v' \quad \text{avec } v' \in V_F \cap V_{F\wedge G}^{\perp}.$$

On a bien entendu que pour tout $u \in V_G \cap V_{F\wedge G}$, $\langle v', u\rangle = 0$, et si F et G sont orthogonales, on a par la Définition 1.6 que pour tout $u \in V_G \cap V_{F\wedge G}^{\perp}$, $\langle v', u\rangle = 0$. Par suite on déduit que $v' \in V_G^{\perp}$ et on déduit

$$P_G(v_F) = P_G P_{F\wedge G}(v_F) = P_{F\wedge G}(v_F) = P_{F\wedge G} P_F(v).$$

Finalement on a montré que $P_G P_F = P_{F\wedge G} P_F$ et donc que $P_G P_F = P_{F\wedge G}$. La conclusion de l'implication dans le sens direct suit par le rôle symétrique de F et G. L'implication réciproque est immédiate en remarquant que la Formule (1.2) implique que $V_G \cap V_{F\wedge G}^{\perp} \subseteq V_F^{\perp}$. □

1.1.3 Structure en blocs et décomposition de Tjur

Une *structure en blocs* est simplement définie comme un ensemble structuré par une famille de partitions.

Définition 1.7. (structure en blocs) *Une structure en blocs est un couple $(\Omega, \mathfrak{F})$ où Ω est un ensemble et $\mathfrak{F}$ une famille de partitions de Ω.*

Une telle structure dépend naturellement des propriétés vérifiées par l'ensemble des partitions $\mathfrak{F}$. Les conditions minimales sur $\mathfrak{F}$ nécessaires à l'énoncé de la décomposition ANOVA ont été données par Tjur [Tju84]. Elles sont au nombre de trois, et une structure en blocs qui vérifie ces conditions est appelée *structure en blocs de Tjur* — en anglais, *Tjur block structure*.

Définition 1.8. (structure en blocs de Tjur) *Une structure en blocs de Tjur est une structure en blocs $(\Omega, \mathfrak{F})$ dans laquelle la famille de partitions $\mathfrak{F}$ vérifie*

(T1) *la partition la plus fine — i.e. E — est dans $\mathfrak{F}$,*
(T2) *$\mathfrak{F}$ est stable pour l'infimum,*
(T3) *les partitions de $\mathfrak{F}$ sont orthogonales deux à deux.*

Finalement, dans ce cadre élémentaire, on peut énoncer la décomposition ANOVA.

Théorème 1.1. [Tjur, 1984] *Soit $(\Omega, \mathfrak{F})$ une structure en blocs de Tjur, alors il existe une unique décomposition de $\mathbb{R}^\Omega$ en somme directe orthogonale*

$$\mathbb{R}^\Omega = \bigoplus_{F\in\mathfrak{F}}^{\perp} W_F \tag{1.3}$$

telle que pour tout $F \in \mathfrak{F}$,

$$V_F = \bigoplus_{G\in\mathfrak{F},\ G\preceq F}^{\perp} W_G,$$

où les V_F sont les s.e.v. de $\mathbb{R}^\Omega$ associés aux partitions $F \in \mathfrak{F}$.

Nous reproduisons dans ses grandes lignes la preuve originale de Tjur — voir [Tju84] pages 42–44 — d'une part parce qu'elle permet de comprendre par quels mécanismes combinatoires on aboutit à la décomposition ANOVA, et d'autre part parce qu'elle introduit des objets utiles par la suite.

Démonstration. **(Existence)** D'abord on a l'égalité

$$I = \prod_{F \in \mathfrak{F}} \Big(P_F + (I - P_F) \Big)$$

où I est l'opérateur identité sur $\mathbb{R}^\Omega$. Puis en développant le terme de droite, il vient

$$I = \sum_{\mathfrak{G} \subseteq \mathfrak{F}} Q_{\mathfrak{G}} \tag{1.4}$$

où chacun des $Q_{\mathfrak{G}}$ est défini par

$$Q_{\mathfrak{G}} = \Big(\prod_{F \in \mathfrak{G}} P_F \Big) \Big(\prod_{F \in \mathfrak{F} \setminus \mathfrak{G}} (I - P_F) \Big).$$

Notons alors que les opérateurs $I - P_F$, $F \in \mathfrak{F}$, sont les projecteurs orthogonaux sur $V_F^\perp$, et comme les projections P_F, $F \in \mathfrak{F}$, commutent deux à deux — par (T3) dans la Définition 1.8 et le Lemme 1.2 — on déduit que les $Q_{\mathfrak{G}}$, $\mathfrak{G} \subseteq \mathfrak{F}$, sont des projecteurs orthogonaux commutant deux à deux également. Définissons alors pour tout $\mathfrak{G} \subseteq \mathfrak{F}$ les sous-espaces vectoriels $W_{\mathfrak{G}} = \{Q_{\mathfrak{G}}(f),\ f \in \mathbb{R}^\Omega\}$ et notons qu'ils sont orthogonaux deux à deux du fait que

$$\mathfrak{G}_1 \neq \mathfrak{G}_2 \Rightarrow Q_{\mathfrak{G}_1} Q_{\mathfrak{G}_2} = 0.$$

On déduit alors de la Formule (1.4) la décomposition en somme directe orthogonale

$$\mathbb{R}^\Omega = \bigoplus_{\mathfrak{G} \subseteq \mathfrak{F}}^{\perp} W_{\mathfrak{G}}. \tag{1.5}$$

Néanmoins, la majorité des sous-espaces $W_{\mathfrak{G}}$, $\mathfrak{G} \subseteq \mathfrak{F}$, sont triviaux. En effet, soit $\mathfrak{G} \subseteq \mathfrak{F}$, on a

(a) si F et F' sont deux partitions de Ω telles que $F \in \mathfrak{G}$, $F' \notin \mathfrak{G}$ et $F \preceq F'$, alors $Q_{\mathfrak{G}} = 0$ car il contient $P_F(I - P_{F'})$ qui s'annule dès lors que $F \preceq F'$; et par suite $W_{\mathfrak{G}} = \{0\}$

(b) si $G = \bigwedge_{F \in \mathfrak{G}} F$ n'appartient pas à $\mathfrak{G}$ alors il appartient néanmoins à $\mathfrak{F}$ — d'après (T2) dans la Définition 1.8 — et on conclut que $Q_{\mathfrak{G}} = 0$ car il contient le produit de $(I - P_G)$ et de $\prod_{F \in \mathfrak{G}} P_F = P_G$

(c) en revenant à la définition des $Q_{\mathfrak{G}}$ et en notant que par (T1) dans la définition des structures en blocs de Tjur, la partition en singletons E est dans $\mathfrak{F}$, on aboutit aisément à $Q_\emptyset = 0$

Par suite, seuls les $\mathfrak{G} \subseteq \mathfrak{F}$ de la forme $\{F \in \mathfrak{F},\ G \preceq F\}$, $G \in \mathfrak{F}$ engendrent des sous-espaces $W_{\mathfrak{G}}$ non triviaux. On note alors simplement $W_G = W_{\{F \in \mathfrak{F},\ G \preceq F\}}$ et $Q_G = Q_{\{F \in \mathfrak{F},\ G \preceq F\}}$ et on déduit de la Formule (1.5) la décomposition en somme directe orthogonale

$$\mathbb{R}^\Omega = \bigoplus_{F \in \mathfrak{F}}^{\perp} W_F.$$

Puis en notant que

$$P_F Q_G = \begin{cases} Q_G & \text{si } G \preceq F \\ 0 & \text{sinon} \end{cases}$$

on a

$$P_F = P_F \Big(\sum_{G \in \mathfrak{F}} Q_G \Big) = \sum_{G \preceq F} Q_G \tag{1.6}$$

et par suite

$$V_F = \bigoplus_{G \preceq F}^{\perp} W_G.$$

(Unicité) Considérons une autre décomposition $\mathbb{R}^\Omega = \bigoplus_{F \in \mathfrak{F}}^{\perp} W'_F$, supposons qu'il existe F dans

$\mathfrak{F}$ tel que $W_F \neq W'_F$ et montrons qu'on aboutit à une absurdité. En effet, on peut toujours trouver $G \preceq F$ tel que $W_G \neq W'_G$ et $W_H = W'_H$ pour tout $H \prec G$. Par suite, comme

$$V_G = \bigoplus_{H \preceq G}^{\perp} W_H = \bigoplus_{H \preceq G}^{\perp} W'_H,$$

on obtient que soit $\{H \in \mathfrak{F} \mid H \prec G\}$ est vide, soit $\bigoplus_{H \prec G}^{\perp} W_H \neq \bigoplus_{H \prec G}^{\perp} W'_H$. Le premier cas mène à $V_G \neq V_G$ et le second à $\bigoplus_{H \prec G}^{\perp} W_H \neq \bigoplus_{H \prec G}^{\perp} W_H$; la conclusion suit. □

Les sous-espaces vectoriels W_F, $F \in \mathfrak{F}$, sont généralement appelés *strates*, et leurs dimensions respectives, notées d_F, sont appelées *degré de liberté* de F. Dans le cadre combinatoire introduit précédemment, on parle couramment de *décomposition en strates* lorsqu'on évoque la décomposition de la Formule (1.3). Ces sous-espaces W_F, $F \in \mathfrak{F}$ peuvent être explicités en fonction des V_G, $G \in \mathfrak{F}$.

Corollaire 1.1. *La strate W_F associée à une partition F de $\mathfrak{F}$ se définit comme*

$$W_F = V_F \cap \Big(\sum_{G \in \mathfrak{F},\ G \prec F} V_G \Big)^{\perp}.$$

Démonstration. On note d'abord que par définition des Q_F, $F \in \mathfrak{F}$, dans la preuve du théorème précédent

$$\begin{aligned} Q_F &= Q_{\{G \in \mathfrak{F},\ F \preceq G\}} \\ &= \Big(\prod_{F \preceq G} P_G \Big) \Big(\prod_{F \not\preceq G} (I - P_G) \Big) \\ &= P_F \Big(\prod_{F \not\preceq G} (P_F - P_{F \wedge G}) \Big) \\ &= P_F \Big(\prod_{G \prec F} (P_F - P_G) \Big). \end{aligned}$$

Ensuite, considérons $f \in \mathbb{R}^{\Omega}$, $F \in \mathfrak{F}$ et $G_0 \prec F$; alors comme les P_G, $G \in \mathfrak{F}$, commutent deux à deux, on a

$$\begin{aligned} Q_F(f) &= (P_F - P_{G_0}) P_F \prod_{G \prec F,\ G \neq G_0} (P_F - P_{G_0})(f) \\ &= (P_F - P_{G_0})(f'), \quad \text{avec } f' \in V_F \end{aligned}$$

et donc $Q_F(f) \in V_{G_0}^{\perp}$. Par suite $Q_F(f) \in \big(\sum_{G \prec F} V_G \big)^{\perp}$, et de la même manière il est aisé de constater que $Q_F(f) \in V_F$. Finalement, on a l'inclusion

$$W_F \subseteq V_F \cap \Big(\sum_{G \in \mathfrak{F},\ G \prec F} V_G \Big)^{\perp}.$$

L'inclusion inverse se déduit également de la commutativité des projecteurs orthogonaux P_F, $F \in \mathfrak{F}$, en notant que si $f \in V_F \cap \big(\sum_{G \in \mathfrak{F},\ G \prec F} V_G \big)^{\perp}$ alors $Q_F(f) = f$. Et la conclusion suit. □

Remarque 1.1. *Dans la preuve du* Théorème *1.1, l'item* (c) *dans la caractérisation des projecteurs orthogonaux $Q_{\mathfrak{G}}$, $\mathfrak{G} \subseteq \mathfrak{F}$ ne figure pas dans la preuve originale. C'est un ajout que l'on fait car $Q_{\emptyset}$ n'est pas automatiquement l'application nulle, elle l'est dans le cas d'une structure en bloc de Tjur parce qu'une telle structure contient la partition en singletons E. D'ailleurs, c'est à cet endroit seulement de la preuve que* (T1) *dans la* Définition 1.8 *opère. En toute rigueur, la preuve originale ne permet pas de conclure à* (1.3).

Remarque 1.2. *Bailey* [Bai96] *donne une démonstration extrêmement réduite de ce théorème en se basant sur la formulation explicite des strates donnée dans le* Corollaire 1.1 — *voir* Théorème 2 *page 56 dans* [Bai96]. *On conseille néanmoins la lecture de la preuve originale de Tjur plutôt que celle de Bailey tant les hypothèses d'application — les propriétés* (T1), (T2) *et* (T3) *de la* Définition 1.8 — *sont masquées dans la version de cette dernière.*

1.1.4 Effet des facteurs et leur quantification

Dans cette section, nous parlons d'*effet de facteurs* mais il s'agit en toute rigueur d'*effet de partitions* induites par leurs facteurs correspondants.

En reprenant la définition des projecteurs orthogonaux P_F, $F \in \mathfrak{F}$, et le Lemme 1.1, on définit naturellement $P_F(f)$ comme l'*effet* des facteurs $G \preceq F$ sur la fonction $f \in \mathbb{R}^\Omega$. Et en introduisant la norme euclidienne induite par le produit scalaire sur $\mathbb{R}^\Omega$, notée simplement $||\cdot||$, on peut quantifier cet effet par

$$\mathrm{SCB}_F(f) = ||P_F(f)||^2 \tag{1.7}$$

où SCB est l'abréviation de la dénomination classique *somme de carrés brute* — en anglais, *crude sum of squares* (CSS). Dans ce cadre la décomposition ANOVA consiste simplement en un moyen d'isoler l'effet de chacun des facteurs dans $P_F(f)$. Plus précisément, à l'aide des projecteurs orthogonaux Q_F sur W_F, $F \in \mathfrak{F}$, on définit naturellement $Q_F(f)$ comme l'effet du facteur F sur la fonction $f \in \mathbb{R}^\Omega$. Cet effet est alors quantifié par

$$\mathrm{SC}_F(f) = ||Q_F(f)||^2$$

où SC est l'abréviation de la dénomination classique *somme de carrés* — en anglais *sum of squares* (SS). Ce formalisme est bien défini, au sens où, les Q_F, $F \in \mathfrak{F}$, étant deux à deux orthogonaux — voir la preuve du Théorème 1.1 — on a

$$\mathrm{SCB}_F(f) = \sum_{G \preceq F} \mathrm{SC}_G(f)$$

et en particulier pour $F = E$

$$||f||^2 = \sum_{F \in \mathfrak{F}} \mathrm{SC}_F(f). \tag{1.8}$$

En pratique, un calcul direct permet d'estimer les quantités $\mathrm{SCB}_F(f)$, $F \in \mathfrak{F}$, et on applique une formule d'inversion de Möbius — voir par exemple [Sta12] — pour calculer l'effet de chacun des facteurs comme suit

$$\mathrm{SC}_F(f) = \sum_{G \preceq F} \mu(G, F) \mathrm{SCB}_G(f)$$

avec la fonction de Möbius définie récursivement

$$\begin{aligned} \mu(F, F) &= 1 && \text{pour tout } F \in \mathfrak{F} \\ \mu(G, F) &= -\sum_{G \preceq H \prec F} \mu(G, H) && \text{pour tous } G \prec F,\ G, F \in \mathfrak{F}. \end{aligned}$$

En particulier, si la famille de partitions $\mathfrak{F}$ est de la forme $\mathfrak{F} = \mathcal{P}\big(\{F_1, \ldots, F_n\}\big)$ alors l'ordre partiel $\preceq$ est simplement la relation d'inclusion $\subseteq$, et $\mu(G, F) = (-1)^{|F|-|G|}$, $F, G \in \mathfrak{F}$, et on aboutit alors à la formule classique

$$\mathrm{SC}_F(f) = \sum_{G \subseteq F} (-1)^{|F|-|G|} \mathrm{SCB}_G(f). \tag{1.9}$$

1.1.5 Généralisation à un espace probabilisé

La généralisation de la décomposition de Tjur donnée par Helland [Hel98] consiste à appliquer le formalisme combinatoire précédent en remplaçant l'ensemble fini Ω par un espace probabilisé en notant que la démonstration du Théorème 1.1 ne met en jeu que des combinaisons de projecteurs orthogonaux, et que le rôle spécifique de Ω et de son espace fonctionnel associé $\mathbb{R}^\Omega$ peuvent être marginalisés.

On considère dorénavant un espace probabilisé $(\Omega, \mathcal{E}, \mathbb{P})$ ainsi que l'espace vectoriel, noté $L^2(\mathcal{E}) = L^2(\Omega, \mathcal{E}, \mathbb{P}; \mathbb{R})$, des variables aléatoires réelles de carré intégrable définies sur $(\Omega, \mathcal{E}, \mathbb{P})$. Sa structure d'espace de Hilbert — pour le produit scalaire usuel $\langle Y_1, Y_2 \rangle_\mathbb{P} = \int Y_1 Y_2 \mathrm{d}\mathbb{P}$, Y_1, $Y_2 \in L^2(\mathcal{E})$ — permet de définir des projecteurs orthogonaux via l'espérance conditionnelle.

Nous reprenons maintenant point par point le formalisme introduit dans la section précédente en l'allégeant néanmoins de la notion de facteur.

Partition Une partition de $(\Omega, \mathcal{E}, \mathbb{P})$ est une sous-tribu de $\mathcal{E}$. Par la suite, on ne parle d'ailleurs plus de partition, mais uniquement de sous-tribu.

Ordre partiel L'ensemble des sous-tribus de $\mathcal{E}$ est naturellement muni d'un ordre partiel induit par l'inclusion. Une sous-tribu $\mathcal{F}$ est dite plus fine que $\mathcal{G}$ si $\mathcal{G} \subseteq \mathcal{F}$. La partition la plus fine est la tribu $\mathcal{E}$, et la plus grossière est la tribu grossière $\{\emptyset, \Omega\}$.

Infimum, supremum On définit l'infimum et le supremum de deux sous-tribus $\mathcal{F}$ et $\mathcal{G}$ de $\mathcal{E}$ respectivement par

$$\begin{aligned} \mathcal{F} \wedge \mathcal{G} &= \mathcal{F} \cap \mathcal{G} \\ \mathcal{F} \vee \mathcal{G} &= \sigma(\mathcal{F} \cup \mathcal{G}) \end{aligned}$$

où $\sigma(\cdot)$ désigne la tribu engendrée. Ces deux opérations sont clairement associatives et commutatives.

Sous-espace vectoriel associé à une tribu À toute sous-tribu $\mathcal{F}$ de $\mathcal{E}$ est associé le sous-espace vectoriel

$$L^2(\mathcal{F}) = \{Y \in L^2(\mathcal{E}) \mid Y \text{ est mesurable par rapport à } \mathcal{F}\}. \tag{1.10}$$

Projecteur orthogonal associé à une tribu À toute sous-tribu $\mathcal{F}$ de $\mathcal{E}$ est associée le projecteur orthogonal $P_{\mathcal{F}}$ sur $L^2(\mathcal{F})$ défini via l'espérance conditionnelle

$$\forall Y \in L^2(\Omega),\ P_{\mathcal{F}}(Y) = \mathbb{E}\big[Y|\mathcal{F}\big].$$

Tribus orthogonales Deux tribus $\mathcal{F}$ et $\mathcal{G}$ de $\mathcal{E}$ sont dites orthogonales si $L^2(\mathcal{F}) \cap L^2(\mathcal{F} \wedge \mathcal{G})^{\perp}$ et $L^2(\mathcal{G}) \cap L^2(\mathcal{F} \wedge \mathcal{G})^{\perp}$ sont des s.e.v. orthogonaux.

Avec ce nouveau cadre, on aboutit à la généralisation du Théorème 1.1 annoncée.

Théorème 1.2. [Helland, 1998] *Soit $\mathfrak{F}$ une famille finie de sous-tribus constituant une structure en blocs de Tjur — voir* Définition 1.8 *— alors il existe une unique décomposition de $L^2(\mathcal{E})$ en somme directe orthogonale*

$$L^2(\mathcal{E}) = \bigoplus_{\mathcal{F} \in \mathfrak{F}}^{\perp} W_{\mathcal{F}}$$

satisfaisant pour toute sous-tribu $\mathcal{F} \in \mathfrak{F}$

$$L^2(\mathcal{F}) = \bigoplus_{\mathcal{G} \subseteq \mathcal{F}}^{\perp} W_{\mathcal{G}}$$

où les $L^2(\mathcal{F})$ sont définis dans la Formule (1.10).

Démonstration. La preuve consiste d'abord à remarquer que la décomposition ANOVA donnée par Tjur peut se réduire à la décomposition des projecteurs orthogonaux P_F, $F \in \mathfrak{F}$ de la Formule (1.6), la décomposition de l'espace euclidien $\mathbb{R}^{\Omega}$ de la Formule (1.3) en étant seulement une conséquence directe. Par suite, la preuve du Théorème 1.1 est indépendante de la structure de l'espace de fonction considéré et ne repose que sur les propriétés de la famille $\mathfrak{F}$; elle s'applique donc sans restriction ici. □

1.1.6 Décomposition ANOVA de variables aléatoires

Il est possible de restreindre la décomposition de Helland à des familles de tribus engendrées par des variables aléatoires indépendantes. Dans ce cas, la décomposition ANOVA obtenue est celle énoncée par Efron et Stein en 1981 [ES81] — voir également [Van98] pages 158–159 — conséquence des travaux initiaux de Hoeffding sur les U-statistiques [Hoe48]. Nous présentons donc ce résultat

comme un corollaire anachronique du Théorème 2. Nous terminons en donnant une version de cette décomposition ANOVA pour des variables aléatoires dépendantes [Hoo07, CGP12] ; dans ce dernier cas, l'espace probabilisé $L^2(\mathcal{E})$ se décompose toujours en somme directe de strates W_F, $F \in \mathfrak{F}$. Néanmoins, les strates ne sont plus orthogonales, et par conséquent, le formalisme de la décomposition de Tjur et Helland ne permettent pas d'aboutir à cette décomposition.

Corollaire 1.2. [Efron & Stein, 1981] *Soient $X_1, \ldots, X_d$ des variables aléatoires indépendantes réelles définies sur un espace probabilisé $(\Omega, \mathcal{A}, \mathbb{P})$, et $\mathcal{E} = \sigma(\mathbf{X})$ la tribu engendrée par le vecteur aléatoire $\mathbf{X} = (X_1, \ldots, X_d)$. Alors toute variable aléatoire réelle de carré intégrable Y définie sur $(\Omega, \mathcal{E}, \mathbb{P})$ se décompose de manière unique sous la forme*

$$Y = \sum_{\mathfrak{u} \subseteq [1:d]} \eta_{\mathfrak{u}}(\mathbf{X}_{\mathfrak{u}}) \quad \mathbb{P}\text{-}p.s.$$

où les $2^d - 1$ variables aléatoires dans le membre de droite sont deux à deux orthogonales.

Démonstration. Pour tout $\mathfrak{u} \subseteq [1:d]$ non vide, considérons la tribu engendrée par les $\mathbf{X}_{\mathfrak{u}}$, $\mathcal{F}_{\mathfrak{u}} = \sigma(\mathbf{X}_{\mathfrak{u}})$ et notons $\mathcal{F}_{\emptyset}$ la tribu grossière. Alors $\mathfrak{F} = \big\{\mathcal{F}_{\mathfrak{u}},\ \mathfrak{u} \subseteq [1:d]\big\}$ est une structure en blocs de Tjur. En effet

(T1) est triviale

(T2) est vérifiée en notant que pour tous $\mathfrak{u}$, $\mathfrak{v}$ dans $[1:d]$, $\mathcal{F}_{\mathfrak{u}} \cap \mathcal{F}_{\mathfrak{v}} = \mathcal{F}_{\mathfrak{u} \cap \mathfrak{v}}$

(T3) est vérifiée en notant que pour tous $\mathfrak{u}$, $\mathfrak{v}$ dans $[1:d]$, $\mathcal{F}_{\mathfrak{u}}$ et $\mathcal{F}_{\mathfrak{v}}$ sont indépendantes conditionnellement à $\mathcal{F}_{\mathfrak{u}} \cap \mathcal{F}_{\mathfrak{v}}$ (voir, e.g., [CMC03]), et par suite pour toutes variables aléatoires Y_1, Y_2 respectivement dans $L^2(\mathcal{F}_{\mathfrak{u}}) \cap L^2(\mathcal{F}_{\mathfrak{u}} \cap \mathcal{F}_{\mathfrak{v}})^{\perp}$ et $L^2(\mathcal{F}_{\mathfrak{v}}) \cap L^2(\mathcal{F}_{\mathfrak{u}} \cap \mathcal{F}_{\mathfrak{v}})^{\perp}$, on a

$$\mathbb{E}\big[Y_1 Y_2\big] = \mathbb{E}\Big[\mathbb{E}\big[Y_1 Y_2 | \mathcal{F}_{\mathfrak{u}} \cap \mathcal{F}_{\mathfrak{v}}\big]\Big] = \mathbb{E}\Big[\mathbb{E}\big[Y_1 | \mathcal{F}_{\mathfrak{u}} \cap \mathcal{F}_{\mathfrak{v}}\big]\mathbb{E}\big[Y_2 | \mathcal{F}_{\mathfrak{u}} \cap \mathcal{F}_{\mathfrak{v}}\big]\Big] = 0$$

et la conclusion suit.

Par suite le Théorème 2 s'applique et on a la décomposition en somme directe orthogonale

$$L^2(\mathcal{E}) = \bigoplus_{\mathfrak{u} \subseteq [1:d]}^{\perp} W_{\mathfrak{u}} \tag{1.11}$$

avec pour tout $\mathfrak{u} \subseteq [1:d]$

$$L^2(\mathcal{F}_{\mathfrak{u}}) = \bigoplus_{\mathfrak{v} \subseteq \mathfrak{u}}^{\perp} W_{\mathfrak{v}}. \tag{1.12}$$

Ensuite, soit $Y \in L^2(\mathcal{E})$, notons $Y = \sum_{\mathfrak{u} \subseteq [1:d]} Y_{\mathfrak{u}}$ la décomposition de Y. Alors la Formule (1.12) implique que

$$\mathbb{E}\big[Y | \mathcal{F}_{\mathfrak{u}}\big] = \sum_{\mathfrak{v} \subseteq \mathfrak{u}} Y_{\mathfrak{v}} \quad \mathbb{P}\text{-}p.s., \quad \mathfrak{u} \subseteq [1:d]$$

et par la formule d'inversion de Möbius donnée en (1.9), on obtient la forme explicite suivante des $Y_{\mathfrak{u}}$, $\mathfrak{u} \subseteq [1:d]$

$$Y_{\mathfrak{u}} = \sum_{\mathfrak{v} \subseteq \mathfrak{u}} (-1)^{|\mathfrak{u}| - |\mathfrak{v}|} \mathbb{E}\big[Y | \mathcal{F}_{\mathfrak{v}}\big] \quad \mathbb{P}\text{-}p.s.$$

et la conclusion suit. □

Dans le cas de variables aléatoires dépendantes, la décomposition en strates qu'on obtient dans la Formule (1.11) du corollaire précédent est toujours valide, mais les strates ne sont plus orthogonales deux à deux. Les travaux sur ces décompositions ont été initiés par Stone [Sto94] et Hooker [Hoo07], et ont fait l'objet d'un développement récent par Chastaing et al. [CGP12].

Théorème 1.3. [Chastaing, Gamboa et Prieur, 2012] *Soient $X_1, \ldots, X_d$ des variables aléatoires dépendantes réelles définies sur un espace probabilisé $(\Omega, \mathcal{A}, \mathbb{P})$, $\mathcal{E} = \sigma(\mathbf{X})$ la tribu engendrée par le*

vecteur aléatoire $\mathbf{X} = (X_1, \dots, X_d)$*, et* $\mathbb{P}_{\mathbf{X}}$ *la mesure image de* $\mathbf{X}$*. Si* $\mathbb{P}_{\mathbf{X}}$ *est absolument continue par rapport à une mesure produit* $\boldsymbol{\nu}$ *et qu'il existe une constante* $0 < M \leq 1$ *telle que pour tout* $\mathfrak{u} \subseteq [1:d]$

$$M \frac{\mathrm{d}\mathbb{P}_{\mathbf{X}_\mathfrak{u}}}{\mathrm{d}\boldsymbol{\nu}_\mathfrak{u}} \frac{\mathrm{d}\mathbb{P}_{\mathbf{X}_{\mathfrak{u}^c}}}{\mathrm{d}\boldsymbol{\nu}_{\mathfrak{u}^c}} \leq \frac{\mathrm{d}\mathbb{P}_{\mathbf{X}}}{\mathrm{d}\boldsymbol{\nu}},$$

alors on a la décomposition de $L^2(\mathcal{E})$ *en somme directe*

$$L^2(\mathcal{E}) = \bigoplus_{\mathfrak{u} \subseteq [1:d]} W_\mathfrak{u}$$

où les strates $W_\mathfrak{u}$*,* $\mathfrak{u} \subseteq [1:d]$*, sont définies par*

$$W_\mathfrak{u} = V_\mathfrak{u} \cap \Big(\sum_{\mathfrak{v} \subsetneq \mathfrak{u}} V_\mathfrak{v} \Big)^\perp$$

où $V_\mathfrak{u} = L^2\big(\sigma(\mathbf{X}_\mathfrak{u})\big)$*.*

Démonstration. Voir la preuve du Théorème 1 dans [CGP12]. □

Dans la décomposition de Efron & Stein, il est possible de quantifier l'effet de chaque variable ou groupe de variables en utilisant la norme induite par le produit scalaire comme cela est fait dans la Section 1.1.4. Plus précisément, pour $\mathfrak{u} \subseteq [1:d]$ non vide, on introduit les quantités

$$\begin{aligned} \underline{\tau}^2_\mathfrak{u} &= \big|\big|P_{\mathcal{F}_\mathfrak{u}}(Y)\big|\big|^2 \\ &= \langle Y, P_{\mathcal{F}_u}(Y)\rangle_\mathbb{P} \end{aligned} \tag{1.13}$$

et

$$\sigma^2_\mathfrak{u} = \big|\big|Q_{\mathcal{F}_\mathfrak{u}}(Y)\big|\big|^2 \tag{1.14}$$

$$= \langle Y, Q_{\mathcal{F}_u}(Y)\rangle_\mathbb{P}. \tag{1.15}$$

On pose $\sigma^2_\emptyset = \underline{\tau}^2_\emptyset = 0$, et pour tout $\mathfrak{u} \subseteq [1:d]$ non vide, on a aisément que

$$\underline{\tau}^2_\mathfrak{u} = \big|\big|\mathbb{E}[Y|\mathcal{F}_\mathfrak{u}]\big|\big|^2 = \mathrm{Var}[\mathbb{E}[Y|\mathcal{F}_\mathfrak{u}]]$$

— car $W_\mathfrak{u} \subseteq L^2(\Omega, \mathcal{F}_\emptyset, \mathbb{P})^\perp$ d'après le Corollaire 1.1. Par linéarité du produit scalaire, (1.13) et (1.15) donnent

$$\underline{\tau}^2_\mathfrak{u} = \sum_{\mathfrak{v} \subseteq \mathfrak{u}} \sigma^2_\mathfrak{u}$$

et en particulier, en notant $\sigma^2 = \mathrm{Var}[Y]$, on a

$$\sigma^2 = \sum_{\mathfrak{u} \subseteq [1:d]} \sigma^2_\mathfrak{u}. \tag{1.16}$$

Puis par la formule d'inversion de Möbius de la Formule (1.9), on a

$$\sigma^2_\mathfrak{u} = \sum_{\mathfrak{v} \subseteq \mathfrak{u}} (-1)^{|\mathfrak{u}|-|\mathfrak{v}|} \underline{\tau}^2_\mathfrak{v}. \tag{1.17}$$

En ce qui concerne la décomposition ANOVA pour variables aléatoires dépendantes, la quantification est moins complète. On définit uniquement les $\sigma^2_\mathfrak{u}$, $\mathfrak{u} \subseteq [1:d]$, par la Formule (1.15) — notons au passage que l'égalité entre (1.14) et (1.15) ne tient plus par défaut d'orthogonalité. Dans ce cas, on a pour tout $\mathfrak{u} \subseteq [1:d]$, $\sigma^2_\mathfrak{u} = \mathrm{Cov}(Y, \eta_\mathfrak{u}(\mathbf{X}_\mathfrak{u}))$ où Y se décompose sur les strates $W_\mathfrak{u}$ (voir Théorème 1.3. ci-avant) comme

$$Y = \sum_{\mathfrak{u} \subseteq [1:d]} \eta_\mathfrak{u}(\mathbf{X}_\mathfrak{u}).$$

Et par suite on obtient trivialement

$$\sigma^2 = \sum_{\mathfrak{u} \subseteq [1:d]} \sigma^2_{\mathfrak{u}}.$$

Nous remarquons qu'en introduisant dans le Corollaire 1.2 la notion de fonction de plusieurs variables — rappelons que Y est mesurable par rapport à $\mathcal{E}$ si et seulement si il existe une fonction borélienne $f : \mathbb{R}^d \to \mathbb{R}$ telle que $Y = f(X_1, \ldots, X_d)$ — on fait apparaître naturellement la notion de produit tensoriel. C'est un outil important dans la décomposition ANOVA comme nous le montrons dans la section suivante.

1.2 Décomposition ANOVA par l'algèbre tensorielle

L'algèbre tensorielle est extrêmement adaptée au traitement de la décomposition ANOVA, comme en témoigne l'article de Takemura en 1983 [Tak83]. Elle permet entre autre de retrouver la décomposition ANOVA du Corollaire 1.2 d'une façon différente. On obtient alors ce que nous appelons *décomposition de Sobol'* en référence au travail initial de Sobol' [Sob93]. Notons que l'intérêt essentiel d'utiliser l'algèbre tensorielle dans le cadre de la décomposition ANOVA est qu'elle permet de définir des décompositions ANOVA *spectrales* relativement à des bases orthonormales.

1.2.1 Décomposition sur un espace sous-jacent fini

Cette section s'inspire en partie de la décomposition énoncée dans le Théorème 3.2 pages 42–43 dans [Col70].

Espaces fonctionnels élémentaires Soient $\Omega_1, \ldots, \Omega_d$ des ensembles finis non vides, $\mathcal{B}_1, \ldots, \mathcal{B}_d$ leur tribu discrète associée — i.e. $\mathcal{B}_i = \mathcal{P}(\Omega_i)$, $i \in [1:d]$ — et $\mu_1, \ldots, \mu_d$ des mesures de probabilité définies sur chacun des $\mathcal{B}_i$, $i \in [1:d]$. Pour tout indice $i \in [1:d]$, on considère l'espace vectoriel $\mathbb{R}^{\Omega_i}$ des fonctions réelles définies sur Ω_i, et on introduit ses deux s.e.v.

$$\begin{aligned} W_{\mu_i,0} &= \Big\{ f \in \mathbb{R}^{\Omega_i} \;\Big|\; f \text{ est constante } \mu_i\text{-}p.s. \Big\}, \\ W_{\mu_i,1} &= \Big\{ f \in \mathbb{R}^{\Omega_i} \;\Big|\; \int_{\Omega_i} f \mathrm{d}\mu_i = 0 \Big\}. \end{aligned}$$

Notons que $W_{\mu_i,0}$ et $W_{\mu_i,1}$ sont supplémentaires l'un de l'autre dans $\mathbb{R}^{\Omega_i}$.

Espace fonctionnel principal Notons

$$\begin{aligned} \Omega &= \Omega_1 \times \cdots \times \Omega_d, \\ \mathcal{B} &= \mathcal{B}_1 \times \cdots \times \mathcal{B}_d = \mathcal{P}(\Omega), \\ \mu &= \mu_1 \otimes \cdots \otimes \mu_d. \end{aligned}$$

On peut alors définir l'espace $\mathbb{R}^\Omega$ comme l'espace des fonctions réelles définies sur Ω. Il est lié naturellement aux $\mathbb{R}^{\Omega_i}$ par l'isomorphisme

$$\mathbb{R}^\Omega \simeq \bigotimes_{i \in [1:d]} \mathbb{R}^{\Omega_i} \tag{1.18}$$

— voir par exemple le Corollaire 1 dans [CO68] page 84. En outre, on munit cet espace du produit scalaire

$$\langle f, g \rangle_\mu = \sum_{\omega \in \Omega} f(\omega) g(\omega) \mu(\omega),$$

ce qui lui confère une structure d'espace de Hilbert.

Sous-espaces vectoriels de $\mathbb{R}^\Omega$ L'identification (1.18) permet d'introduire des sous-espaces vectoriels de $\mathbb{R}^\Omega$ construits à partir du produit tensoriel. En particulier, on définit les espaces $V_{\mu,\mathfrak{u}}$, $\mathfrak{u} \subseteq [1:d]$ par

$$V_{\mu,\mathfrak{u}} = \bigotimes_{i\in[1:d]} U_i$$

avec pour tout $i \in [1:d]$

$$U_i = \begin{cases} \mathbb{R}^{\Omega_i} & \text{si } \ i \in \mathfrak{u} \\ W_{\mu_i,0} & \text{sinon.} \end{cases}$$

Les éléments de $V_{\mu,\mathfrak{u}}$ sont les fonctions $f \in \mathbb{R}^\Omega$ telles que pour tout $\mathbf{x}$ et $\mathbf{x}'$ dans Ω,

$$\mathbf{x}_\mathfrak{u} = \mathbf{x}'_\mathfrak{u} \Rightarrow f(\mathbf{x}) = f(\mathbf{x}') \ \ \mu\text{-}p.s.$$

Plus simplement, ce sont les fonctions qui ne dépendent pas des variables $i \notin \mathfrak{u}$.

Le résultat principal de cette section repose essentiellement sur l'identification qui suit — voir Théorème 8.10 pages 94–95 dans [CO68].

Lemme 1.3. *Soient I et J_i, $i \in I$ une famille d'ensembles finis non vides, et $E_{i,j}$, $i \in I$, $j \in J_i$ une famille d'espaces vectoriels. Pour tout $i \in I$, considérons*

$$E_i = \bigoplus_{j\in J_i} E_{i,j}.$$

Soit $\mathfrak{X}$ l'ensemble des applications χ de I dans $\cup_{i\in I} J_i$ telles que pour tout $i \in I$, $\chi(i) \in J_i$. Pour tout $\chi \in \mathfrak{X}$, considérons

$$E_\chi = \bigotimes_{i\in I} E_{i,\chi(i)}.$$

Alors on a l'isomorphisme

$$\bigotimes_{i\in I} E_i \simeq \bigoplus_{\chi\in\mathfrak{X}} E_\chi.$$

Démonstration. Voir la preuve du Théorème 8.10 [CO68]. □

Théorème 1.4. *Avec les notations précèdentes, il existe une unique décomposition de $\mathbb{R}^\Omega$ en somme directe orthogonale*

$$\mathbb{R}^\Omega = \overset{\perp}{\bigoplus_{\mathfrak{u}\subseteq[1:d]}} W_{\mu,\mathfrak{u}} \tag{1.19}$$

telle que

$$V_{\mu,\mathfrak{u}} = \overset{\perp}{\bigoplus_{\mathfrak{v}\subseteq\mathfrak{u}}} W_{\mu,\mathfrak{v}}. \tag{1.20}$$

Démonstration. **(Existence)** Pour tout $\mathfrak{u} \subseteq [1:d]$ on définit les strates

$$W_{\mu,\mathfrak{u}} = \bigotimes_{i\in\mathfrak{u}} W_{\mu_i,\mathbf{1}_\mathfrak{u}(i)}.$$

Après avoir remarqué que pour tout $i \in [1:d]$, les s.e.v. de $\mathbb{R}^{\Omega_i}$, $W_{\mu_i,0}$ et $W_{\mu_i,1}$ sont supplémentaires dans $\mathbb{R}^{\Omega_i}$, on applique le Lemme 1.3 aux espaces vectoriels E_i, $i \in [1:d]$, suivants

$$E_i = W_{\mu_i,0} \oplus W_{\mu_i,1}.$$

On obtient que

$$\bigotimes_{i\in[1:d]} \mathbb{R}^{\Omega_i} = \bigoplus_{\mathfrak{u}\subseteq[1:d]} W_{\mu,\mathfrak{u}}$$

et par l'identification (1.18) on aboutit à (1.19). Puis pour $\mathfrak{u} \subseteq [1:d]$, en appliquant le Lemme 1.3 aux espaces vectoriels E_i, $i \in [1:d]$, suivants

$$\begin{aligned} E_i &= W_{\mu_i,0} \oplus W_{\mu_i,1} \quad \text{si } i \in \mathfrak{u} \\ E_i &= W_{\mu_i,0} \qquad\qquad\quad \text{sinon} \end{aligned}$$

on aboutit à (1.20).

(Unicité) On le montre aisément en appliquant une preuve identique à celle de l'unicité de la décomposition dans le Théorème 1.1.

(Orthogonalité) Soient $\mathfrak{u}$ et $\mathfrak{v}$ deux sous ensembles distincts de $[1:d]$; on montre que pour toutes fonctions $f \in W_{\mu,\mathfrak{u}}$ et $g \in W_{\mu,\mathfrak{v}}$, on a $\langle f, g\rangle_\mu = 0$. Dans un premier temps, on vérifie cette propriété pour

$$f = \bigotimes_{i\in[1:d]} f_i \;\text{ avec } f_i \in W_{\mu_i,\mathbf{1}_\mathfrak{u}(i)}$$

et

$$g = \bigotimes_{i\in[1:d]} g_i \;\text{ avec } g_i \in W_{\mu_i,\mathbf{1}_\mathfrak{u}(i)}.$$

En effet, dans ce cas, comme $\mathfrak{u}$ et $\mathfrak{v}$ sont distincts, il existe $j \in [1:d]$ tel que $j \in \mathfrak{u}\Delta\mathfrak{v}$. Par suite $f_j g_j$ est de moyenne nulle et on conclut par le théorème de Fubini. La conclusion concernant l'orthogonalité suit en notant que les f et g considérés plus haut engendrent respectivement $W_{\mu,\mathfrak{u}}$ et $W_{\mu,\mathfrak{v}}$. □

1.2.2 Décomposition de Sobol'

La construction de la décomposition précédente se généralise de façon naturelle à la dimension infinie. Pour la suite, considérons p tel que $1 \leq p < +\infty$.

Espaces fonctionnels élémentaires On considère les espaces $L^p(\mu_i) = L^p(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mu_i; \mathbb{R})$, $i \in [1:d]$, où $\mathcal{B}(\mathbb{R})$ désigne la tribu borélienne sur $\mathbb{R}$, et μ_i, $i \in [1:d]$, une famille de mesures de probabilité sur $\mathbb{R}$, et on introduit leurs deux s.e.v.

$$\begin{aligned} W_{\mu_i,0} &= \Big\{ f \in L^p(\mu_i) \;\Big|\; f \text{ est constante } \mu_i\text{-}p.s. \Big\} \\ W_{\mu_i,1} &= \Big\{ f \in L^p(\mu_i) \;\Big|\; \int_{\Omega_i} f \,\mathrm{d}\mu_i = 0 \Big\}. \end{aligned}$$

Notons que $W_{\mu_i,0}$ et $W_{\mu_i,1}$ sont supplémentaires l'un de l'autre dans $L^p(\mu_i)$.

Espace fonctionnel principal Considérons l'espace vectoriel $L^p(\mu) = L^p(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \mu; \mathbb{R})$ où $\mu = \mu_1 \otimes \cdots \otimes \mu_d$. Il est lié naturellement aux $L^p(\mu_i)$ par l'isomorphisme

$$L^p(\mu) \simeq \bigotimes_{i\in[1:d]} L^p(\mu_i) \tag{1.21}$$

— voir par exemple le Théorème VI-20 page 212 dans [Vil08]. En outre, si $p = 2$, $L^p(\mu)$ est un espace de Hilbert pour le produit scalaire

$$\langle f, g\rangle_\mu = \int_{\mathbb{R}^d} fg \,\mathrm{d}\mu.$$

Sous-espaces vectoriels de $L^p(\mu)$ L'identification (1.21) permet d'introduire des sous-espaces vectoriels de $L^p(\mu)$ construits à partir du produit tensoriel. En particulier, on définit les espaces $V_{\mu,\mathfrak{u}}$, $\mathfrak{u} \subseteq [1:d]$ par

$$V_{\mu,\mathfrak{u}} = \bigotimes_{i\in[1:d]} U_i$$

avec pour tout $i \in [1:d]$

$$U_i = \begin{cases} L^p(\mu_i) & \text{si } \ i \in \mathfrak{u} \\ W_{\mu_i,0} & \text{sinon.} \end{cases}$$

Dans ce cadre, on obtient une décomposition ANOVA qu'on désigne comme la *décomposition de Sobol'* en référence à la décomposition énoncée par Sobol' en 1990 — voir Théorème 1 page 408 dans [Sob93] — dont le résultat ci-dessous n'est qu'une généralisation.

Théorème 1.5. (Décomposition de Sobol') *Avec les notations précèdentes, il existe une unique décomposition de $L^p(\mu)$ en somme directe*

$$L^p(\mu) = \bigoplus_{\mathfrak{u}\subseteq[1:d]} W_{\mu,\mathfrak{u}} \tag{1.22}$$

telle que

$$V_{\mu,\mathfrak{u}} = \bigoplus_{\mathfrak{v}\subseteq\mathfrak{u}} W_{\mu,\mathfrak{v}}. \tag{1.23}$$

De plus, pour $p = 2$, les sous-espaces vectoriels $W_{\mu,\mathfrak{u}}$, $\mathfrak{u} \subseteq [1:d]$ sont orthogonaux deux à deux.

Démonstration. La preuve est constructive, et pour tout $\mathfrak{u} \subseteq [1:d]$, on introduit les strates

$$W_{\mu,\mathfrak{u}} = \bigotimes_{i\in\mathfrak{u}} W_{\mu_i,\mathbf{1}_{\mathfrak{u}}(i)}.$$

Ensuite, le lemme 1.3 n'étant pas conditionné à la dimension finie des espaces vectoriels, il s'applique sans restriction ici et la preuve est identique à celle du Théorème 1.4. □

Remarque 1.3. *On retrouve rigoureusement l'énoncé de la décomposition ANOVA de Efron & Stein — voir* Corollaire 1.2 *— en notant que Y est $\sigma(\mathbf{X})$-mesurable si et seulement si il existe une application borélienne f de $\mathbb{R}^d$ dans $\mathbb{R}$ telle que $Y = f(\mathbf{X})$.*

Exemple 1.1. *Lorsque μ est la mesure uniforme sur l'hypercube unité $[0,1]^d$ et $p = 1$, on retrouve le* Théorème 1 *dans* [Sob93].

1.2.3 Indices de Sobol' et dimensions effectives

Nous revenons maintenant en détail sur la quantification des différents termes d'une décomposition ANOVA relative à des variables aléatoires indépendantes. Par la Remarque 1.3, une telle décomposition s'obtient de manière équivalente par le Théorème 1.2 ou 1.5. En outre, nous avons vu précédemment que cette quantification se fait naturellement en utilisant la norme euclidienne de l'espace de Hilbert considéré — voir Formules (1.13–1.17) —, et que la Formule (1.16) suggère de normaliser les quantités obtenues par la variance de Y, notée σ^2, qui ne dépend d'aucun sous-ensemble $\mathfrak{u} \subseteq [1:d]$. Cette normalisation mène à la définition des *indices de Sobol'* ou *indices de sensibilité basés sur la variance* dont nous donnons un résumé dans la définition suivante.

Définition 1.9. (Indices de Sobol') *Soit $\sigma^2 = \sum_{\mathfrak{v}\subseteq[1:d]} \sigma^2_{\mathfrak{u}}$ la décomposition de la variance correspondant à la décomposition ANOVA de d variables aléatoires indépendantes — voir Section 1.1.6. Alors pour $\mathfrak{u} \subseteq [1:d]$, on définit*

(i) l'indice de Sobol' élémentaire

$$S_{\mathfrak{u}} = \frac{\sigma^2_{\mathfrak{u}}}{\sigma^2}.$$

*Si $\mathfrak{u} = \{i\}$, $S_{\{i\}}$ quantifie l'importance relative de l'*effet principal *dû à la variable X_i. Si $|\mathfrak{u}| > 1$, $S_{\mathfrak{u}}$ quantifie l'importance relative de l'*interaction *d'ordre $|\mathfrak{u}|$ entre toutes les variables X_i, $i \in \mathfrak{u}$.*

(ii) l'indice de Sobol' descendant, *ou* global

$$\underline{S}_{\mathfrak{u}} = \frac{\tau^2_{\mathfrak{u}}}{\sigma^2},$$

où on rappelle

$$\underline{\tau}^2_{\mathfrak{u}} = \sum_{\mathfrak{v}\subseteq\mathfrak{u}} \sigma^2_{\mathfrak{v}}.$$

Il quantifie l'importance relative de tous les effets principaux dus à chacun des X_i, $i \in \mathfrak{u}$, et de toutes les interactions entre plusieurs variables parmi les X_i, $i \in \mathfrak{u}$.

(iii) l'indice de Sobol' ascendant, *ou* total

$$\overline{S}_{\mathfrak{u}} = \frac{\overline{\tau}^2_{\mathfrak{u}}}{\sigma^2},$$

où on définit

$$\overline{\tau}^2_{\mathfrak{u}} = \sum_{\mathfrak{u}\cap\mathfrak{v}\neq\emptyset} \sigma^2_{\mathfrak{v}}.$$

Il quantifie l'importance relative de tous les effets principaux et des interactions totalement ou partiellement dus aux variables X_i, $i \in \mathfrak{u}$.

Notons que pour tout $\mathfrak{u} \subseteq [1:d]$, $\overline{\tau}^2_{\mathfrak{u}} = \sigma^2 - \underline{\tau}^2_{\mathfrak{u}^c}$; et $0 \leq \sigma^2_u \leq \underline{\tau}^2_{\mathfrak{u}} \leq \overline{\tau}^2_{\mathfrak{u}} \leq \sigma^2$.

Ces notions permettent de quantifier l'importance relative des variables ou groupes de variables d'entrée sur la variable de sortie. Une autre notion utile pour définir la structure des variables ou groupes de variables non négligeables est celle de *dimension effective* [CMO97] qui se dédouble en deux définitions.

Définition 1.10. (Dimensions effectives) *La fonction f a pour dimension effective s au sens de la* superposition *si*

$$\sum_{|\mathfrak{u}|\leq s} \sigma^2_{\mathfrak{u}} \geq (1-\varepsilon)\sigma^2$$

et pour dimension effective s au sens de la troncature *si*

$$\sum_{\mathfrak{u}\subseteq 1:s} \sigma^2_{\mathfrak{u}} \geq (1-\varepsilon)\sigma^2$$

où ε est un paramètre fixé arbitrairement ; généralement, on considère $\varepsilon = 0,01$.

1.2.4 Décompositions spectrales

La décomposition de Sobol', pour $p = 2$, s'adapte particulièrement bien au cadre des bases orthonormales dans des espaces de Hilbert. La décomposition due à Sobol' en 1990 [Sob93] n'est d'ailleurs qu'une généralisation d'une décomposition obtenue en exploitant les décompositions en ondelettes de Haar, vingt ans plus tôt [Sob69] ; citons également les travaux de Cukier et al. [CLS78] exploitant les décompositions en série de Fourier dans les années 1970. De manière plus générale, la notion importante dans ce cas particulier est celle de *base adaptée* à une décomposition en somme directe.

Définition 1.11. (base adaptée) *Soient E un espace vectoriel, et $F_1, \ldots, F_d$ une famille de s.e.v. de E tels que*

$$E = F_1 \oplus \cdots \oplus F_d. \tag{1.24}$$

Une base $(\Phi_i)_{i\in I}$ de E est dite adaptée *à la décomposition* (1.24) *si il existe une partition $\{I_1, \ldots, I_d\}$ de I telle que pour tout $k \in [1:d]$, $(\Phi_i)_{i\in I_k}$ est une base de F_k.*

En particulier, on a la propriété suivante.

Proposition 1.1. *Pour tout $i \in [1:d]$, soit $(\Phi_{ij})_{j\in\mathbb{N}}$ une base orthonormale de $L^2(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mu_i; \mathbb{R})$ telle que $\Phi_{i0} = 1$. Alors $(\Phi_{1j_1} \otimes \cdots \otimes \Phi_{dj_d})_{\mathbf{j}\in\mathbb{N}^d}$ est une base adaptée à la décomposition ANOVA sur $L^2(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \mu; \mathbb{R})$ énoncée dans le* Théorème 1.5.

Démonstration. C'est une conséquence directe de la définition des strates dans la décomposition de Sobol' et du lemme suivant. □

Lemme 1.4. *Soient* $(\Phi_{1i})_{i\in\mathbb{N}}$ *et* $(\Phi_{2i})_{i\in\mathbb{N}}$ *des bases orthonormales respectives des espaces de Hilbert* $\mathbb{H}_1$ *et* $\mathbb{H}_2$*, alors* $(\Phi_{1i_1}\otimes\Phi_{2i_2})_{\mathbf{i}\in\mathbb{N}^2}$ *est une base orthonormale de* $\mathbb{H}_1\otimes\mathbb{H}_2$*.*

Démonstration. Voir par exemple Alinéa 6 pages 56–57 dans [CH53]. □

Une base orthonormale adaptée à une décomposition de Sobol' permet donc d'expliciter chaque composante de la décomposition de manière unique sur la famille de fonctions formant la base. En outre, elle permet de donner une version spectrale des mesures d'importance $\sigma^2_{\mathfrak{u}}$, $\mathfrak{u}\subseteq[1:d]$ et par suite des indices de Sobol'. Plus précisément, soient $\mathrm{c}_{\mathbf{k}}$, $\mathbf{k}\in\mathbb{N}^d$, les coefficients de la décomposition d'une fonction réelle $f\in L^2(\mathbb{R}^d,\mathcal{B}(\mathbb{R}^d),\mu;\mathbb{R})$ sur une base orthonormale $(\Phi_{1j_1}\otimes\cdots\otimes\Phi_{dj_d})_{\mathbf{j}\in\mathbb{N}^d}$, i.e.

$$f(\mathbf{x})=\sum_{\mathbf{k}\in\mathbb{N}^d}\mathrm{c}_{\mathbf{k}}\mathbf{\Phi}_{\mathbf{k}}(\mathbf{x})$$

alors

$$\sigma^2_{\mathfrak{u}}=\sum_{\mathbf{k}\in\mathbb{N}^*_{\mathfrak{u}}}|\mathrm{c}_{\mathbf{k}}|^2$$

où

$$\mathbb{N}^*_{\mathfrak{u}}=\{(k_1,\ldots,k_d)\mid\forall i\in\mathfrak{u},\ k_i\in\mathbb{N}^*\text{ et }\forall i\in\mathfrak{u}^c,\ k_i=0\}.$$

Les principaux exemples d'applications sont les décompositions en série trigonométrique (voir en particulier le Chapitre 5) ou suivant des polynômes orthogonaux associés à des lois de probabilités classiques qui conduisent à des décompositions en chaos polynomial généralisés, dits *de Wiener-Askey* — voir e.g. [Hoc72] pour des généralités sur les polynômes orthogonaux, et [Bla09] pour une étude applicative en analyse de sensibilité — dont la correspondance est reportée dans la Table 1.1.

	distribution	polynômes	support
continue	gaussienne	Hermite	$\mathbb{R}$
	exponentielle	Laguerre	$\mathbb{R}_+$
	gamma	Laguerre généralisés	$\mathbb{R}_+$
	beta	Jacobi	$[a,b]$
	uniforme	Legendre	$[a,b]$
discrète	Poisson	Charlier	$\{0,1,2,\ldots\}$
	binomiale	Krawtchouk	$\{0,1,\ldots,N\}$
	binomiale négative	Meixner	$\{0,1,2,\ldots\}$
	hypergéométrique	Hahn	$\{0,1,\ldots,N\}$

TABLE 1.1 – Distributions et familles de polynômes orthogonaux pour les chaos de Wiener-Askey.

1.3 Notes

1) De maniére analogue à la décomposition ANOVA, d'autres décompositions fonctionelles reposent sur la notion de projecteurs orthogonaux commutant deux à deux. Dans le domaine de l'approximation fonctionnelle, on peut citer par exemple la *décomposition ancrée* — en anglais, *anchored decomposition* — (voir e.g.[Gri05, KSWW12]), qu'on trouve également dans la littérature sous le nom de *cut-HDMR* (voir e.g. [RA99]).

2) Les conclusions des Théorèmes 1.1 et 1.2 — i.e. les décompositions de Tjur et de Helland — rendent explicite une propriété de la décomposition ANOVA qui n'apparaît ni dans le Corollaire 1.2

— i.e. la décomposition de Efron & Stein — ni dans la décomposition de Sobol' originale [Sob93]. En effet, les Théorèmes 1.1 et 1.2 établissent que l'espace étudié $\mathbb{R}^\Omega$ (resp. $L^2(\mathcal{E})$) se décompose en somme directe orthogonale de strates W_F, $F \in \mathfrak{F}$ (resp. $\mathcal{F} \in \mathfrak{F}$) i.e. que tout élément de cet espace se décompose de manière unique sur une famille de s.e.v. orthogonaux. Mais en outre, ils affirment que cette décomposition en somme directe — i.e. la donnée particulière des s.e.v. — est unique sous la contrainte que

$$\forall F \in \mathfrak{F},\ V_F = \bigoplus_{G \in \mathfrak{F}, F \preceq F}^{\perp} W_G \quad (\text{resp. } \forall \mathcal{F} \in \mathfrak{F},\ V_{\mathcal{F}} = \bigoplus_{\mathcal{G} \in \mathfrak{F}, \mathcal{G} \subseteq \mathcal{F}}^{\perp} W_{\mathcal{G}}).$$

3) Dans le domaine de l'intégration numérique, la notion de dimension effective a principalement permis de caractériser les fonctions dont le calcul de l'intégrale par une méthode de Monte Carlo peut être optimisé par une méthode de quasi-Monte Carlo. En effet, sans hypothèse particulière sur l'intégrande, une méthode de quasi-Monte Carlo ne peut théoriquement être plus efficace qu'une méthode de Monte Carlo que pour des tailles d'échantillons très grandes qui croissent avec la dimension (voir e.g. [Thi00]). Néanmoins, certaines applications, en particulier à la finance (voir e.g. [PT95]), montrent que les méthodes de quasi-Monte Carlo peuvent surpasser la méthode de Monte Carlo pour des dimensions élevées — $d = 360$ — et des tailles d'échantillons raisonnables. Dans ce cas, l'intégrande se révèle être la somme de fonctions de petites dimensions [CMO97]. Réciproquement Owen [Owe02] a montré que pour des intégrations numériques quasi-Monte Carlo relativement à des (t, m, s)-net randomisés [Owe97], si une fonction f est moins bien intégrée par une méthode de Monte Carlo, alors cette fonction possède une petite dimension effective au sens de la superposition.

Parallèlement à ces notions de dimension effective a été développée une théorie de l'intégration numérique sur des espaces de Hilbert à noyau reproduisant (RKHS) pondérés, comme les espaces de Sobolev ou de Korobov pondérés (voir e.g. [KSS12] pour une revue récente). Dans ce cadre, l'efficacité de la méthode de quasi-Monte Carlo — une quadrature sur un sous-groupe fini du tore, également appelée *lattice rule* en anglais (voir e.g. [SJ94]) — dépend des poids qui paramétrisent le RKHS. Comme nous le notons dans la Remarque 5.3, les cas particuliers dans lesquels la méthode de quasi-Monte Carlo montre des performances remarquables s'expliquent essentiellement par la dimension effective des fonctions appartenant à ces RKHS particuliers. Owen [Owe12b] a récemment formalisé ce lien dans des espaces de Sobolev pondérés en introduisant une nouvelle notion de dimension effective.

4) En se basant sur la décomposition ANOVA et sur les indices de Sobol' élémentaires, Liu & Owen [LO06] présentent une *distribution de la dimension* d'une fonction f, en considérant une variable aléatoire U prenant ses valeurs dans $\mathcal{P}\big([1:d]\big)$, avec pour tout $\mathfrak{u} \subseteq [1:d]$, $\mathbb{P}(U = \mathfrak{u}) = S_{\mathfrak{u}}(f)$. En définissant les moments successifs de $|U|$, ils résument les caractéristiques de la fonction f considérée. Par exemple, si le moment d'ordre 1, $\mu^{(1)} = \mathbb{E}\big[|U|\big]$, est égal à 1 alors f est additive. Le moment d'ordre 1, $\mu^{(1)}$ s'exprime explicitement en fonction des indices de Sobol' ascendants d'ordre 1

$$\mu^{(1)} = \sum_{i=1}^{d} \overline{S}_{\{i\}}.$$

5) Le chapitre qui précède, bien qu'il privilégie l'abstraction mathématique, ne doit pas faire oublier que la décomposition ANOVA d'une variable aléatoire $Y = f(X_1, \ldots, X_d)$ de carré intégrable où les X_i sont des v.a. indépendantes, est un objet bien concret :

$$Y = f_\emptyset + f_1(X_1) + \cdots + f_d(X_d) + f_{12}(X_1, X_2) + \cdots + f_{1\ldots d}(X_1, \ldots, X_d),$$

dont chacun des termes s'exprime de manière explicite comme

$$f_{\mathfrak{u}}(\mathbf{X}_{\mathfrak{u}}) = \sum_{\mathfrak{v} \subseteq \mathfrak{u}} (-1)^{|\mathfrak{u}|-|\mathfrak{v}|} \mathbb{E}\big[Y | \sigma(\mathbf{X}_{\mathfrak{v}})\big], \quad \mathfrak{u} \subseteq [1:d].$$

Chapitre 2

Estimation des indices de Sobol'

Nous avons pu constater dans le chapitre précédent que les indices de Sobol' sont définis par des intégrales. Par conséquent l'approche la plus directe pour évaluer ces indices consiste à procéder à une intégration numérique. Nous nous intéressons dans ce chapitre uniquement à ces approches directes, en omettant volontairement les méthodes consistant à approcher le modèle d'étude par un *métamodèle* puis à évaluer les indices de Sobol' du métamodèle ; la réponse donnée par ces méthodes indirectes consistant essentiellement à troquer un problème d'*intégration numérique* pour un problème d'*approximation numérique*. Même si la différence entre ces deux approches n'est pas toujours marquée, nous nous concentrons ici principalement sur les méthodes d'intégration.

Elles se divisent en deux familles très distinctes : d'une part les méthodes exploitant la décomposition ANOVA classique mettant en avant les espérances conditionnelles, et d'autre part les méthodes exploitant une décomposition ANOVA spectrale relative à une base orthonormale particulière. Dans la première famille, nous présentons successivement la méthode de Sobol' qui est la méthode de Monte Carlo naturelle pour l'estimation des indices de Sobol' (Section 2.1), puis la méthode de McKay qui, bien qu'elle soit peu utilisée en pratique, constitue un apport technique considérable dans les aspects combinatoires de l'estimation des indices de Sobol' (Section 2.2). Nous abordons ensuite la famille des méthodes relatives à une décomposition ANOVA spectrale. Nous commençons par une revue des méthodes Fourier Amplitude Sensitivity Test (FAST), Random Balance Design (RBD) et RBD-FAST telles qu'elles ont été introduites jusqu'à l'année 2010 (Section 2.3). Ces approches, toutes basées sur une décomposition en série de Fourier, n'étaient jusqu'à présent pas rigoureusement établies d'un point de vue mathématique. Nous avons complété leurs bases théoriques, et les avons généralisées dans cette thése (voir Chapitres 4 et 5). Enfin, nous terminons en présentant comment appréhender l'estimation des indices de Sobol' en considérant une décomposition ANOVA spectrale quelconque (Section 2.4). Une section de notes clôt le chapitre.

2.1 Méthode de Sobol'

Dans cette section, nous reprenons le cadre du chapitre précédent et nous considérons un modèle $Y = f(X_1, \dots, X_d)$ où les X_i, $i \in [1:d]$ sont des variables aléatoires indépendantes et f une fonction réelle telle que $\mathbb{E}[Y^2] < +\infty$. On désigne par $\mathbf{X}^j$, et $\mathbf{Z}^j$, $j \in [1:n]$ des vecteurs aléatoires indépendants identiquement distribués suivant la loi de $\mathbf{X}$. Nous rappelons également que pour tout $\mathfrak{u} \subseteq [1:d]$, la notation $\mathbf{X}_\mathfrak{u} : \mathbf{Z}_{-\mathfrak{u}}$ — respectivement $\mathbf{X}^j_\mathfrak{u} : \mathbf{Z}^j_{-\mathfrak{u}}$ — désigne le vecteur dont les composantes indicées par $i \in \mathfrak{u}$ sont les X_i, et dont les autres composantes sont les Z_i — respectivement les X^j_i et les Z^j_i. Enfin, pour tout $\mathfrak{u} \subseteq [1:d]$ on introduit les notations

$$\begin{aligned} Y_\mathfrak{u} &= f(\mathbf{X}_\mathfrak{u} : \mathbf{Z}_{-\mathfrak{u}}) \\ Y^j_\mathfrak{u} &= f(\mathbf{X}^j_\mathfrak{u} : \mathbf{Z}^j_{-\mathfrak{u}}), \quad j \in [1:n]. \end{aligned}$$

Les indices de Sobol' se définissant à l'aide d'intégrales, une méthode naturelle d'estimation consiste à évaluer ces intégrales par la méthode de Monte Carlo. La méthode ainsi obtenue est généralement appelée *méthode de Sobol'* et quelquefois *Sobol' Pick and Freeze* (SPF). Les caractéristiques d'une telle méthode sont donc celles de la méthode de Monte Carlo classique :

- hypothèses d'application peu restrictives, typiquement une condition d'intégrabilité sur la fonction f
- vitesse de convergence indépendante de la dimension, mais seulement en $O(n^{-1/2})$ pour une taille d'échantillon égale à n.

La méthode de Sobol' possède en outre une caractéristique spécifique due à la définition de ses différents estimateurs, au sens où elle nécessite deux échantillons distincts pour estimer un seul indice de Sobol'. Comme nous le présentons par la suite, certaines optimisations permettent néanmoins de pallier cet inconvénient. L'exposé qui suit s'oriente principalement sur la question de l'estimation des indices $\underline{S}_{\mathfrak{u}}$ et $\overline{S}_{\mathfrak{u}}$, $\mathfrak{u} \subseteq [1:d]$, ainsi que de leurs numérateurs respectifs $\underline{\tau}^2_{\mathfrak{u}}$ et $\overline{\tau}^2_{\mathfrak{u}}$; l'estimation de leur dénominateur σ^2 étant généralement éludée en considérant l'estimateur de la variance empirique

$$\dot{\sigma}^2_n = \frac{1}{n}\sum_{j=1}^{n}\left(f(\mathbf{X}^j) - \frac{1}{n}\sum_{k=1}^{n} f(\mathbf{X}^k)\right)^2. \tag{2.1}$$

Le premier estimateur de $\underline{\tau}^2_{\mathfrak{u}}$ a été introduit par Sobol' en 1990 [Sob93] et indépendemment discuté à la même période par Hora et Iman [HI86, IH90a] et Ishigami et Homma [IH89]. Il est défini par

$$\dot{\underline{\tau}}^2_{\mathfrak{u},n} = \frac{1}{n}\sum_{j=1}^{n} f(\mathbf{X}^j) f(\mathbf{X}^j_{\mathfrak{u}} : \mathbf{Z}^j_{-\mathfrak{u}}) - \left(\frac{1}{n}\sum_{j=1}^{n} f(\mathbf{X}^j)\right)^2, \quad \mathfrak{u} \subseteq [1:d]. \tag{2.2}$$

Quant à la mesure d'importance $\overline{\tau}^2_{\mathfrak{u}}$ et son estimateur Monte Carlo associé, ils ont été introduits par Homma et Saltelli [HS96]. L'estimateur est défini par

$$\dot{\overline{\tau}}^2_{\mathfrak{u},n} = \frac{1}{n}\sum_{j=1}^{n} f(\mathbf{X}^j)^2 - \frac{1}{n}\sum_{j=1}^{n} f(\mathbf{X}^j) f(\mathbf{X}^j_{-\mathfrak{u}} : \mathbf{Z}^j_{\mathfrak{u}}). \tag{2.3}$$

Les deux estimateurs définis dans les Formules (2.2) et (2.3), ainsi que l'estimateur de la variance (2.1) ont subi des modifications jusqu'à très récemment. Nous les présentons au fil des trois sections thématiques suivantes. La Section 2.1.1 traite de l'estimation de l'erreur commise lors du calcul de l'indice de Sobol', la Section 2.1.2 présente la question de l'optimisation combinatoire en vue de réduire le coût de calcul, et enfin la Section 2.1.3 est consacrée à l'accélération de la convergence des différents estimateurs.

2.1.1 Estimation de l'erreur

La question du calcul de l'erreur d'estimation des indices de Sobol' est traitée sous deux formes différentes dès le milieu des années 1990. D'une part Homma et Saltelli [HS96] proposent de calculer approximativement l'*erreur probable* commise lors de l'estimation des indices de Sobol'. Comme nous le constaterons par la suite, leur travail est étroitement lié à la construction d'intervalles de confiance pour des estimateurs asymptotiquement normaux. D'autre part, Archer et al. [ASS97] dérivent des intervalles de confiance par la technique du bootstrap (voir e.g. [ET93]). Dans les deux cas, l'analyse est orientée vers les indices de Sobol' $\underline{S}_{\mathfrak{u}}$, $\mathfrak{u} \subseteq [1:d]$ mais elle se généralise naturellement aux $\overline{S}_{\mathfrak{u}}$, $\mathfrak{u} \subseteq [1:d]$. Plus récemment, Janon et al. [JKL$^+$12] ont complété la démarche approximative de Homma et Saltelli en montrant rigoureusement, par une application de la Delta méthode (voir e.g. [Van98]) que l'estimateur Monte Carlo des indices de Sobol' descendants est asymptotiquement normal. Nous revenons en détail sur l'approche originale de Homma et Saltelli [HS96] ainsi que sur les résultats théoriques obtenus par Janon et al. [JKL$^+$12] via la Delta méthode. Nous omettons volontairement la méthode de bootstrap développée par Archer et al. [ASS97] qui n'est pas spécifique à l'estimation des indices de Sobol'.

Approche originale de Homma et Saltelli

En négligeant les erreurs d'estimation de la moyenne de Y

$$\varepsilon(\mathfrak{u}) = \mathbb{E}[Y] - \frac{1}{n}\sum_{j=1}^{n} f(\mathbf{X}^j_{\mathfrak{u}} : \mathbf{Z}^j_{-\mathfrak{u}}), \quad \mathfrak{u} \subseteq [1:d],$$

et

$$\varepsilon'(\mathfrak{u}) = \mathbb{E}[Y] - \frac{1}{n}\sum_{j=1}^{n} f(\mathbf{X}^j), \quad \mathfrak{u} \subseteq [1:d]$$

— ce qui revient à supposer que la moyenne de Y est connue — les estimateurs définis dans les Formules (2.1) et (2.2) peuvent être respectivement approchés par les moyennes empiriques

$$\dot{m}_n(\sigma^2) = \frac{1}{n}\sum_{j=1}^{n} \big(f(\mathbf{X}^j) - \mathbb{E}[Y]\big)^2$$

et

$$\dot{m}_n(\underline{\tau}^2_{\mathfrak{u}}) = \frac{1}{n}\sum_{j=1}^{n} \big(f(\mathbf{X}^j) - \mathbb{E}[Y]\big)\big(f(\mathbf{X}^j_{\mathbf{u}} : \mathbf{Z}^j_{-\mathbf{u}}) - \mathbb{E}[Y]\big), \quad \mathfrak{u} \subseteq [1:d].$$

Par suite, sous cette approximation, le théorème Central Limit permet de conclure à la normalité asymptotique des estimateurs $\dot{\sigma}^2$ et $\underline{\dot{\tau}}^2_{\mathfrak{u}}$. En posant

$$\dot{s}_n(\sigma^2) = \left(\frac{1}{n-1}\sum_{j=1}^{n} \Big(\big(f(\mathbf{X}^j) - \mathbb{E}[Y]\big)^2 - \dot{m}_n(\sigma^2)\Big)^2\right)^{1/2}$$

et pour tout $\mathfrak{u} \subseteq [1:d]$

$$\dot{s}_n(\underline{\tau}^2_{\mathfrak{u}}) = \left(\frac{1}{n-1}\sum_{j=1}^{n} \Big(\big(f(\mathbf{X}^j) - \mathbb{E}[Y]\big)\big(f(\mathbf{X}^j_{\mathbf{u}} : \mathbf{Z}^j_{-\mathbf{u}}) - \mathbb{E}[Y]\big) - \dot{m}_n(\underline{\tau}^2_{\mathfrak{u}})\Big)^2\right)^{1/2},$$

on obtient alors les intervalles de confiance asymptotiques au seuil α

$$\begin{aligned}
I_{\alpha,n}(\sigma^2) &= \left[\dot{\sigma}^2_n - \delta(\dot{\sigma}^2_n)\ ,\ \dot{\sigma}^2_n + \delta(\dot{\sigma}^2_n)\right], \text{ avec } \delta(\dot{\sigma}^2_n) = u_{1-\alpha/2}\frac{\dot{s}_n(\sigma^2)}{\sqrt{n}},\\
I_{\alpha,n}(\underline{\tau}^2_{\mathfrak{u}}) &= \left[\underline{\dot{\tau}}^2_{\mathfrak{u},n} - \delta(\underline{\dot{\tau}}^2_{\mathfrak{u},n})\ ,\ \underline{\dot{\tau}}^2_{\mathfrak{u},n} + \delta(\underline{\dot{\tau}}^2_{\mathfrak{u},n})\right], \text{ avec } \delta(\underline{\dot{\tau}}^2_{\mathfrak{u},n}) = u_{1-\alpha/2}\frac{\dot{s}_n(\underline{\tau}^2_{\mathfrak{u}})}{\sqrt{n}}
\end{aligned}$$

où $u_{1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ de la loi normale centrée réduite. Les $\delta(\underline{\dot{\tau}}^2_{\mathfrak{u},n})$ et $\delta(\dot{\sigma}^2_n)$ sont appelés *erreurs probables* par Homma et Saltelli [HS96], elles permettent aux auteurs de déduire une erreur probable pour l'estimateur $\underline{\dot{S}}_{\mathfrak{u},n} = \underline{\dot{\tau}}^2_{\mathfrak{u},n}/\dot{\sigma}^2_n$, par un raisonnement approximatif. En effet, en considérant que, si les inégalités

$$\begin{aligned}
\dot{\sigma}^2_n - \delta(\dot{\sigma}^2_n) &\leq \sigma^2 \leq \dot{\sigma}^2_n + \delta(\dot{\sigma}^2_n),\\
\underline{\dot{\tau}}^2_{\mathfrak{u},n} - \delta(\underline{\dot{\tau}}^2_{\mathfrak{u},n}) &\leq \underline{\tau}^2_{\mathfrak{u}} \leq \underline{\dot{\tau}}^2_{\mathfrak{u},n} + \delta(\underline{\dot{\tau}}^2_{\mathfrak{u},n})
\end{aligned}$$

sont vérifiées avec une forte probabilité alors il en est de même pour

$$\frac{\underline{\dot{\tau}}^2_{\mathfrak{u},n} - \delta(\underline{\dot{\tau}}^2_{\mathfrak{u},n})}{\dot{\sigma}^2_n + \delta(\dot{\sigma}^2_n)} \leq \underline{S}_{\mathfrak{u}} \leq \frac{\underline{\dot{\tau}}^2_{\mathfrak{u},n} + \delta(\underline{\dot{\tau}}^2_{\mathfrak{u},n})}{\dot{\sigma}^2_n - \delta(\dot{\sigma}^2_n)}.$$

Puis en notant que

$$\begin{aligned}
\frac{\underline{\dot{\tau}}^2_{\mathfrak{u},n} - \delta(\underline{\dot{\tau}}^2_{\mathfrak{u},n})}{\dot{\sigma}^2_n + \delta(\dot{\sigma}^2_n)} &= \frac{\underline{\dot{\tau}}^2_{\mathfrak{u},n}\dot{\sigma}^2_n - \delta(\underline{\dot{\tau}}^2_{\mathfrak{u},n})\dot{\sigma}^2_n - \underline{\dot{\tau}}^2_{\mathfrak{u},n}\delta(\dot{\sigma}^2_n) + \delta(\underline{\dot{\tau}}^2_{\mathfrak{u},n})\delta(\dot{\sigma}^2_n)}{(\dot{\sigma}^2_n)^2 - \delta(\dot{\sigma}^2_n)^2}\\
&\approx \underline{\dot{S}}_{\mathfrak{u},n} - \frac{\delta(\underline{\dot{\tau}}^2_{\mathfrak{u},n})\dot{\sigma}^2_n + \underline{\dot{\tau}}^2_{\mathfrak{u},n}\delta(\dot{\sigma}^2_n)}{(\dot{\sigma}^2_n)^2}
\end{aligned}$$

et de manière identique que

$$\frac{\underline{\dot{\tau}}^2_{\mathfrak{u},n} + \delta(\underline{\dot{\tau}}^2_{\mathfrak{u},n})}{\dot{\sigma}^2_n - \delta(\dot{\sigma}^2_n)} \approx \underline{\dot{S}}_{\mathfrak{u},n} + \frac{\delta(\underline{\dot{\tau}}^2_{\mathfrak{u},n})\dot{\sigma}^2_n + \underline{\dot{\tau}}^2_{\mathfrak{u},n}\delta(\dot{\sigma}^2_n)}{(\dot{\sigma}^2_n)^2},$$

on peut définir l'erreur probable de l'estimateur $\dot{\underline{S}}_{\mathfrak{u},n}$ par

$$\delta(\dot{\underline{S}}_{\mathfrak{u},n}) = \frac{\delta(\dot{\underline{\tau}}^2_{\mathfrak{u},n}) + \dot{\underline{S}}_{\mathfrak{u},n}\delta(\dot{\sigma}^2_n)}{\dot{\sigma}^2_n}.$$

Dans la dérivation précédente, l'obstacle majeur à une construction rigoureuse d'un intervalle de confiance pour $\dot{\underline{S}}_{\mathfrak{u},n}$ est principalement le passage au quotient qui, en toute rigueur, ne permet pas de conserver la propriété de normalité asymptotique. L'outil qui peut pallier cette difficulté est la Delta méthode, elle a été appliquée par Janon et al. très récemment [JKL$^+$12].

Delta méthode et normalité asymptotique

Janon et al. [JKL$^+$12] introduisent un nouvel estimateur des indices de Sobol' $\underline{S}_{\mathfrak{u}}$, $\mathfrak{u} \subseteq [1:d]$ et en réintroduisent un second [MNM06]. Nous les notons respectivement $\widetilde{\underline{S}}_{\mathfrak{u},n} = \widetilde{\underline{\tau}}^2_{\mathfrak{u},n}/\widetilde{\sigma}^2_n$ et $\widehat{\underline{S}}_{\mathfrak{u},n} = \widehat{\underline{\tau}}^2_{\mathfrak{u},n}/\widehat{\sigma}^2_n$, et ils sont définis par

$$\begin{aligned}
\widetilde{\sigma}^2_n &= \frac{1}{n}\sum_{j=1}^n \left(f(\mathbf{X}^j) - \frac{1}{n}\sum_{k=1}^n f(\mathbf{X}^k)\right)^2 \\
\widetilde{\underline{\tau}}^2_{\mathfrak{u},n} &= \frac{1}{n}\sum_{j=1}^n f(\mathbf{X}^j) f(\mathbf{X}^j_{\mathfrak{u}} : \mathbf{Z}^j_{-\mathfrak{u}}) - \left(\frac{1}{n}\sum_{j=1}^n f(\mathbf{X}^j)\right)\left(\frac{1}{n}\sum_{j=1}^n f(\mathbf{X}^j_{\mathfrak{u}} : \mathbf{Z}^j_{-\mathfrak{u}})\right), \quad \mathfrak{u} \subseteq [1:d] \\
\widehat{\sigma}^2_n &= \frac{1}{n}\sum_{j=1}^n \left(\frac{f(\mathbf{X}^j)^2 + f(\mathbf{X}^j_{\mathfrak{u}} : \mathbf{Z}^j_{-\mathfrak{u}})^2}{2}\right) - \left(\frac{1}{n}\sum_{j=1}^n \frac{f(\mathbf{X}^j) + f(\mathbf{X}^j_{\mathfrak{u}} : \mathbf{Z}^j_{-\mathfrak{u}})}{2}\right)^2 \\
\widehat{\underline{\tau}}^2_{\mathfrak{u},n} &= \frac{1}{n}\sum_{j=1}^n f(\mathbf{X}^j) f(\mathbf{X}^j_{\mathfrak{u}} : \mathbf{Z}^j_{-\mathfrak{u}}) - \left(\frac{1}{n}\sum_{j=1}^n \frac{f(\mathbf{X}^j) + f(\mathbf{X}^j_{\mathfrak{u}} : \mathbf{Z}^j_{-\mathfrak{u}})}{2}\right)^2, \quad \mathfrak{u} \subseteq [1:d].
\end{aligned}$$

Du fait de la loi forte des grands nombres, il est aisé de remarquer que les estimateurs $\dot{\underline{S}}_{\mathfrak{u},n}$, $\widetilde{\underline{S}}_{\mathfrak{u},n}$ et $\widehat{\underline{S}}_{\mathfrak{u},n}$ sont fortement consistants — i.e. ils convergent presque sûrement vers l'indice de Sobol' $\underline{S}_{\mathfrak{u}}$. Janon et al. ont montré, en outre, qu'ils sont asymptotiquement normaux.

Théorème 2.1. [Janon, Klein, Lagnoux, Nodet et Prieur, 2012] *Si $\mathbb{E}[Y^4]$ est fini, alors pour tout $\mathfrak{u} \subseteq [1:d]$, on a*

$$\begin{aligned}
\sqrt{n}(\dot{\underline{S}}_{\mathfrak{u},n} - \underline{S}_{\mathfrak{u}}) &\xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,\sigma_1), \\
\sqrt{n}(\widetilde{\underline{S}}_{\mathfrak{u},n} - \underline{S}_{\mathfrak{u}}) &\xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,\sigma_2), \\
\sqrt{n}(\widehat{\underline{S}}_{\mathfrak{u},n} - \underline{S}_{\mathfrak{u}}) &\xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,\sigma_3)
\end{aligned}$$

avec

$$\sigma_1 = \sigma_2 = \frac{\mathrm{Var}\Big[\big(Y - \mathbb{E}[Y]\big)\big(Y_{\mathfrak{u}} - \mathbb{E}[Y]\big) - \underline{S}_{\mathfrak{u}}\big(Y - \mathbb{E}[Y]\big)^2\Big]}{\mathrm{Var}[Y]^2}$$

et

$$\sigma_3 = \frac{\mathrm{Var}\Big[\big(Y - \mathbb{E}[Y]\big)\big(Y_{\mathfrak{u}} - \mathbb{E}[Y]\big) - \underline{S}_{\mathfrak{u}}/2\Big(\big(Y - \mathbb{E}[Y]\big)^2 + \big(Y_{\mathfrak{u}} - \mathbb{E}[Y]\big)^2\Big)\Big]}{\mathrm{Var}[Y]^2}.$$

Démonstration. Pour les deux estimateurs introduits par Janon et al., nous renvoyons à la preuve de la Proposition 2.2 dans [JKL$^+$12]. Pour l'estimateur $\dot{\underline{S}}_{\mathfrak{u},n}$, il est aisé de conclure en s'inspirant de cette même preuve. □

En outre, en reprenant les notations du théorème précédent, on a toujours $\sigma_3 \leq \sigma_1$, et l'égalité est stricte dès lors que $\underline{S}_{\mathfrak{u}} = 0$ ou 1. Ce résultat est complété par la propriété d'efficacité asymptotique de $\widehat{\underline{S}}_{\mathfrak{u},n}$ parmi une famille d'estimateurs très généraux — voir Proposition 2.5 dans [JKL$^+$12].

2.1.2 Optimisations combinatoires

Comme nous l'avons remarqué en introduction de cette section, les indices de Sobol' sont coûteux à calculer par la méthode de Sobol' car leur estimation nécessite un certain nombre d'échantillons distincts. En effet, par définition, les $\underline{\dot{\tau}}^2_{\mathfrak{u},n}$ — respectivement les $\dot{\overline{\tau}}^2_{\mathfrak{u},n}$ — sont des quantités qui nécessitent l'utilisation des deux échantillons $\{\mathbf{X}^j\}_{j\in[1:n]}$ et $\{\mathbf{X}^j_{\mathfrak{u}} : \mathbf{Z}^j_{-\mathfrak{u}}\}_{j\in[1:n]}$ — respectivement $\{\mathbf{X}^j\}_{j\in[1:n]}$ et $\{\mathbf{X}^j_{-\mathfrak{u}} : \mathbf{Z}^j_{\mathfrak{u}}\}_{j\in[1:n]}$ — et les estimateurs de $\sigma^2_{\mathfrak{u}}$, $\mathfrak{u}\subseteq[1:d]$, dérivés des estimateurs respectifs des $\underline{\tau}^2_{\mathfrak{v}}$, $\mathfrak{v}\subseteq\mathfrak{u}$, par la formule de Möbius nécessitent théoriquement $2^{|\mathfrak{u}|}$ échantillons. Il est néanmoins trivial de constater qu'il est possible d'estimer tous les indices $\underline{\tau}^2_{\{i\}}$, $i\in[1:d]$, ou tous les indices $\overline{\tau}^2_{\{i\}}$, $i\in[1:d]$, à l'aide uniquement de $d+1$ échantillons en mutualisant l'échantillon $\{\mathbf{X}^j\}_{j\in[1:n]}$. Des résultats plus généraux ont été énoncés par Saltelli [Sal02], et une analyse extensive de ces questions combinatoires a été récemment présentée par Owen [Owe12c]. Le travail de ce dernier permet en particulier de retrouver les résultats originaux de Saltelli. Nous donnons un aperçu non exhaustif des notions et des résultats introduits par Owen.

Définition des indices de Sobol' généralisés

L'objet central de tout ce qui suit est l'indice de Sobol' généralisé (ISG) défini par

$$\sum_{\mathfrak{u}\subseteq[1:d]}\sum_{\mathfrak{v}\subseteq[1:d]}\lambda(\mathfrak{u},\mathfrak{v})\mathbb{E}\big[f(\mathbf{X}_{\mathfrak{u}} : \mathbf{Z}_{-\mathfrak{u}})f(\mathbf{X}_{\mathfrak{v}} : \mathbf{Z}_{-\mathfrak{v}})\big]$$

où les $\lambda(\mathfrak{u},\mathfrak{v})$ sont des coefficients réels, et où il est important de noter que

$$\mathbb{E}\big[f(\mathbf{X}_{\mathfrak{u}} : \mathbf{Z}_{-\mathfrak{u}})f(\mathbf{X}_{\mathfrak{v}} : \mathbf{Z}_{-\mathfrak{v}})\big] = \mu^2 + \underline{\tau}^2_{-(\mathfrak{u}\Delta\mathfrak{v})}.$$

L'ensemble de ces indices forme un espace vectoriel de dimension 2^{2d} alors que les indices de Sobol' élémentaires, ascendants ou descendants sont chacuns au nombre de 2^d. Il est donc nécessaire de classer les ISG suivant un certain nombre de propriétés particulières. Ainsi, un ISG est dit *bilinéaire* dès lors que pour tous $\mathfrak{u}$ et $\mathfrak{v}$ inclus dans $[1:d]$, $\lambda(\mathfrak{u},\mathfrak{v}) = \lambda_1(\mathfrak{u})\lambda_2(\mathfrak{v})$. Un ISG bilinéaire tel que pour tout $\mathfrak{u}\subseteq[1:d]$, $\lambda_1(\mathfrak{u}) = \lambda_2(\mathfrak{u})$ est dit *carré*. Un ISG bilinéaire tel que pour tout $\mathfrak{v}\subseteq[1:d]$,

$$\lambda_2(\mathfrak{v}) = \begin{cases} 1 & \text{si } \mathfrak{v}=\emptyset \\ 0 & \text{sinon} \end{cases}$$

est dit *simple*. Enfin, un ISG vérifiant $\sum_{\mathfrak{u}\in[1:d]}\sum_{\mathfrak{v}\in[1:d]}\lambda(\mathfrak{u},\mathfrak{v}) = 0$ est un *contraste* ; il a la particularité de ne pas dépendre de μ^2 et par conséquent permet de définir des estimateurs Monte Carlo non biaisés.

Coût d'un estimateur

Nous supposons ici que chaque terme, $f(\mathbf{X}_{\mathfrak{u}} : \mathbf{Z}_{-\mathfrak{u}})$ ou $f(\mathbf{X}_{\mathfrak{v}} : \mathbf{Z}_{-\mathfrak{v}})$ dans le développement d'un ISG est estimé par un échantillon de taille n. L'évaluation du coût d'un ISG se réduit donc à connaître le nombre de termes distincts dans le développement de ce dernier. Notons $\boldsymbol{\lambda} = \big(\lambda(\mathfrak{u},\mathfrak{v})\big)_{\mathfrak{u},\mathfrak{v}\in[1:d]}$ la matrice carrée de taille 2^d, et définissons

$$C_{\mathfrak{u}}(\boldsymbol{\lambda}) = \begin{cases} 1 & \text{si } \exists\ \mathfrak{v}\in[1:d] \text{ tel que } \lambda(\mathfrak{u},\mathfrak{v})\neq 0 \\ 0 & \text{sinon.} \end{cases}$$

Alors le coût de l'ISG basé sur $\boldsymbol{\lambda}$ est

$$C(\boldsymbol{\lambda}) = \sum_{\mathfrak{u}\subseteq[1:d]}\big(C_{\mathfrak{u}}(\boldsymbol{\lambda}) + C_{\mathfrak{u}}(\boldsymbol{\lambda}^\intercal) - C_{\mathfrak{u}}(\boldsymbol{\lambda})C_{\mathfrak{u}}(\boldsymbol{\lambda}^\intercal)\big).$$

Pour les détails concernant ce dénombrement, nous renvoyons à [Owe12c] Section 3.3.

	$\emptyset$	j	$-j$	k	$-k$	$[1:d]$
$\emptyset$	$[1:d]$	$-j$	j	$-k$	k	$\emptyset$
j	$-j$	$[1:d]$	$\emptyset$	$-\{j,k\}$	$\{j,k\}$	j
$-j$	j	$\emptyset$	$[1:d]$	$\{j,k\}$	$-\{j,k\}$	$-j$
$[1:d]$	$\emptyset$	j	$-j$	k	$-k$	$[1:d]$

TABLE 2.1 – Table des ensembles $-(\mathfrak{u}\Delta\mathfrak{v})$. $\mathfrak{u}$ et $\mathfrak{v}$ sont respectivement lus sur la première colonne et la première ligne du tableau.

Optimisations combinatoires

Owen [Owe12c] montre que l'estimation des indices de Sobol' élémentaires peut être allégée considérablement en notant que la part de variance $\sigma^2_{\mathfrak{u}}$ s'exprime comme un ISG bilinéaire dont le coût est au plus $2^{\lfloor|\mathfrak{u}|/2\rfloor+1}$, alors que le coût normal obtenu en utilisant la formule de Möbius est $2^{|\mathfrak{u}|}$. Plus explicitement, il énonce le résultat suivant.

Théorème 2.2. [Owen, 2012] *Soit $\mathfrak{u} \subseteq [1:d]$, alors pour tout $\mathfrak{v}_1 \subseteq \mathfrak{u}$, on note $\mathfrak{v}_2 = \mathfrak{u} \setminus \mathfrak{v}_1$, et on a*

$$\sigma^2_{\mathfrak{u}} = \sum_{\mathfrak{w}_1 \in \mathfrak{v}_1} \sum_{\mathfrak{w}_2 \in \mathfrak{v}_2} (-1)^{|\mathfrak{v}_1 - \mathfrak{w}_1| + |\mathfrak{v}_2 - \mathfrak{w}_2|} \mathbb{E}\big[f(\mathbf{X}_{\mathfrak{w}_1} : \mathbf{Z}_{-\mathfrak{w}_1}) f(\mathbf{X}_{\mathfrak{w}_2 \cup -\mathfrak{u}} : \mathbf{Z}_{-(\mathfrak{w}_2 \cup -\mathfrak{u})})\big].$$

Démonstration. Nous renvoyons à la preuve du Théorème 3 dans [Owe12c]. □

Sur le plan de l'estimation des indices de Sobol' ascendants et descendants, on retrouve de manière naturelle les résultats de Saltelli [Sal02]. En rappelant que

$$\mathbb{E}\big[f(\mathbf{X}_{\mathfrak{u}} : \mathbf{Z}_{-\mathfrak{u}}) f(\mathbf{X}_{\mathfrak{v}} : \mathbf{Z}_{-\mathfrak{v}})\big] = \mu^2 + \underline{\tau}^2_{-(\mathfrak{u}\Delta\mathfrak{v})}$$

et en établissant la table des ensembles $-\mathfrak{u}\Delta\mathfrak{v}$ suivante on aboutit aux deux résultats de Saltelli.

Théorème 2.3. [Saltelli, 2002] *Les $d+2$ échantillons $Y^j_{\mathfrak{u}} = f(\mathbf{X}^j_{\mathfrak{u}} : \mathbf{Z}^j_{-\mathfrak{u}})$, $j = [1:n]$ avec $\mathfrak{u} \in \{\emptyset, \{1\}, \{2\}, \ldots, \{d\}, [1:d]\}$ permettent d'estimer tous les indices de Sobol' descendants $\underline{\tau}^2_j$, $j \in [1:d]$, et ascendants $\overline{\tau}^2_j$, $j \in [1:d]$ et $\overline{\tau}^2_{\{j,k\}}$, $j,k \in [1:d]$.*

Théorème 2.4. [Saltelli, 2002] *Les $2d+2$ échantillons $Y^j_{\mathfrak{u}} = f(\mathbf{X}^j_{\mathfrak{u}} : \mathbf{Z}^j_{-\mathfrak{u}})$, $j = [1:n]$ avec $\mathfrak{u} \in \{\emptyset, \{1\}, \ldots, \{d\}, -\{1\}, \ldots, -\{d\}, [1:d]\}$ permettent d'estimer tous les indices de Sobol' descendants $\underline{\tau}^2_j$, $j \in [1:d]$ et $\underline{\tau}^2_{\{j,k\}}$, $j,k \in [1:d]$, et ascendants $\overline{\tau}^2_j$, $j \in [1:d]$ et $\overline{\tau}^2_{\{j,k\}}$, $j,k \in [1:d]$.*

Le formalisme combinatoire précédent peut laisser espérer d'autres résultats intéressants quant à la minimisation du nombre d'échantillons distincts lors de l'estimation d'un indice ou de plusieurs indices de Sobol' ; Owen [Owe12c] donne d'ailleurs des applications autres que les deux principales présentées ci-dessus. De notre côté, nous proposons une autre approche combinatoire pour optimiser la méthode de Sobol'. Dans notre travail, la combinatoire n'opère plus de façon *globale* en combinant plusieurs échantillons, mais de manière *locale* en ne considérant que deux échantillons distincts et en manipulant leur structure interne. Dans ce cadre, on est amené à considérer des *hypercubes latins* [MCB79], des hypercubes latins basés sur des *tableaux orthogonaux* [Owe92, Tan93] et enfin, les *répliqués* [McK95] de ceux-ci. Notre travail figure au Chapitre 6 sous la forme d'une prépublication intitulée *Estimating Sobol' indices combining Monte Carlo estimators and Latin hypercube sampling* [TP12b], et les notions précédentes sont discutées dès la Section 2.2.

2.1.3 Accélération de la convergence

La question de l'accélération de la convergence de la méthode de Sobol' est abondamment discutée dans la littérature, principalement pour $\underline{\tau}^2_{\mathfrak{u}}$. Elle se pose sous deux formes principales.

La première approche consiste à accélérer la convergence de la méthode en améliorant l'échantillon, i.e. typiquement en considérant des *suites à discrépance faible* [Nie92] en lieu et place des échantillons *aléatoires*[1] où tous les points sont supposés être des réalisations indépendantes de $\mathbf{X}$. Cette méthode

1. Dans la réalité, les échantillons *aléatoires*, ne font que simuler l'aléatoire, et en toute rigueur, on devrait parler d'échantillons *pseudo-aléatoires* (voir par exemple [Thi00] Section 2.2.)

de quasi-Monte Carlo est évoquée par Sobol' dès 1990 [Sob93] lorsqu'il mentionne l'éventualité d'une utilisation de séquences *quasi-aléatoires*. Elle a été reprise dans la littérature un certain nombre de fois comme par exemple dans [SAA+10] où sont exploitées les séquences LP_τ [Sob67]. Cette approche possède néanmoins plusieurs inconvénients. D'abord, les vitesses de convergence ne sont valides que si la fonction considérée est à variation bornée au sens de Hardy et Krause (voir par exemple [Nie92]) et comme le fait remarquer Thiémard [Thi00], certaines fonctions de faible complexité ne le sont pas, comme par exemple

$$f(x,y) = \begin{cases} 1 & \text{si } y \leq x \\ 0 & \text{sinon.} \end{cases}$$

Ensuite, la taille minimale de l'échantillon nécessaire pour obtenir une vitesse de convergence supérieure à celle de la méthode de Monte Carlo augmente avec la dimension (voir [Thi00]). Et enfin, le calcul d'une borne d'erreur de l'estimation suppose de disposer d'une majoration de la variation de la fonction ainsi que de la discrépance à l'origine de la suite quasi-aléatoire considérées. Or le calcul de telles majorations reste un problème souvent non trivial (voir [Thi00]).

La seconde approche consiste à améliorer l'estimateur lui-même afin de minimiser sa variance. Elle est discutée dès 1990 par Sobol' [Sob93] qui mentionne que lorsque l'espérance de Y est grande, on constate empiriquement que le recentrage des données par une constante c proche de la moyenne — typiquement la moyenne empirique des données — améliore la précision de l'estimateur. Cette technique est discutée par Sobol' et Myshetskaya [SM07] et reprise par Owen récemment [Owe12a]. On constate sur des exemples son effet positif.

Dans un second temps, Homma et Saltelli [HS96] remarquent que si pour $\mathfrak{u} \subseteq \mathcal{D}$, $\underline{\tau}^2_{\mathfrak{u}} = 0$ alors l'estimateur $\dot{\underline{\tau}}^2_{\mathfrak{u},n}$ vérifie

$$\begin{aligned} \dot{\underline{\tau}}^2_{\mathfrak{u},n} &= \frac{1}{n}\sum_{j=1}^{n} f(\mathbf{X}^j) f(\mathbf{X}^j_{\mathbf{u}} : \mathbf{Z}^j_{-\mathbf{u}}) - \left(\frac{1}{n}\sum_{j=1}^{n} f(\mathbf{X}^j)\right)^2 \\ &= \frac{1}{n}\sum_{j=1}^{n} f(\mathbf{X}^j) f(\mathbf{Z}^j) - \left(\frac{1}{n}\sum_{j=1}^{n} f(\mathbf{X}^j)\right)^2 \text{ p.s.,} \end{aligned}$$

et donc que l'estimateur corrigé

$$\ddot{\underline{\tau}}^2_{\mathfrak{u},n} = \frac{1}{n}\sum_{j=1}^{n} f(\mathbf{X}^j) f(\mathbf{X}^j_{\mathbf{u}} : \mathbf{Z}^j_{-\mathbf{u}}) - \frac{1}{n}\sum_{j=1}^{n} f(\mathbf{X}^j) f(\mathbf{Z}^j)$$

est "exact" pour les parts de variance $\underline{\tau}^2_{\mathfrak{u}}$ nulles, et potentiellement meilleur pour les $\underline{\tau}^2_{\mathfrak{u}}$ proches de zéro. Cet estimateur a été mentionné à de nombreuses reprises dans la littérature comme dans [Mau02], [Sal02], [SM07] et dernièrement [Owe12a]. En pratique, il s'avère particulièrement efficace pour estimer les indices de Sobol' proches de zéro.

Enfin, Owen a dernièrement introduit un nouvel estimateur dont le coût est 4 — i.e. il utilise quatre échantillons différents — mais dont la précision apparaît empiriquement supérieure à tous les estimateurs précédemment cités pour des indices de Sobol' proches de zéro. Il est défini par

$$\breve{\underline{\tau}}^2_{\mathfrak{u},n} = \frac{1}{n}\sum_{j=1}^{n} \left(f(\mathbf{X}^j) - f(\mathbf{Z}^j_{\mathbf{u}} : \mathbf{X}^j_{-\mathbf{u}})\right)\left(\left(f(\mathbf{X}^j_{\mathbf{u}} : \mathbf{Y}^j_{-\mathbf{u}}) - f(\mathbf{Y}^j)\right)\right).$$

2.1.4 En résumé

La méthode de Sobol', qui a été introduite en 1990 [Sob93], repose sur des bases mathématiques rigoureuses et a bénéficié jusqu'à très récemment de développements importants. Elle possède principalement deux caractéristiques. D'abord, elle s'applique à une classe considérable de fonctions — elle ne suppose en effet qu'une simple hypothèse d'intégrabilité — alors que la plupart des autres méthodes supposent une certaine régularité. Ceci a pour conséquence sa seconde caractéristique : la vitesse de convergence de ses estimateurs est seulement en $O(n^{-1/2})$. Avec l'introduction des intervalles de confiance asymptotiques [JKL+12], elle permet d'avoir une connaissance essentielle sur la précision de l'estimation des indices de Sobol' qu'elle fournit.

Cette méthode doit donc être considérée comme une méthode sûre — la validité de ses résultats ne dépend pas d'une quelconque hypothèse de régularité — présentant une vitesse de convergence modeste. Enfin, le principal frein à son application en pratique est le nombre important d'échantillons qu'elle requiert pour l'estimation des principaux indices de Sobol'. Dans cette optique, nous introduisons au Chapitre 6 de cette thèse une nouvelle approche d'optimisation combinatoire.

2.2 Méthode de McKay

La caractéristique principale de cette méthode [McK95] consiste en sa technique d'échantillonnage singulière par *hypercubes latins répliqués*. Toutefois la méthode de McKay ne se résume pas à cette particularité, et ses estimateurs peuvent se définir indépendamment de la notion d'hypercube latin et de réplication. Dans une première section, nous présentons donc les estimateurs originaux dus à McKay pour des échantillons aléatoires, faisant ainsi apparaître la méthode de McKay comme une simple méthode de Monte Carlo. Dans le paragraphe suivant, nous présentons la méthode de McKay proprement dite en ayant au préalable introduit les notions de réplication, d'hypercubes latins et d'hypercubes latins répliqués. Cela nous permet d'exposer la principale cararctéristique de cette méthode qui consiste à pouvoir estimer tous les indices de Sobol' descendants d'ordre 1 à un coût indépendant de la dimension.

Nous conservons le formalisme proposé en introduction de la Section 2.1. En outre, pour $r \geq 2$, on désigne par $\mathbf{Z}^k$, $k = [1:r]$, des vecteurs aléatoires indépendants identiquement distribués suivant la loi de $\mathbf{X}$. On note également pour tout $k \in [1:r]$,

$$Y_{\mathrm{u}}^k = f(\mathbf{X}_{\mathrm{u}} : \mathbf{Z}_{-\mathrm{u}}^k)$$

et

$$\overline{Y}_{\mathrm{u}} = \frac{1}{r}\sum_{k=1}^{r} Y_{\mathrm{u}}^k.$$

2.2.1 L'estimateur Monte Carlo de la méthode de McKay

Avec les notations précédentes, le théorème de la variance totale — voir par exemple Appendice A.4 p 74 dans [McK95] — s'énonce comme suit

$$\mathrm{Var}\big[\overline{Y}_{\mathrm{u}}\big] = \mathrm{Var}\Big[\mathbb{E}\big[\overline{Y}_{\mathrm{u}}\big|\mathbf{X}_{\mathrm{u}}\big]\Big] + \mathbb{E}\Big[\mathrm{Var}\big[\overline{Y}_{\mathrm{u}}\big|\mathbf{X}_{\mathrm{u}}\big]\Big].$$

Puis en notant que, conditionnellement à $\mathbf{X}_{\mathrm{u}}$, les Y_{u}^k sont indépendants et ont même loi, il vient aisément

$$\mathrm{Var}\big[\overline{Y}_{\mathrm{u}}\big] = \mathrm{Var}\Big[\mathbb{E}\big[Y\big|\mathbf{X}_{\mathrm{u}}\big]\Big] + \frac{1}{r}\mathbb{E}\Big[\mathrm{Var}\big[Y\big|\mathbf{X}_{\mathrm{u}}\big]\Big] \ .$$

Par suite, pour n'importe quel $r \geq 2$ on a

$$\underline{\tau}_{\mathrm{u}}^2 = \mathrm{Var}\big[\overline{Y}_{\mathrm{u}}\big] - \frac{1}{r}\mathbb{E}\Big[\mathrm{Var}\big[Y\big|\mathbf{X}_{\mathrm{u}}\big]\Big]. \tag{2.4}$$

Cette expression permet de définir naturellement un estimateur Monte Carlo de $\underline{\tau}_{\mathrm{u}}^2$, en opérant une double sommation pour la double intégrale qui figure dans le second terme du membre de droite de (2.4). En conservant les mêmes notations utilisées jusqu'à présent, on note

$$\begin{aligned} Y_{\mathrm{u}}^{j,k} &= f(\mathbf{X}_{\mathrm{u}}^j, \mathbf{Z}_{\mathrm{u}}^{j,k}) \\ \overline{Y}_{\mathrm{u}}^{j\cdot} &= \frac{1}{r}\sum_{k=1}^{r} Y_{\mathrm{u}}^{j,k} \\ \overline{Y}_{\mathrm{u}}^{\cdot\cdot} &= \frac{1}{rn}\sum_{k=1}^{r}\sum_{j=1}^{n} Y_{\mathrm{u}}^{j,k} \ , \end{aligned} \tag{2.5}$$

et on estime les deux termes de droite dans (2.4) par les quantités

$$\frac{1}{n}\sum_{j=1}^{n}(\overline{Y}_{\mathrm{u}}^{j\cdot}-\overline{Y}_{\mathrm{u}}^{\cdot\cdot})^2 \quad \text{et} \quad \frac{1}{r^2 n}\sum_{k=1}^{r}\sum_{j=1}^{n}(Y_{\mathrm{u}}^{j,k}-\overline{Y}_{\mathrm{u}}^{j\cdot})^2.$$

Par suite, l'estimateur de McKay pour $\underline{\tau}_{\mathrm{u}}^2$ s'écrit

$$\dot{\underline{\tau}}_{\mathrm{u},n,r}^{2,MK}=\frac{1}{n}\sum_{j=1}^{n}(\overline{Y}_{\mathrm{u}}^{j\cdot}-\overline{Y}_{\mathrm{u}}^{\cdot\cdot})^2-\frac{1}{r^2 n}\sum_{k=1}^{r}\sum_{j=1}^{n}(Y_{\mathrm{u}}^{j,k}-\overline{Y}_{\mathrm{u}}^{j\cdot})^2.$$

Remarque 2.1. *Cet estimateur nécessite d'effectuer une intégration numérique double suivant les indices $j \in [1:n]$ et $k \in [1:r]$. Il est donc nécessaire pour obtenir une convergence vers la valeur théorique ciblée de considérer une valeur de r conséquente. On s'en persuade d'autant plus en constatant que pour le cas limite $r = 2$, on a*

$$\dot{\underline{\tau}}_{\mathrm{u},n,2}^{2,MK}=\widehat{\underline{\tau}}_{\mathrm{u},n}^{2}+\frac{1}{8n}\sum_{j=1}^{n}\left(Y_{\mathrm{u}}^{j,1}+Y_{\mathrm{u}}^{j,2}\right)^2$$

où $\widehat{\underline{\tau}}_{\mathrm{u},n}^{2}$ est un estimateur de la méthode de Sobol' introduit par [MNM06] (*voir Formule* (2.4)).

Pour l'estimateur de σ^2, on a

$$\frac{1}{r}\sum_{k=1}^{r}\mathrm{Var}\left[Y_{\mathrm{u}}^{k}\right]=\mathrm{Var}[Y],$$

puis en notant

$$\overline{Y_{\cdot k}}=\frac{1}{n}\sum_{j=1}^{n}Y_{\mathrm{u}}^{j,k},$$

on définit naturellement l'estimateur de variance "par groupes"

$$\dot{\sigma}_{n,r,gr}^{2,MK}=\frac{1}{rn}\sum_{k=1}^{r}\sum_{j=1}^{n}(Y_{\mathrm{u}}^{j,k}-\overline{Y}_{\cdot k})^2.$$

Toutefois, McKay lui préfère l'estimateur de la moyenne empirique qui ne regroupe pas les termes suivant les r groupes

$$\dot{\sigma}_{n,r}^{2,MK}=\frac{1}{rn}\sum_{k=1}^{r}\sum_{j=1}^{n}(Y_{\mathrm{u}}^{j,k}-\overline{Y}_{\mathrm{u}}^{\cdot\cdot})^2.$$

On peut en effet considérer que ce dernier est asymptotiquement non biaisé du fait de la formule suivante

$$\dot{\sigma}_{n,r,gr}^{2,MK}-\dot{\sigma}_{n,r}^{2,MK}=\frac{1}{r}\sum_{k=1}^{r}(\overline{Y}_{\mathrm{u}}^{\cdot k}-\overline{Y}_{\mathrm{u}}^{\cdot\cdot})^2\underset{r\to\infty}{\longrightarrow}0.$$

Finalement McKay introduit l'estimateur suivant

$$\dot{\underline{S}}_{\mathrm{u},n,r}^{MK}=\frac{\frac{1}{n}\sum_{j=1}^{n}(\overline{Y}_{\mathrm{u}}^{j\cdot}-\overline{Y}_{\mathrm{u}}^{\cdot\cdot})^2-\frac{1}{r^2 n}\sum_{k=1}^{r}\sum_{j=1}^{n}(Y_{\mathrm{u}}^{j,k}-\overline{Y}_{\mathrm{u}}^{j\cdot})^2}{\frac{1}{rn}\sum_{k=1}^{r}\sum_{j=1}^{n}(Y_{\mathrm{u}}^{j,k}-\overline{Y}_{\mathrm{u}}^{\cdot\cdot})^2}. \tag{2.6}$$

En l'état, c'est un estimateur qui ne présente aucun intérêt, car sa dépendance vis-à-vis de r en fait un estimateur, a priori, encore plus coûteux que ceux de la méthode de Sobol'. Néanmoins, en mettant en œuvre un échantillonnage particulier, McKay introduit une méthode qui permet de calculer tous les indices de Sobol' $\underline{S}_{\{i\}}$, $i \in [1:d]$, pour un échantillon de taille totale égale à rn, indépendant de la dimension d.

2.2.2 La méthode de McKay

Nous restreignons dorénavant les discussions à des variables aléatoires $X_1,\dots, X_d$ uniformément distribuées sur l'hypercube unité. Lorsque ce n'est pas le cas, on considère qu'on peut se ramener numériquement à ce cadre par une transformation inverse. Nous commençons par introduire les notions d'hypercube latin, de réplication et d'hypercubes latins répliqués.

Définition 2.1. (hypercube latin) *Soient d et n dans $\mathbb{N}^*$, et considérons Π_n l'ensemble de toutes les permutations de $\{1,\dots,n\}$. On dit que $(\mathbf{X}^j)_{j\in[1:n]}$ est un hypercube latin de taille n dans $[0,1]^d$ — et nous notons $(\mathbf{X}^j)_j \sim \mathcal{LH}(n,d)$ — si pour tout $j \in 1:d$*

$$\mathbf{X}^j = \left(\frac{\pi_1(j) - U_{1,j}}{n}, \dots, \frac{\pi_d(j) - U_{d,j}}{n}\right)$$

où les π_i et les $U_{i,j}$ sont des variables aléatoires indépendantes respectivement uniformément distribuées sur Π_n et $[0,1]$.

Quant à la réplication elle consiste à obtenir un échantillon B à partir d'un échantillon A en permutant les valeurs des points de l'échantillon A, coordonnée par coordonnée. Pour les hypercubes latins, on obtient la définition suivante

Définition 2.2. (hypercubes latins répliqués) *Soient d et n dans $\mathbb{N}^*$, et considérons Π_n l'ensemble de toutes les permutations de $\{1,\dots,n\}$. On dit que $(\mathbf{X}^{j,1})_{j\in[1:n]},\dots,(\mathbf{X}^{j,r})_{j\in[1:n]}$ sont des hypercubes latins répliqués de taille n dans $[0,1]^d$ — et nous notons $(\mathbf{X}^{j,k})_{j,k} \sim \mathcal{RLH}(n,d,r)$ — si pour tout $j \in [1:n]$ et $k \in [1:r]$*

$$\mathbf{X}^{j,k} = \left(\frac{\pi_1^k(j) - U_{1,\pi_1^k(j)}}{n}, \dots, \frac{\pi_d^k(j) - U_{d,\pi_d^k(j)}}{n}\right)$$

où les π_i^k et les $U_{i,j}$ sont des variables aléatoires indépendantes respectivement uniformément distribuées sur Π_n et $[0,1]$.

Cette définition est étendue aux hypercubes latins basés sur des tableaux orthogonaux de force supérieure à 1 au Chapitre 6 de cette thèse. On considère maintenant des hypercubes latin répliqués $\mathbf{X}^{j,k}$, $j \in [1:n]$, $k \in [1:r]$. Et en notant que quelque soit $i \in [1:d]$ et quelque soit $j \in [1:n]$, on a

$$X_i^{(\pi_i^k)^{-1}(j),k} = X_i^{(\pi_i^k)^{-1}(j),1} = \frac{j - U_{i,j}}{n}$$

est constant pour tout $k \in [1:r]$, on peut déduire que pour tout $i \in [1:d]$,

$$y_{\{i\}}^{(\pi_i^k)^{-1}(j),k} = f\left(\mathbf{x}^{(\pi_i^k)^{-1}(j),k}\right) = f\left(\mathbf{x}_i^{(\pi_i^k)^{-1}(j),1} : \mathbf{x}_{-\{i\}}^{(\pi_i^k)^{-1}(j),k}\right).$$

La méthode de McKay consiste alors à appliquer l'estimateur introduit précédemment dans la Formule (2.6) pour $\mathfrak{u} = \{i\}$, en remplaçant les échantillons aléatoires $y_{\{i\}}^{j,k}$ définis dans la Formule (2.5) par les échantillons stratifiés $y_{\{i\}}^{(\pi_i^k)^{-1}(j),k}$ définis ci-dessus. Il n'y a donc pas besoin de réévaluer la fonction f sur de nouveaux échantillons pour chaque i, on se contente de réutiliser les mêmes mais en ordonnant les points de façon différente.

2.2.3 En résumé

Finalement, même si la dépendance vis-à-vis de r donne un aspect négatif à cette méthode et semble l'avoir fait tomber dans l'oubli, le fait de pouvoir estimer tous les indices de Sobol' d'ordre 1 à un coût indépendant de la dimension d est remarquable. C'est d'ailleurs l'idée qui sous-tend la méthode que nous proposons au Chapitre 6 de cette thèse. Dans son rapport, McKay généralise sa méthode à l'estimation de tout indice de Sobol' descendant en utilisant toujours des hypercubes latins répliqués. Mais dès lors que $|\mathfrak{u}| \geq 2$, il ne conserve pas la propriété d'indépendance vis-à-vis de la dimension d. Il est toutefois possible de la récupérer en introduisant des hypercubes latins basés sur des tableaux orthogonaux de force supérieure à 1. Cette généralisation est d'un intérêt très relatif pour la méthode de McKay qui ne s'applique pas vraiment en pratique à notre connaissance, mais elle est importante dans la perspective de notre travail et elle est discutée au Chapitre 6.

2.3 Méthodes FAST, RBD et RBD-FAST

La présentation qui suit ne prend pas en compte nos contributions pour les méthodes citées dans le titre de cette section. Nos travaux figurent aux Chapitres 4 et 5 de cette thèse sous la forme d'un article et d'une prépublication respectivement intitulés *Bias correction for the estimation of sensitivity indices based on random balance designs* — à paraître dans *Reliability Enginnering and System Safety* — et *Variance-based sensitivity analysis using harmonic analysis.*

La méthode FAST a été introduite par quatre articles successifs de Cukier et al. [CFS+73, CSS75, CLS78] et de Schaibly et Shuler [SS73] au cours des années 1970, initialement afin d'étudier la cinétique des concentrations dans des réactions chimiques ; puis elle a été enrichie par Saltelli et al. [STC99]. L'approche singulière des travaux initiaux consiste principalement à quantifier les variations d'une fonction multivariée en analysant sa structure harmonique. Même si le développement de cette méthode est antérieur aux formalisations de la décomposition ANOVA dues à Efron & Stein [ES81] et Sobol' [Sob93], les mesures d'importances que FAST estime sont rigoureusement celles définies dans le Chapitre 1 — voir Définition 1.9 — comme les indices de Sobol' élémentaires [2]. L'idée principale de la méthode est de réduire le coût de calcul des intégrales multiples qui constituent les indices de Sobol', par des intégrales simples. Cette simplification repose sur un théorème ergodique de Weyl [Wey38] et sur l'utilisation des séries de Fourier. Ces considérations théoriques font l'objet de la première section. Nous consacrons ensuite une section à la présentation des méthodes RBD et RBD-FAST qui, toutes deux, introduites par Tarantola et al. [TGM06], exploitent le cadre théorique sous-jacent de la méthode FAST.

Dans ce qui suit, nous restreignons le modèle $Y = f(X_1, \ldots, X_d)$ à des variables aléatoires X_1, $\ldots X_d$ uniformes. Nous discutons cette hypothèse par la suite dans la sous-section 2.3.1. Nous rappelons par ailleurs que dans la notation des intégrales, l'ensemble d'intégration, s'il n'est pas spécifié, est l'hypercube unité dont la dimension est celle de l'intégrande.

2.3.1 Méthode FAST

Pour $\mathbf{k} = (k_1, \ldots, k_d) \in \mathbb{Z}^d$, nous notons

$$\mathrm{c}_{\mathbf{k}}(f) = \mathbb{E}\big[f(\mathbf{X}) \exp(-2i\pi \mathbf{k} \cdot \mathbf{X})\big] = \int f(\mathbf{x}) \exp(-2i\pi \mathbf{k} \cdot \mathbf{x}) \mathrm{d}\mathbf{x} \tag{2.7}$$

le $\mathbf{k}$-ème coefficient de Fourier multidimensionnel de f.

Approximation du théorème ergodique de Weyl et définition des estimateurs

Le point de départ de la méthode FAST est une réduction de la dimension de l'intégrale multiple figurant dans la Formule (2.7) ; elle est obtenue par une approximation du théorème ergodique qui suit.

Théorème 2.5. [Weyl, 1938] *Soit g une fonction définie sur l'hypercube unité $[0,1]^d$, bornée et intégrable au sens de Riemann. Pour tout $i \in [1:d]$, considérons les fonctions $x_i(t) = \{\omega_i t\}$, $t \in \mathbb{R}$, où les ω_i sont des réels linéairement indépendants sur $\mathbb{Q}$ et où $\{\cdot\}$ désigne la partie fractionnaire. Alors*

$$\int g(\mathbf{x}) d\mathbf{x} = \lim_{T \to +\infty} \frac{1}{2T} \int_{-T}^{T} g\big(x_1(t), \ldots, x_d(t)\big).$$

En particulier, pour tout $\mathbf{k} \in \mathbb{Z}^d$ et $g : \mathbf{x} \mapsto f(\mathbf{x}) \exp(-2i\pi \mathbf{k} \cdot \mathbf{x})$, ce résultat donne

$$\mathrm{c}_{\mathbf{k}}(f) = \lim_{T \to +\infty} \frac{1}{2T} \int_{-T}^{T} f \circ \mathbf{x}(t) \exp\big(-2i\pi(\mathbf{k} \cdot \boldsymbol{\omega})t\big) dt.$$

2. Saltelli et Bolado [SB98] ont montré empiriquement que les effets principaux estimés dans la méthode FAST sont comparables aux indices de Sobol' élémentaires d'ordre 1. D'autre part, nous explicitons cette équivalence d'un point de vue théorique dans le Chapitre 5 de cette thèse en la généralisant aux indices de Sobol' élémentaires de tous ordres, lesquels étaient déjà considérés par Cukier et al. [CLS78], Section 5.C

Par suite, afin de contourner le problème de l'intégration sur un ensemble infini, la courbe paramétrée $\mathbf{x}(t) = \{\boldsymbol{\omega} t\}$ est remplacée par $\mathbf{x}^*(t)$ définie par

$$x_i^*(t) = G_i\big(\sin(2\pi\omega_i t)\big), \quad i \in [1:d]$$

où les ω_i sont cette fois-ci des entiers positifs, et les G_i sont des fonctions définies sur $[-1,1]$ à image dans $[0,1]$. Finalement, comme les courbes paramétrées x_i^* sont 1-périodiques, on obtient l'approximation des coefficients de Fourier suivante

$$\mathrm{c}_{\mathbf{k}}(f) \approx \int_0^1 f \circ \mathbf{x}^*(t) \exp\big(-2i\pi(\mathbf{k}\cdot\boldsymbol{\omega})t\big)dt$$

et en discrétisant l'intégrale sur une grille régulière, par une méthode des rectangles, on obtient l'estimateur

$$\dot{\mathrm{c}}_{\mathbf{k},n}(f) = \frac{1}{n}\sum_{j=0}^{n-1} f \circ \mathbf{x}^*\Big(\frac{j}{n}\Big) \exp\Big(-2i\pi(\mathbf{k}\cdot\boldsymbol{\omega})\frac{j}{n}\Big).$$

On remarque alors que $\dot{\mathrm{c}}_{\mathbf{k},n}(f)$ n'est autre que le $\mathbf{k}\cdot\boldsymbol{\omega}$-ème coeffcient de Fourier du signal unidimensionnel $\big(f \circ \mathbf{x}^*(j/n)\big)_{j\in[0:n-1]}$. Pour la suite, l'estimateur $\dot{\mathrm{c}}_{\mathbf{k},n}(f)$ est donc noté $\dot{\mathrm{c}}_{\mathbf{k}\cdot\boldsymbol{\omega},n}(f \circ \mathbf{x}^*)$. Les estimateurs de la variance totale σ^2 et des variances partielles d'ordre 1 $\sigma^2_{\{i\}}$, $i \in [1:d]$ introduits dans la méthode FAST sont alors

$$\dot{\sigma}^{2,FAST}_{\{i\},n} = 2\sum_{k=1}^{N_i} |\dot{\mathrm{c}}_{k\omega_i,n}(f \circ \mathbf{x}^*)|^2, \tag{2.8}$$

$$\dot{\sigma}^{2,FAST}_{n} = \sum_{k=1}^{n-1} |\dot{\mathrm{c}}_{k,n}(f \circ \mathbf{x}^*)|^2 \tag{2.9}$$

et

$$\dot{S}^{FAST}_{\{i\},n} = \frac{2\sum_{k=1}^{N_i} |\dot{\mathrm{c}}_{k\omega_i,n}(f \circ \mathbf{x}^*)|^2}{\sum_{k=1}^{n-1} |\dot{\mathrm{c}}_{k,n}(f \circ \mathbf{x}^*)|^2}$$

où les N_i sont des ordres de troncatures historiquement fixés à 4 ou 6 (voir par exemple [SS73, STC99, TGM06]), mais l'expérience montre que des valeurs supérieures peuvent être nécessaires. On note également que par la formule de Parseval, l'estimateur de σ^2 de (2.9) est exactement l'estimateur de variance empirique des $f \circ \mathbf{x}^*(j/n)$, $j \in [0:n-1]$. Enfin, comme remarqué dans [CLS78] (voir Appendice 5.C), on peut définir les estimateurs des parts de variances partielles de tout ordre $\sigma^2_{\mathfrak{u}}$, $\mathfrak{u} \subseteq [1:d]$, de la même manière que dans (2.8) en considérant les combinaisons linéaires de fréquences ω_i (voir Chapitre 5 de cette thèse, Section 5.3.1).

Problèmes liés au traitement du signal

La validité des estimateurs définis dans (2.8) est vérifiée par l'expérience dès les travaux initiaux en 1973 (voir [SS73]). Néanmoins, elle peut être remise en cause si certaines conditions, liées à la théorie du traitement du signal, ne sont pas respectées. Nous en développons les deux principaux aspects, la taille de l'échantillon et le choix des fréquences ω_i.

Tout d'abord, le choix du nombre de simulations n est régi par le théorème de Nyquist-Shannon qui s'énonce comme suit

« La fréquence d'échantillonnage d'un signal doit être égale ou supérieure au double de la fréquence maximale contenue dans ce signal. »

Ce critère implique que la taille de l'échantillon est bornée inférieurement,

$$n \geq 2N_h\omega_{\max} \ .$$

Dans le cas contraire, on s'expose au phénomène de *repliement de spectre* — ou *aliasing* — i.e. que les composantes des fréquences supérieures à la fréquence critique $\lfloor n \rfloor/2$ ne sont pas correctement

identifiées et sont attribuées à tort à des fréquences inférieures. Par conséquent, l'aliasing mène principalement à une surestimation des indices de Sobol'.

D'autre part, le fait d'utiliser des fréquences entières conduit à des dépendances linéaires entre ces dernières, i.e. qu'il existe des p-uplets $(a_1, \ldots, a_p)$ d'entiers relatifs tels que

$$\sum_{i=1}^{p} a_i \omega_i = 0 \ .$$

Cela signifie que des interférences peuvent apparaître et fausser l'estimation des indices de Sobol', i.e. des combinaisons linéaires de certaines fréquences peuvent être identifiées comme une harmonique d'une autre fréquence. Pour se prémunir de ce phénomène, Schaibly et Shuler [SS73] proposent d'imposer que les fréquences soient linéairement indépendantes "jusqu'à l'ordre M", i.e.

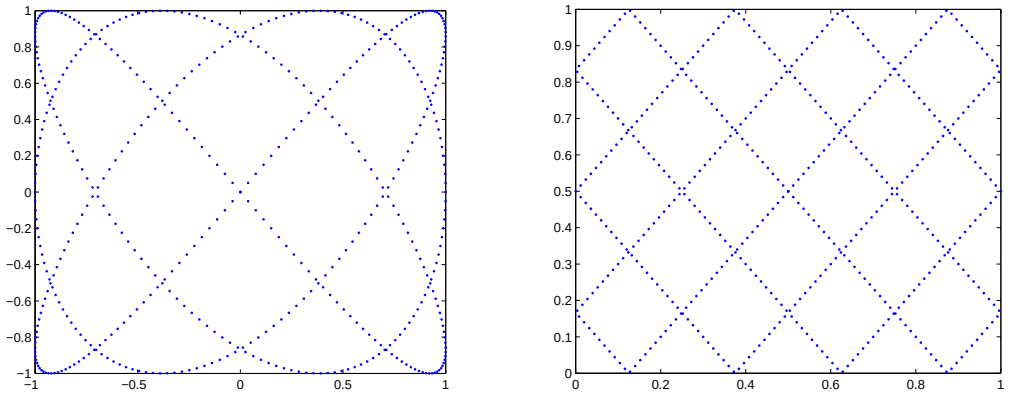
$$\sum_{i=1}^{p} a_i \omega_i \neq 0 \quad \text{pour} \quad \sum_{i=1}^{p} |a_i| \leq M + 1 \ .$$

Ils font aussi l'hypothèse que pour des fréquences suffisamment grandes, les interférences sont négligeables. La constante M doit être fixée par l'utilisateur ; il est évident que plus M est grand et plus $\omega_{\max}$ sera grand. En 1973, Schaibly et Shuler la fixe à 4 et Saltelli et al. confirment ce choix en 1999 en imposant $M = N_h$, avec $N_h = 4$ ou 6.

Développements de Saltelli et al. (1999)

L'ensemble des contributions effectuées par Saltelli et al. [STC99] sont quelquefois référencées sous l'abréviation EFAST, pour Extended FAST. Nous les résumons dans cette section.

L'échantillonnage suivant la courbe paramétrée $\mathbf{x}^*$ engendre de nombreuses contraintes car les répartitions marginales des points de l'échantillon $\mathbf{x}^*(j/n)$, $j \in [0 : n-1]$ sont liées entre elles. Néanmoins une question importante dans la méthode FAST a été de pouvoir générer des échantillons respectant les distributions marginales des variables d'entrée. Une fois les fréquences ω_i fixées, les seuls degrés de liberté disponibles sont les fonctions G_i. Dans le cas où les fonctions G_i sont égales à l'identité (voir figure 2.1 (a)), on peut constater que la répartition marginale des points semble très éloignée d'une distribution uniforme. La solution de ce problème est donnée par Cukier et al.



(a) $G_1 = G_2 = \mathrm{Id}$ (b) $G_1 = G_2 = \frac{1}{2} + \frac{1}{\pi}\arcsin$

FIGURE 2.1 – Répartition de 500 points pour les fréquences $\omega_1 = 3$ et $\omega_2 = 4$.

[CLS78]. Ils montrent que, si l'on note f_{X_i} les densités des variables d'entrées X_i, l'échantillon généré dans la méthode FAST respecte[3] ces distributions à condition que chacune des fonctions $G_i(y)$ vérifie

3. i.e. que la longueur totale de l'arc contenu dans la bande délimitée par u_i et $u_i + du_i$ est égale à $f_{X_i}(u_i)du_i$. L'échantillonnage respecte, de ce fait, les lois marginales, mais les points restent liés entre eux par la courbe d'échantillonnage.

l'équation différentielle suivante

$$\pi(1-y^2)^{1/2} f_{X_i}(G_i(y)) \frac{dG_i(y)}{dy} = 1 \ , \ y \in [-1,1]$$

avec la condition au bord $G_i(0) = 0$. Dans le cas de paramètres d'entrée uniformément distribués, Saltelli et al. [STC99] montrent que la solution est

$$G_i = \frac{1}{2} + \frac{1}{\pi} \arcsin \quad \text{(voir figure 2.1 (b))}$$

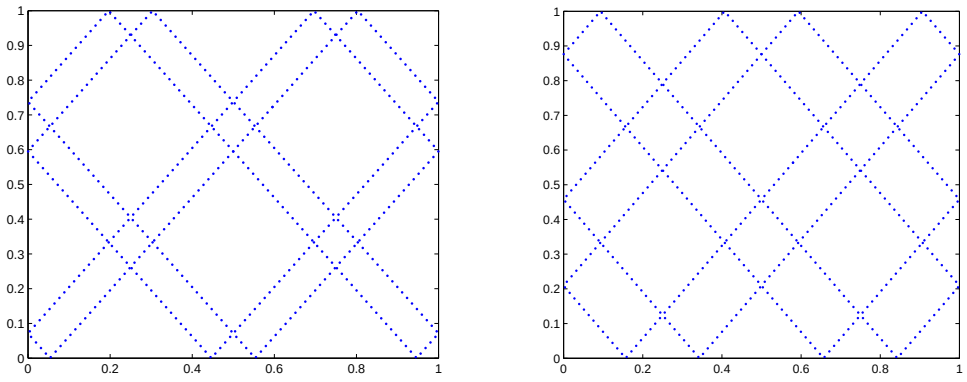
i.e.

$$x_i(t) = \frac{1}{2} + \frac{1}{\pi} \arcsin(\sin(2\pi\omega_i t)) \ .$$

En outre, Saltelli et al. [STC99] éliminent un défaut majeur de la méthode FAST qui est que, une fois les fréquences ω_i et les fonctions G_i fixées, l'échantillonnage ne peut se faire que suivant la courbe $\mathbf{x}^*$. Ce qui signifie que la majorité de l'espace d'état des variables ne peut être échantillonné. Plus précisément, ils constatent que le fait de déphaser les fonctions sinus comme suit

$$x_i(t) = G_i(\sin(2\pi\omega_i t + \varphi_i)) \quad \text{pour } i = 1, \ldots, p \ ,$$

ne modifie aucunement la théorie développée dans FAST. Au contraire, ce déphasage permet de considérer — à fréquences ω_i et fonctions G_i constantes — de nouvelles courbes d'échantillonnage (voir figure 2.2). La nouvelle approche proposée consiste alors à estimer les indices de Sobol' en



(a) $\varphi_1 = 0.3, \ \varphi_2 = 0.7$ (b) $\varphi_1 = 0.25, \ \varphi_2 = 0.2$

FIGURE 2.2 – Répartition de 500 points pour les fréquences $\omega_1 = 3$ et $\omega_2 = 4$, $G_1 = G_2 = \frac{1}{2} + \frac{1}{\pi}\arcsin$.

faisant la moyenne des indices calculés sur un nombre N_r de courbes d'échantillonnages aléatoirement déphasées ; la taille de l'échantillon devient donc

$$n = (2N_h\omega_{\max} + 1)N_r \ .$$

FAST devient ainsi une méthode entièrement probabiliste dont on peut juger la précision en calculant la variance empirique de l'estimateur des indices de Sobol'. Les détails techniques se réduisent alors à la question : "comment fixer la valeur N_r ?". Saltelli et al. émettent des suggestions concernant ce choix en fonction de la fréquence maximale $\omega_{\max}$; pour plus de détails nous renvoyons à leur article [STC99].

Enfin, Saltelli et al. [STC99] introduisent un nouvel estimateur pour les indices ascendants d'ordre 1. L'idée est de considérer l'indice total S_{T_i}, non pas par sa définition directe — somme des indices de tous ordres contenant la i-ème variable — mais par sa définition alternative,

$$\overline{S}_{\{i\}} = 1 - \frac{\sigma^2_{\{i\}^c}}{\sigma^2}.$$

Et cette part de variance $\sigma^2_{\{i\}^c}$ est identifiée par les fréquences qui sont combinaisons linéaires de toutes les fréquences initiales ω_k avec $k \neq i$. Ainsi l'estimateur proposé est

$$\dot{\sigma}^{2,FAST}_{\{i\}^c,n} = \sum_{k \in K_i} |\dot{c}_{k,n}(f \circ \mathbf{x}^*)|^2$$

où K_i désigne l'ensemble des fréquences, combinaisons linaires jusqu'à un ordre N_i des fréquences initiales ω_k avec $k \neq i$,

$$K_i \triangleq \left\{ \sum_{k \neq i} a_k \omega_k \ \text{ tels que } \ \sum_{k \neq i} |a_k| \leq N_i \right\}.$$

L'efficacité de cet estimateur est toutefois soumise au fait que les fréquences, combinaisons linéaires jusqu'à l'ordre N_i des fréquences initiales contenant ω_i, doivent se situer à l'extérieur de l'ensemble K_i sous peine de créer des interférences et d'engendrer une surestimation probable des indices de Sobol' ascendants. Pour éviter ce problème, Saltelli et al. proposent de choisir la fréquence d'intérêt ω_i assez élevée et de cantonner les autres à des valeurs réduites (voir [STC99] pour quelques exemples de jeux de fréquences). Ceci se traduit par deux conséquences principales :

1) les petites valeurs des fréquences autres que la fréquence d'intérêt ω_i mènent à un échantillonnage médiocre.
2) le plan d'expérience — régi par le jeu de fréquences initiales et donc par la position de la "grande" fréquence — est spécifique à chaque variable d'intérêt. Il faut donc d échantillons pour estimer les d indices de Sobol' ascendants d'ordre 1.

Les solutions à la conséquence 1) sont, soit d'augmenter le nombre de rééchantillonnages N_r — introduit dans la nouvelle approche — et ainsi de mieux explorer l'espace d'état des variables, soit d'accroître la fréquence d'intérêt ω_i afin de pouvoir choisir les autres fréquences plus librement. Malheureusement, dans les deux cas, cela mène à une augmentation de la taille de l'échantillon.

2.3.2 Méthodes RBD et RBD-FAST

Les méthodes RBD et RBD-FAST ont été introduites par Tarantola et al. [TGM06] dans le but de contourner les difficultés inhérentes au choix des fréquences ω_i, $i \in [1:d]$, dans la méthode FAST. Initialement, ces deux méthodes restent cantonnées à l'estimation des indices de Sobol' élémentaires d'ordre 1. Néanmoins, par sa définition la méthode hybride RBD-FAST permet naturellement d'estimer les indices de Sobol' élémentaires de tous ordres (voir [Mar09, TP12a]), et elle peut également être appliquée à l'estimation des indices de Sobol' ascendants (voir [Mar09]). En outre, nous présentons, au Chapitre 5 de cette thèse une généralisation de la méthode RBD aux indices de Sobol' descendants de tous ordres.

La construction de la méthode RBD se fait à partir du cadre de FAST avec $\varphi_1 = \cdots = \varphi_d = 0$, $\omega_1 = \cdots = \omega_d = \omega \in \mathbb{N}^*$ — généralement fixé à 1 — et en appliquant des permutations aléatoires aux coordonnées des points $\mathbf{x}^*(\frac{j}{n})$. Plus précisément, soient $\pi_1, \ldots, \pi_d$ des permutations aléatoires de $\{0, \ldots, n-1\}$ et Π_n l'ensemble des $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_d)$. Soit $\boldsymbol{\pi} \in \Pi_n$, considérons la fonction $\mathbf{x}^\times = (x_1^\times, \ldots, x_d^\times)$ définie sur $\{0, \frac{1}{n}, \ldots, \frac{n-1}{n}\}$ telle que pour tout $i \in [1:d]$ et $j \in 0 : n-1$,

$$x_i^\times\Big(\frac{j}{n}\Big) = \frac{1}{\pi} \arcsin\Big(\sin\Big(2\pi\omega \frac{\pi_i(j)}{n}\Big)\Big) + \frac{1}{2}.$$

On définit alors

$$\mathbf{x}^{\times,i}\Big(\frac{j}{n}\Big) = \mathbf{x}^\times\Big(\frac{\pi_i^{-1}(j)}{n}\Big)$$

où π_i^{-1} désigne la permutation inverse de π_i (voir Figure 2.3). Et par suite, Tarantola et al. [TGM06]

introduisent les estimateurs $\dot{\sigma}^{2,RBD}_{\{i\},n}$, $\dot{\sigma}^{2,RBD}_{n}$ et $\dot{S}^{RBD}_{\{i\},n}$

$$\begin{aligned}
\dot{\sigma}^{2,RBD}_{\{i\},n} &= 2\sum_{k=1}^{N_i} \dot{c}_{k\omega,n}(f \circ \mathbf{x}^{\times,i}), \\
\dot{\sigma}^{2,RBD}_{n} &= \sum_{k=1}^{n-1} \dot{c}_{k,n}(f \circ \mathbf{x}^{\times}), \\
\dot{S}^{RBD}_{\{i\},n} &= \frac{2\sum_{k=1}^{N_i} \dot{c}_{k\omega,n}(f \circ \mathbf{x}^{\times,i})}{\sum_{k=1}^{n-1} \dot{c}_{k,n}(f \circ \mathbf{x}^{\times})}
\end{aligned}$$

Comme dans la méthode FAST, on note également que par la formule de Parseval, l'estimateur de σ^2 est exactement l'estimateur de variance empirique des $f \circ \mathbf{x}^{\times}(j/n)$, $j \in [0 : n-1]$.

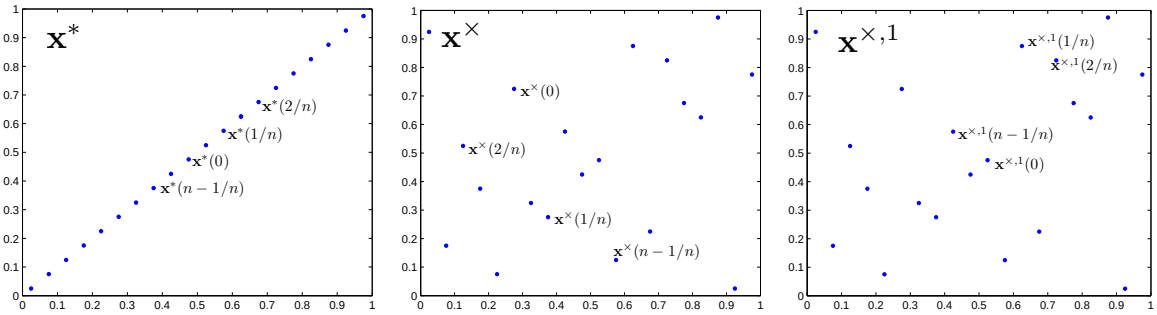


FIGURE 2.3 – Illustration des plans utilisés dans la méthode RBD.

La méthode hybride RBD-FAST consiste à combiner les deux façons d'échantillonner. On commence par scinder les variables d'entrées en plusieurs sous-groupes de tailles réduites à l'intérieur desquels on échantillonne par la méthode FAST — i.e. on attribue des fréquences libres d'interférences à l'intérieur de chaque sous-groupe — puis on applique une permutation à chacun des sous-groupes (voir Table 2.2). Cela permet selon les auteurs [TGM06], de conserver la précision de la méthode FAST et la faiblesse du coût de la méthode RBD.

6 paramètres :	$\underbrace{\omega_1 \ \omega_2 \ \omega_3}_{\pi_1}$	$\underbrace{\omega_1 \ \omega_2 \ \omega_3}_{\pi_2}$	
6 paramètres :	$\underbrace{\omega_1 \ \omega_2}_{\pi_1}$	$\underbrace{\omega_1 \ \omega_2}_{\pi_2}$	$\underbrace{\omega_1 \ \omega_2}_{\pi_3}$
7 paramètres :	$\underbrace{\omega_1 \ \omega_2 \ \omega_3}_{\pi_1}$	$\underbrace{\omega_1 \ \omega_2 \ \omega_3}_{\pi_2}$	$\underbrace{\omega_4}_{\pi_3}$
7 paramètres :	$\underbrace{\omega_1 \ \omega_2 \ \omega_3}_{\pi_1}$	$\underbrace{\omega_4 \ \omega_5}_{\pi_2}$	$\underbrace{\omega_4 \ \omega_5}_{\pi_3}$

TABLE 2.2 – Exemple d'attribution des permutations et des fréquences pour chaque paramètre dans la méthode RBD-FAST.

Cette méthode n'est pas investiguée de manière plus rigoureuse par la suite. On note cependant qu'au regard de notre travail proposé au Chapitre 5 de cette thèse, la méthode RBD-FAST est étroitement liée à la notion de *supercube latin* [Owe98] construit sur des sous-groupes cycliques du tore unité.

2.3.3 En résumé

Les méthodes FAST, RBD et RBD-FAST se définissent principalement comme des méthodes spectrales relatives à la base orthonormale constituée des exponentielles complexes. Même si leurs estimateurs des indices de Sobol' ont pour origine un énoncé mathématique rigoureux — le théorème

ergodique de Weyl — elles apparaîssent néanmoins comme des méthodes approximatives dont l'analyse de l'erreur d'estimation est difficile à appréhender rigoureusement, et le champs d'hypothèses nécessaires à leurs applications reste flou. Cette problématique fait l'objet d'un travail théorique que nous présentons au Chapitre 5.

2.4 Méthodes relatives à une décomposition ANOVA spectrale

Dans le cadre d'une décomposition ANOVA spectrale relative à une base orthonormale $\Phi_{\mathbf{k}} = \phi_{1k_1} \otimes \cdots \otimes \phi_{dk_d}$, $\mathbf{k} \in \mathbb{N}^d$ comme décrite dans la Section 1.2.4, nous rappelons qu'estimer les indices de Sobol' revient à calculer les $|\mathrm{c}_{\mathbf{k}}|^2$ non-négligeables dans la décomposition de Y

$$Y = \sum_{\mathfrak{u} \subseteq [1:d]} \sum_{\mathbf{k} \in K_{\mathfrak{u}}} \mathrm{c}_{\mathbf{k}} \Phi_{\mathbf{k}}(\mathbf{X}_{\mathfrak{u}}).$$

Cette estimation des coefficients spectraux peut se faire principalement par régression linéaire, interpolation polynomiale et quasi-régression. Nous les résumons dans les sections qui suivent.

2.4.1 Régression linéaire

Soient $\mathbf{X}^j$, $j \in [1:n]$ une famille de vecteurs aléatoires indépendants distribués suivant la loi de $\mathbf{X}$, et soit $K \subseteq \mathbb{N}^d$ un ensemble de troncature. Notons $Y^j = f(\mathbf{X}^j)$, et considérons le modèle linéaire

$$Y^j = \sum_{\mathbf{k} \in K} \beta_{\mathbf{k}} \mathbf{\Phi}_{\mathbf{k}}(\mathbf{X}^j) + \varepsilon^j, \quad j \in [1:n],$$

où les paramètres $\beta_{\mathbf{k}}$ et les termes d'erreur ε^j sont des nombres réels. Alors l'estimateur des moindres carrés ordinaires (MCO) issu de la régression linéaire suivant les fonctions des bases $\mathbf{\Phi}_{\mathbf{k}}$, $\mathbf{k} \in K$ est le vecteur $\boldsymbol{\beta}$ qui minimise la norme énergie de l'erreur $\boldsymbol{\varepsilon}$:

$$\dot{\boldsymbol{\beta}} = \mathrm{argmin}_{\boldsymbol{\beta}} \frac{1}{n} \sum_{j=1}^{n} (\varepsilon^j)^2.$$

On pose alors $\mathcal{M}(\mathbf{\Phi})$ la matrice à n lignes et $p = |K|$ colonnes de terme général $\mathbf{\Phi}_{\mathbf{k}}(\mathbf{X}^j)$. Si $\mathcal{M}(\mathbf{\Phi})$ est régulière alors l'estimateur MCO est donné par

$$\dot{\boldsymbol{\beta}} = \left(\mathcal{M}(\mathbf{\Phi})^{\intercal} \mathcal{M}(\mathbf{\Phi})\right)^{-1} \mathcal{M}(\mathbf{\Phi}) \mathbf{Y} \tag{2.10}$$

où $\mathbf{Y} = (Y^1, \ldots, Y^n)^{\intercal}$. Par suite, pour $\mathfrak{u} \subseteq [1:d]$, on définit les estimateurs suivants :

$$\dot{\sigma}^{2,REG}_{\mathfrak{u}} = \sum_{\mathbf{k} \in K \cap \mathbb{N}_{\mathfrak{u}}} |\dot{\beta}_{\mathbf{k}}|^2$$

où $\mathbb{N}_{\mathfrak{u}} = \{(k_1, \ldots, k_d) \mid \forall i \in \mathfrak{u},\ k_i \in \mathbb{N}^* \text{ et } \forall i \in \mathfrak{u}^c,\ k_i = 0\}$, et on aboutit à des estimateurs pour les indices de Sobol' en considérant en outre la variance empirique des Y^j, $j \in [1:n]$ pour estimer σ^2.

Remarque 2.2. *D'un point de vue théorique, l'approche par régression consiste à donner une réponse par optimisation dans l'espace fonctionnel — par minimisation de l'erreur quadratique — à une question d'optimisation posée dans l'espace dual, au sens où on veut minimiser l'erreur d'estimation des indices de Sobol' i.e. l'erreur d'estimation des coefficients spectraux. Par suite, si le dictionnaire de fonctions $\mathbf{\Phi}_{\mathbf{k}}$, $\mathbf{k} \in K$ est insuffisant pour expliquer le modèle f, alors l'optimisation par les moindres carrés ordinaires mène à une erreur minimale dans l'espace fonctionnel puisqu'on minimise l'erreur quadratique, mais à une erreur qu'on ne contrôle malheureusement pas dans l'espace dual.*

Au regard de la remarque précédente, la question du choix de l'ensemble de troncature K apparaît donc comme un problème crucial, et ne semble pouvoir être traîtée que dans un cadre adaptatif par un enrichissement du dictionnaire de fonctions $\mathbf{\Phi}_{\mathbf{k}}$ et également une augmentation de la taille de l'échantillon pour assurer l'inversibilité de $\mathcal{M}(\mathbf{\Phi})$ et un conditionnement correct (voir e.g. [Bla09,

BS10]) Néanmoins, le coût de ce type d'approches adaptatives peut rapidement devenir prohibitif si la méthode ne converge pas suffisamment vite, car la complexité algorithmique d'une régression est en $O(np^2)$ en temps et $O(p^2)$ en espace, ce qui est loin d'être négligeable si la taille de l'échantillon n et le nombre de fonctions de bases p augmentent conjointement. Nous terminons en prolongeant notre remarque précédente.

Remarque 2.3. *En présence de composantes spectrales $\mathbf{\Phi_k}$ de grande complexité — i.e. $|\{k_i,\ k_i \neq 0\}|$ et/ou $\max_{i\in[1:d]} |k_i|$ grands — on peut raisonnablement concevoir que ces méthodes adaptatives n'arrivent jamais à enrichir suffisamment le dictionnaire de fonctions de bases en un temps raisonnable et aboutissent à une optimisation "biaisée" dans l'espace dual comme expliqué dans la remarque précédente. Par conséquent, le calcul des indices de Sobol' par régression linéaire suivant une décomposition spectrale ne peut être envisageable que sous des hypothèses restrictives concernant la structure spectrale de la fonction étudiée i.e. des connaissances a priori sur la décomposition ANOVA de cette dernière. Ceci constitue un paradoxe au sens où cela revient à expliquer en partie de manière a priori l'objet étudié, i.e. la décomposition ANOVA du modèle.*

2.4.2 Interpolation polynomiale

Nous résumons tout d'abord en quoi consiste l'interpolation polynomiale de fonctions de plusieurs variables, puis nous détaillons les aspects pratiques en lien avec le calcul des indices de Sobol' dans un dernier paragraphe.

Interpolation lagrangienne en une variable Soient $x^0, \dots, x^n \in \mathbb{R}$, $n+1$ points distincts, et leurs $n+1$ valeurs associées $y^j = f(x^j)$. Alors, il existe un unique polynôme P_n de degré n tel que pour tout $j \in [0:n]$, $P_n(x^j) = f(x^j)$. Il est défini par

$$P_n(x) = \sum_{j=0}^{n} f(x^j) l_j(x)$$

où les l_j sont les polynômes de Lagrange

$$l_j(x) = \prod_{k\neq j,\ k=0}^{n} \frac{x - x^k}{x^j - x^k}.$$

Interpolation lagrangienne multivariée par produit tensoriel Considérons maintenant une fonction f à d variables et $\mathcal{X}^1_{n_1}, \dots, \mathcal{X}^d_{n_d}$ une famille d'ensembles de points distincts $x_i^0, \dots, x_i^{n_i}$, $i \in [1:d]$. Notons $\mathcal{P}_n$ l'ensemble des polynômes d'une variable de degré au plus n, et $l^{(i)}_{k_i}$ les polynômes de Lagrange relatifs à la i-ème variable. Alors il existe un unique polynôme $P_{\mathbf{n}}(\mathbf{x})$ dans $\mathcal{P}_{n_1} \otimes \cdots \otimes \mathcal{P}_{n_d}$ tel que pour tout $\mathbf{x}' \in \mathcal{X}^1_{n_1} \otimes \cdots \otimes \mathcal{X}^d_{n_d}$, $P_{\mathbf{n}}(\mathbf{x}') = f(\mathbf{x}')$. Il est défini par

$$P_{\mathbf{n}}(\mathbf{x}) = \sum_{k_1=0}^{n_1} \cdots \sum_{k_d=0}^{n_d} f(x_1^{k_1}, \dots, x_d^{k_d}) l^{(1)}_{k_1} \otimes \cdots \otimes l^{(d)}_{k_d}(\mathbf{x}). \tag{2.11}$$

Cette approche multidimensionnelle peut se révéler extrêmement coûteuse en grande dimension, car le nombre de points d'interpolation est égal au produit $n = n_1 \times \cdots \times n_d$, et devient rapidement irréaliste en grande dimension. Rappelons par exemple qu'en dimension 20, si il y a 4 abscisses d'interpolation par variable — i.e. $n_1 = \cdots = n_d = 4$ — on obtient $4^{20} > 10^{12}$ points d'interpolations et que 4 abscisses par variable peut être largement insuffisant, même pour des fonctions de faible complexité. Pour cette raison, en pratique on s'oriente vers l'algorithme de tensorisation "partielle" de Smolyak [Smo63].

Algorithme de Smolyak La méthode d'interpolation introduite par Smolyak [Smo63] consiste à considérer des polynômes de la forme

$$P_{q,d}(\mathbf{x}) = \sum_{q-d+1\leq|\mathbf{k}|\leq q} (-1)^{q-|\mathbf{k}|} \binom{d-1}{q-|\mathbf{k}|} f(x_1^{k_1}, \dots, x_d^{k_d}) l^{(1)}_{k_1} \otimes \cdots \otimes l^{(d)}_{k_d}(\mathbf{x}) \tag{2.12}$$

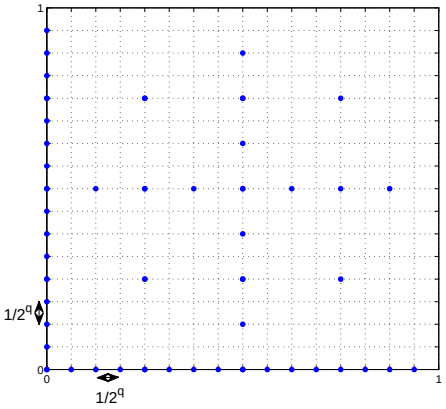


FIGURE 2.4 – Grille creuse pour l'interpolation de fonctions périodiques, avec $q = 4$, $d = 2$ et $\mathcal{X}^1_{k_1} = \mathcal{X}^2_{k_2} = \{0,\ 2^{-k_1}, \dots,\ (k_1 - 1) \times 2^{-k_1}\}$.

où $|\mathbf{k}| = k_1 + \cdots + k_d$ et des ensembles de points d'interpolation en grille "creuse" (voir Figure 2.4)

$$\mathcal{H}(q,d) = \bigcup_{q-d+1 \leq |\mathbf{k}| \leq q} \mathcal{X}^1_{k_1} \otimes \cdots \otimes \mathcal{X}^d_{k_d}.$$

Notons

$$\mathcal{P}_{q,d} = \sum_{|\mathbf{k}|=q} \mathcal{P}_{k_1-1} \otimes \cdots \otimes \mathcal{P}_{k_d-1}.$$

Alors on a le résultat suivant (voir e.g. [BNR00]) :

Proposition 2.1. *Soit f une fonction continue sur un compact $L = L_1 \times \cdots \times L_d$, et supposons que pour tout $i \in [1:d]$, $\mathcal{X}^i_1 \subseteq \mathcal{X}^i_2 \subseteq \cdots \subseteq L_i$. Alors il existe un unique polynôme d'interpolation $P_{q,d}$ dans $\mathcal{P}_{q,d}$ tel que pour tout $\mathbf{x}' \in \mathcal{H}(q,d)$, $P_{q,d}(\mathbf{x}') = f(\mathbf{x}')$. Ce polynôme est défini par la Formule (2.12).*

En pratique Dans la pratique, si les variables aléatoires d'entrée X_i $i \in [1:d]$ sont absolument continues par rapport à la mesure de Lebesgue, ou par rapport à une mesure d'équiprobabilité sur un ensemble $\{1, \dots, N\}$, avec une densité de probabilité classique w_i reliée à une famille de polynômes orthogonaux (Legendre, Hermite, etc., voir Table 1.1), alors on peut choisir les points d'interpolation comme des points dont les composantes sont les racines de ces polynômes orthogonaux — i.e. les points de quadrature de Gauss. Dans ce cas, le polynôme d'interpolation totalement tensorisé de la Formule (2.11) se décompose sur la base tensorisée des polynômes orthogonaux considérés, notés $\mathbf{\Phi_k}$, $\mathbf{k} \in \mathbb{N}^d$,

$$P_{\mathbf{n}}(\mathbf{x}) = \sum_{k_1=0}^{n_1} \cdots \sum_{k_d=0}^{n_d} \widehat{\mathrm{c}}_{\mathbf{k}} \mathbf{\Phi_k}(\mathbf{x}),$$

où les $\widehat{\mathrm{c}}_{\mathbf{k}}$ sont donnés par la formule de quadrature de Gauss

$$\widehat{\mathrm{c}}_{\mathbf{k}} = \sum_{j=0}^{n} w^j f(\mathbf{x}^j) \mathbf{\Phi_k}(\mathbf{x}^j),$$

où les poids de quadrature de Gauss w_j sont définis à partir des polynômes de Lagrange

$$w^j = \prod_{i=1}^{d} \int_{\mathrm{supp}(X_i)} l^{(i)}_j(x) w_i(x) \mathrm{d}x.$$

Les estimateurs des quantités $\sigma^2_{\mathfrak{u}}$, $\mathfrak{u} \subseteq [1:d]$, et σ^2 sont alors définies par

$$\dot{\sigma}^{2,IP}_{\mathfrak{u}} = \sum_{\mathbf{k} \in K \cap \mathbb{N}_{\mathfrak{u}}} |\widehat{\mathrm{c}}_{\mathbf{k}}|^2$$

et

$$\dot{\sigma}^{2,IP} = \sum_{\mathbf{k}\in K\cap\mathbb{N}^*} |\widehat{\mathrm{c}}_{\mathbf{k}}|^2$$

avec $K = [0:n_1]\times\cdots\times[0:n_d]$ et $\mathbb{N}_{\mathfrak{u}} = \{(k_1,\ldots,k_d) \mid \forall i\in\mathfrak{u},\ k_i\in\mathbb{N}^* \text{ et } \forall i\in\mathfrak{u}^c,\ k_i = 0\}$.

Alternativement, on peut choisir les points d'interpolation comme des points dont les composantes sont les racines des dérivées des polynômes orthogonaux — i.e. les points de quadratures de Gauss-Lobatto. Dans ce cas, on peut construire des ensembles de points d'interpolation emboîtés et satisfaire les hypothèses de la Proposition 2.1. On aboutit alors à une formulation similaire à la précédente où le polynôme d'interpolation $\mathcal{P}_{q,d}$ se décompose sur la base tensorisée des polynômes orthogonaux, et ses coefficients sont définis par un schéma de quadrature de Smolyak (voir e.g. [Bla09] Chap. 3 pour plus de détails).

Nous retiendrons simplement que, tout comme l'approche par régression linéaire, l'approche par interpolation polynomiale consiste en une optimisation dans l'espace fonctionnel alors que la question posée — i.e. le calcul des indices de Sobol' — concerne l'analyse dans le domaine spectral. Les deux remarques précédentes 2.2 et 2.3 s'appliquent donc également dans ce cas. Notons en outre, qu'il est également possible de considérer des polynômes trigonométriques d'interpolation ; cette approche est étroitement liée à la méthode FAST et est abordée au Chapitre 5.

2.4.3 Quasi-régression

Comme pour la régression linéaire, nous revenons à un cadre probabiliste, et nous considérons $\mathbf{X}^j$, $j\in[1:n]$ une famille de vecteurs aléatoires indépendants distribués suivant la loi de $\mathbf{X}$. La quasi-régression [AO01] consiste à considérer l'estimateur de régression introduit dans la Formule (2.10) en négligeant l'information donnée par la matrice $\left(\mathcal{M}(\mathbf{\Phi})^\intercal\mathcal{M}(\mathbf{\Phi})\right)^{-1}$ et en la remplaçant par sa valeur asymptotique $1/n$. Par conséquent, les coefficients $|\mathrm{c}_{\mathbf{k}}|^2$ sont simplement évalués par l'estimateur

$$\widehat{\mathrm{c}}^2_{\mathbf{k},n} = \left(\frac{1}{n}\sum_{j=1}^n f(\mathbf{X}^j)\mathbf{\Phi}_{\mathbf{k}}(\mathbf{X}^j)\right)^2;$$

certains parlent quelquefois de méthode de *projection*. On peut alors dériver un estimateur non biaisé (voir [LO02])

$$\frac{n}{n-1}\left[\left(\frac{1}{n}\sum_{j=1}^n f(\mathbf{X}^j)\mathbf{\Phi}_{\mathbf{k}}(\mathbf{X}^j)\right)^2 - \frac{1}{n^2}\sum_{j=1}^n f^2(\mathbf{X}^j)\mathbf{\Phi}^2_{\mathbf{k}}(\mathbf{X}^j)\right].$$

Cependant, plus encore que la quantité $|\mathrm{c}_{\mathbf{k}}|^2$, ce sont les quantités normalisées $\gamma^2_{\mathbf{k}} = |\mathrm{c}_{\mathbf{k}}|^2/\sigma^2$ qui sont intéressantes dans l'analyse de sensibilité. Pour ces dernières, on peut définir un estimateur comme le quotient de l'estimateur de la Formule (2.4.3) par l'estimateur de variance empirique

$$\widehat{\gamma}^2_{\mathbf{k},n} = \frac{\left(\frac{1}{n}\sum_{j=1}^n Y^j\mathbf{\Phi}_{\mathbf{k}}(\mathbf{X}^j)\right)^2}{\frac{1}{n}\sum_{j=1}^n (Y^j)^2 - \left(\frac{1}{n}\sum_{j=1}^n Y^j\right)^2}.$$

Dans ce cas, comme pour l'estimation par la méthode de Sobol', on peut montrer un théorème central limit pour ces quantités normalisées. En effet, en notant que l'estimateur $\widehat{\gamma}^2_{\mathbf{k},n}$ est invariant par toute translation des Y^j par $-\mathbb{E}[Y]$, puis considérant le vecteur

$$W^j = \left((Y^j-\mathbb{E}[Y])\mathbf{\Phi}(\mathbf{X}^j), (Y^j-\mathbb{E}[Y]), (Y^j-\mathbb{E}[Y])^2\right)^\intercal$$

et la fonction

$$\Psi(x,y,z) = \frac{x^2}{z-y^2},$$

on a que $\widehat{\gamma}^2_{\mathbf{k},n} = \Psi(\overline{W}_n)$ avec $\overline{W}_n = (W^1+\cdots+W^n)/n$. Par suite, $\sqrt{n}(\overline{W}_n - (\mathrm{c}_{\mathbf{k}},0,\sigma^2)^\intercal)$ converge en loi vers une loi normale centrée de matrice de covariance Σ, et la Delta méthode implique que $\sqrt{n}(\widehat{\mathrm{c}}^0_{\mathbf{k},n} - \mathrm{c}^0_{\mathbf{k}})$ converge en loi vers une loi normale centrée de variance $g^\intercal\Sigma g$ où

$$g = \nabla\Psi(\mathrm{c}_{\mathbf{k}},0,\sigma^2).$$

Un rapide calcul nous indique alors que cette variance asymptotique est

$$g^\intercal \Sigma g = \frac{4\mathrm{c}_{\mathbf{k}}^2}{(\sigma^2)^2}\mathrm{Var}\Big[Y\mathbf{\Phi}_{\mathbf{k}}(\mathbf{X}) - \frac{\mathrm{c}_{\mathbf{k}}}{2\sigma^2}(Y-\mathbb{E}[Y])^2\Big].$$

Plus généralement, on peut montrer que l'estimateur

$$\widehat{\gamma}_{K,n}^2 = \frac{\sum_{\mathbf{k}\in K}\left(\frac{1}{n}\sum_{j=1}^n Y^j\mathbf{\Phi}_{\mathbf{k}}(\mathbf{X}^j)\right)^2}{\frac{1}{n}\sum_{j=1}^n (Y^j)^2 - \left(\frac{1}{n}\sum_{j=1}^n Y^j\right)^2}$$

est tel que $\sqrt{n}(\widehat{\gamma}_{K,n}^2 - \gamma_{K,n}^2)$ converge en loi vers une loi normale centrée de variance

$$\frac{4}{(\sigma^2)^2}\mathrm{Var}\Big[\sum_{\mathbf{k}\in K}\mathrm{c}_{\mathbf{k}}Y\mathbf{\Phi}_{\mathbf{k}}(\mathbf{X}) - \frac{\sum_{\mathbf{k}\in K}\mathrm{c}_{\mathbf{k}}^2}{2\sigma^2}(Y-\mathbb{E}[Y])^2\Big],$$

avec $\gamma_{K,n}^2 = \sum_{\mathbf{k}\in K}\mathrm{c}_{\mathbf{k}}^2/\sigma^2$.

Pour finir notons que la quasi-régression ne souffre pas des problèmes rencontrés dans la régression linéaire et l'interpolation polynomiale mis en avant dans les Remarques 2.2 et 2.3.

2.4.4 En résumé

Les méthodes spectrales se divisent principalement en deux sous-familles. D'une part, elles comprennent des méthodes "optimisées" : régression et interpolation, et d'autre part une méthode que l'on peut qualifier de "non-optimisée" : la quasi-régression. Les deux approches optimisées peuvent être assimilées à des techniques de métamodélisation au sens où toutes les deux cherchent à obtenir une estimation des indices de Sobol' en passant au préalable par une approximation du modèle considéré. A contrario, la quasi-régression permet l'estimation des indices de Sobol' sans procéder à une approximation du modèle, et par conséquent en étant indépendante de la complexité des différents effets du modèle. Dans nos applications, nous privilégierons la quasi-régression d'autant plus qu'elle autorise, via la Delta méthode, de dériver des intervalles de confiance asymptotiques.

2.5 Notes

1) L'estimateur Monte Carlo couramment utilisé pour les $\overline{\tau}_{\mathfrak{u}}$, $\mathfrak{u}\subseteq[1:d]$ n'est pas celui introduit par Homma et Saltelli [HS96] que nous présentons dans la Formule (2.3). Dans la pratique, Saltelli et al. [SAA+10] ont montré qu'il est préférable d'utiliser la version de Jansen [Jan99]

$$\ddot{\overline{\tau}}_{\mathfrak{u},n}^2 = \frac{1}{2n}\sum_{i=1}^n \big(f(\mathbf{X}^i) - f(\mathbf{X}^i_{-\mathbf{u}} : \mathbf{Z}^i_{\mathbf{u}})\big)^2.$$

En tant qu'indice de Sobol' généralisé (voir Section 2.1.2), cet estimateur est un *contraste* — tout comme l'estimateur de la Formule (2.3) — et il est en outre positif, ce qui implique qu'il estime de manière exact les indices de Sobol' théoriquement nuls.

2) Un certain nombre d'auteurs parmi lesquels Glen et Isaacs [GI12] appréhendent les indices de Sobol' descendants en considérant la définition

$$\underline{S}_{\mathfrak{u}} = \mathbb{E}\left[\frac{f(\mathbf{X})-\mathbb{E}[Y]}{\sigma}\,\frac{f(\mathbf{X}_{\mathfrak{u}} : \mathbf{Z}_{\mathfrak{u}^c}) - \mathbb{E}[Y]}{\sigma}\right]$$

et les estiment par une simple méthode de Monte Carlo sur les sorties du modèle renormalisées. Plusieurs variantes d'estimateurs sont proposées dans [GI12] suivant des corrections de biais différentes.

3) De nombreuses autres méthodes ont été proposées pour l'estimation des indices de Sobol'. On citera par exemple les approches par polynômes locaux ou fonctionnelles de densité [Da-07] qui ont été développées pour l'estimation des indices de Sobol' dans le cas où les variables d'entrée du modèle sont corrélées. Les approches par krigeage — i.e. processus gaussien — sont également discutées dans la littérature (voir e.g. [Mar08]).

4) Dans ce chapitre, nous avons présenté l'estimation des indices de Sobol' essentiellement comme une question d'intégration numérique. On remarque néanmoins que les méthodes de quasi-Monte Carlo sont très faiblement représentées, et en particulier des suites à discrépance faible comme les (t, m, s)-nets [Nie92] brillent par leur absence. Il existe pourtant de nombreux résultats théoriques concernant ces suites, et des développements tels que la randomisation de Owen [Owe97], qui mènent aux *scrambled* (t, m, s)-nets, permettent de définir des méthodes d'intégration très performantes en lien avec la décomposition en ondelettes. Ces travaux constituent certainement un point de départ au développement de nouvelles méthodes d'estimation des indices de Sobol'.

5) Le choix de l'ensemble de troncature $\mathbf{K}$ dans les approches spectrales n'est pas précisé. Il se fait généralement de manière empirique, en sélectionnant uniquement les composantes dépassant un certain seuil. Ce seuillage, dans le cas d'une approche par quasi-régression, peut se faire plus intelligemment en prenant en compte la variance de l'estimateur (voir en particulier [JO03]). Notons en outre que ce type d'approche est évoqué au sujet de la méthode RBD par Tarantola et Koda [TK10] ; ils font référence en particulier aux travaux de Donoho et al. [DJKP95].

6) La dérivation d'un théorème central limit (TCL) pour l'estimation des indices de Sobol' par quasi-régression a été introduite initialement en 2011 par Xu et Gertner [XG11a]. Dans cet article, en s'appuyant sur la Delta méthode, les auteurs introduisent un TCL pour l'estimation des indices de Sobol' par quasi-régression dans le cas particulier d'une base trigonométrique.

Chapitre 3

Méthode de Morris et méthode basée sur les dérivées

L'exploitation des dérivées partielles d'un modèle en analyse de sensibilité remonte principalement aux années 1970 (voir e.g. [Ham94] pour des considérations historiques). On la trouve dans la littérature sous le nom de *méthode directe* ou encore d'*analyse de sensibilité différentielle*. Elle est basée sur l'approximation d'une fonction multivariée par son développement de Taylor à l'ordre 1. En effet, dans le cas d'un accroissement de toutes les variables d'entrées par un incrément de même amplitude relative ε, i.e.

$$y^0 = f(x_1^0, \dots, x_d^0) \quad \text{et} \quad y = f\big(x_1^0(1+\varepsilon), \dots, x_d^0(1+\varepsilon)\big),$$

il vient facilement

$$\frac{y - y^0}{y^0} \approx \varepsilon \sum_{i=1}^{d} \underbrace{\left(\frac{\partial f}{\partial x_i}(\mathbf{x}^0)\right)\frac{x_i^0}{y^0}}_{\mu_i}.$$

Ainsi, la variation relative de la sortie au point $\mathbf{x}^0$ se décompose comme une somme de mesures d'importance μ_i, chacune quantifiant la variation due à la i-ème variable. Le calcul des dérivées partielles au point $\mathbf{x}^0$ peut s'effectuer analytiquement en dérivant le *modèle adjoint* de f, ou si la complexité du modèle est telle que le modèle adjoint n'est pas accessible, on opte généralement pour une approximation par différences finies. Notons que l'approche par le modèle adjoint possède un coût indépendant de la dimension du modèle, car une seule simulation du modèle adjoint permet de disposer des dérivées partielles d'ordre 1 dans toutes les directions — on estime généralement le coût d'un modèle adjoint à 5 ou 6 fois le coût du modèle direct. Néanmoins, c'est une méthode intrusive qui requiert d'effectuer des manipulations à l'intérieur du code de calcul utilisé [1]. A l'opposé, l'approche par différences finies constitue une méthode non intrusive, mais possède un coût en $O(d)$.

Cette approche par dérivées partielles possède essentiellement un inconvénient. En effet, l'approximation obtenue par le développement de Taylor à l'ordre 1 ne tient que sous hypothèse de linéarité et d'additivité. Si le modèle ne possède pas ces caractéristiques extrêmement restrictives, cette méthode n'est valide que localement autour du points $\mathbf{x}_0$. Pour en faire une méthode de sensibilité globale, au sens où on étudie les variations du modèle en fonction de ses paramètres sur leurs domaines de définition tout entier, il est nécessaire d'analyser les dérivées en plusieurs points et d'opérer un traitement statistique des résultats. Cette nouvelle façon d'utiliser les dérivées en faisant une analyse "locale globalisée" a été introduite par Morris [Mor91] dans le cadre des méthodes de *criblage* — en anglais, *screening* — et a été considérablement développée depuis 2007 [CCS07, SK09, SK10, KRFPS09, LIPG12]. Dans un premier temps, nous présentons la méthode originale de Morris, ainsi que la contribution de Campolongo et al. [CCS07]. Puis dans une seconde section nous abordons les développements plus

1. Cette dérivation peut généralement se faire par différentiateur automatique si le code ne présente pas de parties trop complexes. Même dans ce cas, la méthode est qualifiée d'intrusive au sens où le modèle n'est pas considéré comme une boîte noire abstraite, mais doit être lu en détail par le différentiateur automatique.

récents et plus rigoureux qui ont mené à la définition formelle *d'indices de sensibilité basés sur les dérivées.*

Pour commencer, le cadre d'étude est purement déterministe, et on considère un modèle $y = f(x_1, \ldots, x_d)$ où f est une fonction réelle définie sur l'hypercube unité dont les dérivées partielles d'ordre 1 existent.

3.1 Méthode de Morris

3.1.1 Description de la méthode de Morris

La méthode de Morris repose principalement sur la notion *d'effet élémentaire* qui n'est rien d'autre qu'une approximation d'une dérivée partielle d'ordre 1 par un schéma aux différences finies. Pour introduire cette notion, on discrétise tout d'abord l'ensemble de définition des x_i suivant une grille régulière à p niveaux

$$\Omega = \left\{0, \frac{1}{p-1}, \ \ldots, \ \frac{p-2}{p-1}, \ 1\right\}^d$$

où p est un entier naturel non nul. Soit Δ un multiple de $\frac{1}{p-1}$ et $i \in [1:d]$; on définit l'ensemble,

$$\Omega_{\Delta,i} = \left\{\mathbf{x} \in \Omega \mid (x_1, \ldots, \ x_{i-1}, \ x_i + \Delta, \ x_{i+1}, \ldots, \ x_d) \in \Omega\right\}$$

et pour $\mathbf{x} \in \Omega_{\Delta,i}$, on introduit

$$d_i(\mathbf{x}^0) = \frac{f(x_1^0, \ldots, x_{i-1}^0, x_i^0 + \Delta, x_{i+1}^0, \ldots, x_d^0) - f(\mathbf{x}^0)}{\Delta}$$

appelé *effet élémentaire* du i-ème facteur au point $\mathbf{x}$.

Le calcul de cet accroissement permet d'évaluer l'impact de chacun des paramètres x_i sur la sortie au point $\mathbf{x}^0$. Les facteurs sont incrémentés de la quantité Δ suivant un schéma One-At-a-Time (OAT), i.e. à tour de rôle. À ce stade, l'information obtenue n'a qu'un intérêt purement local puisque l'effet élémentaire est calculé en un point $\mathbf{x}^0$ particulier. On s'intéresse donc à la variable aléatoire $d_i(\mathbf{X})$ — où implicitement $\mathbf{X}$ suit une loi uniforme sur l'hypercube unité — dont l'espérance, notée $\mu_i(\mathbf{X})$, et l'écart-type, noté $\sigma_i(\mathbf{X})$, constituent des mesures globales de l'influence du i-ème facteur. L'information contenue dans ces deux quantités permet d'exhiber une partition de la famille des paramètres d'entrée résumée dans la Table 3.1. Notons que si la distribution des $d_i(\mathbf{X})$ est symétrique par rapport à 0, alors l'espérance $\mu_i(\mathbf{X})$ est nulle quelle que soit l'amplitude des valeurs absolues des $d_i(\mathbf{X})$. Pour cette raison Campolongo et al. [CCS07] ont proposé de considérer plutôt la quantité $\mu_i^*(\mathbf{X}) = \mathbb{E}\big[|d_i(\mathbf{X})|\big]$.

	Écart-type faible	Écart-type élevé
Espérance faible	négligeable	influent (effet non-linéaire et/ou interaction avec d'autres facteurs)
Espérance élevée	influent (effet linéaire et additif)	

TABLE 3.1 – Classification des facteurs à l'issue de la méthode de Morris.

Ces espérances et ces écart-types sont estimés à l'aide d'un échantillon de r effets élémentaires où r est généralement fixé à quelques dizaines [CCS07]. Pour chaque calcul d'un effet élémentaire, on choisit une direction i, puis on tire un point uniformément dans $\Omega_{\Delta,i}$. On peut évaluer l'efficacité de la méthode par la définition d'une grandeur proche d'un rendement — appelée *economy* par Morris [Mor91] ; il s'agit de la fraction du nombre d'effets élémentaires obtenus par le nombre d'évaluations du modèle qui ont été nécessaires. Un rapide calcul de la méthode exposée ci-dessus montre que son rendement est de $\frac{1}{2}$. On peut améliorer ce résultat en utilisant une méthode plus économique, c'est l'objet de la section suivante.

3.1.2 Description de la méthode économique

Le rendement de la méthode précédente est accru en diminuant le nombre de simulations du modèle tout en conservant le nombre d'effets élémentaires obtenus. Le cadre de simulation reste inchangé : l'ensemble des entrées est modélisé par un hypercube unité discrétisé de dimension d, noté Ω, et chaque entrée possède la même probabilité d'être sélectionnée. Comme le montre Morris, cette dernière hypothèse conduit à un cadre plus restrictif. Ainsi, on se place dans le cas où

$$p \text{ est pair} \quad \text{et} \quad \Delta = \frac{p}{2(p-1)}. \tag{3.1}$$

La base de la méthode repose sur une famille de matrices particulières. Soit $\mathcal{B}(d)$ l'ensemble des matrices ayant $d+1$ lignes et d colonnes, dont les éléments sont 0 ou 1, et qui possèdent la propriété que, pour chaque colonne i, il existe exactement deux lignes qui diffèrent uniquement par leur i-ème composante ; par exemple,

$$B = \begin{bmatrix} 0 & 0 & \cdots & 0 & 0 \\ 1 & 0 & \ddots & \ddots & 0 \\ 1 & 1 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 & 0 \\ 1 & \ddots & \ddots & 1 & 0 \\ 1 & 1 & \cdots & 1 & 1 \end{bmatrix}$$

Considérons la matrice ΔB et traduisons chacune de ses lignes comme une entrée du modèle ; il est alors aisé de constater que réaliser des simulations en ces $d+1$ entrées permet d'obtenir d effets élémentaires, un pour chaque facteur. Le rendement d'une telle méthode, $\frac{d}{d+1}$, tend vers 1 lorsque d tend vers $+\infty$; dans tous les cas, il est supérieur au rendement de la méthode classique, égal à $\frac{1}{2}$.

La génération aléatoire de ces matrices se fait par le biais de la formule suivante,

$$B^* \triangleq \left[J_{d+1,1} x^* + \frac{\Delta}{2} \Big((2B - J_{d+1,d}) D^* + J_{d+1,d} \Big) \right] P^* \tag{3.2}$$

où $J_{m,k}$ sont des matrices m par k remplies de 1,
x^* est une entrée choisie aléatoirement de manière équiprobable dans $\{0, \frac{1}{p-1}, \cdots, 1-\Delta\}^d$,
D^* est une matrice diagonale de dimension d dont les éléments diagonaux 1 et -1 sont choisis avec la même probabilité,
P^* est une matrice de permutation de taille d tirée de manière équiprobable.

Avec une telle matrice, on montre aisément que, pour i fixé entre 1 et d, chaque effet élémentaire $d_i(\mathbf{x})$ calculé possède la même probabilité d'être choisi ; chaque variable aléatoire $d_i(\mathbf{X})$ est donc échantillonnée de manière uniforme.

Remarque 3.1. *Dans l'expression* (3.2)*, Morris [Mor91] fait remarquer que la matrice P^* n'est pas nécessaire pour obtenir l'équiprobabilité théorique, sa présence permet simplement de rajouter un "brassage aléatoire" supplémentaire.*

Remarque 3.2. *On peut encore généraliser cette méthode en élargissant la famille de matrices $\mathcal{B}(d)$; le nombre de lignes n'est alors plus réduit à $d+1$ et surtout, le passage d'une ligne à l'autre peut se faire en modifiant plus d'une composante, ce qui mène à un screening non-OAT. La motivation d'une telle généralisation réside essentiellement dans le fait que le rendement peut alors devenir supérieur à 1.*

3.1.3 Représentation graphique et lecture des résultats

Les quantités $\mu_i(\mathbf{X})$ — ou $\mu_i^*(\mathbf{X})$ — et $\sigma_i(\mathbf{X})$ sont représentées dans le demi-plan supérieur, les moyennes μ_i en abscisse et les écart-types σ_i en ordonnée. On fait également figurer les deux droites définies par

$$|y| = 2\frac{\sigma_i}{\sqrt{r}}\ , \tag{3.3}$$

où r est le nombre d'effets élémentaires calculés par facteur (pour des illustrations de la méthode de Morris, voir les Figures 3.1 et 3.2 dans le premier paragraphe des notes de ce chapitre). Morris explique que toute variable d'entrée se situant à l'extérieur du cône formé par ces deux droites — i.e. $|\mu_i(\mathbf{X})| \geq 2\sigma_i/\sqrt{r}$ — peut être considérée de moyenne "significativement" différente de 0. Ce dernier point reste discutable, car le raisonnement de Morris se base, a priori, sur le théorème de la limite centrale qui est un résultat valide uniquement asymptotiquement. Or, le nombre limité d'effets élémentaires calculés permet difficilement de se placer sous hypothèse asymptotique.

3.2 Indices de sensibilité basés sur les dérivées

La définition des indices de sensibilité basés sur les dérivées (ISBD) peut se voir comme une relecture rigoureuse de la méthode de Morris. Plus précisèment, lorsque les dérivées partielles d'ordre 1 de f existent et sont de carré intégrable, on s'intéresse à

$$d_i(\mathbf{x}^0) = \left(\frac{\partial f}{\partial x_i}(\mathbf{x}^0)\right)^2 \tag{3.4}$$

et en particulier à l'espérance $\nu_i = \mathbb{E}[d_i(\mathbf{X})]$. Dans la pratique, ces quantités sont estimées par une méthode de Monte Carlo ou de quasi-Monte Carlo (voir [KRFPS09]), et on ne s'éloigne que très peu de l'approche de Morris. Cet indice synthétique a été introduit récemment par Sobol' et Kucherenko [SK09], alors que certaines variantes avaient été discutées antérieurement par Sobol' et Gresham [SG95] (voir [KRFPS09]). Dans le cas général, il est essentiellement introduit pour fournir, via l'inégalité de Poincaré (voir e.g. [BBCG08]), un majorant des indices de Sobol' ascendants d'ordre 1. Nous présentons ces résultats dans la première section et nous discutons de sa généralisation aux indices ascendants d'ordre quelconque dans la seconde section.

3.2.1 Majoration des indices ascendants d'ordre 1

Cette majoration a été énoncée initialement par Sobol' et Kucherenko [SK09] pour des variables d'entrée uniformément distribuées et généralisées par Lamboni et al. [LIPG12] à des mesures de Boltzmann, i.e. des mesures absolument continues par rapport à la mesure de Lebesgue et dont la densité est de la forme

$$\rho(x) = c\exp\big(-v(x)\big), \quad x \in \mathbb{R} \tag{3.5}$$

où v est une fonction continue et c une constante de normalisation. Ces deux résultats s'énoncent comme suit

Théorème 3.1. [Sobol' et Kucherenko, 2009] *Avec les notations et sous les hypothèses qui précèdent, si $\mathbf{X}_i$ suit une loi uniforme sur $[0,1]$, alors*

$$\overline{S}_i \leq \frac{1}{\pi^2\sigma^2}\nu_i. \tag{3.6}$$

Théorème 3.2. [Lamboni, Iooss, Popelin et Gamboa, 2012] *Avec les notations et sous les hypothèses qui précèdent, si $\mathbf{X}_i$ suit une loi de Boltzmann, alors*

$$\overline{S}_i \leq \frac{4C_i^2}{\sigma^2}\nu_i, \tag{3.7}$$

où C_i est la constante de Cheeger (voir Table 3.2 pour quelques valeurs classiques)

$$C_i = \sup_{x\in\mathbb{R}} \frac{\min\big(F_i(x), 1-F_i(x)\big)}{\rho_i(x)} \tag{3.8}$$

avec F_i et ρ_i qui sont respectivement la fonction de répartition et la densité de X_i.

Dans la pratique, plus l'indice de Sobol' est petit, plus la majoration est fine.

loi de probabilité	constante de Cheeger
normale $\mathcal{N}(\mu, \sigma^2)$	$\frac{\sigma}{2}$
exponentielle $\mathcal{E}(\lambda)$, $\lambda > 0$	$\frac{1}{\lambda}$
Gumbel $\mathcal{G}(\mu, \beta)$, échelle $\beta > 0$	$\frac{\beta}{\log(2)}$
Weibull $\mathcal{W}(k, \lambda)$, $k \geq 1$, échelle $\lambda > 0$	$\frac{\lambda}{k}\log(2)^{(1-k)/k}$.

TABLE 3.2 – Constantes de Cheeger pour quelques lois de Boltzmann.

3.2.2 Majoration des indices ascendants d'ordre quelconque

Sobol' et Kucherenko [SK10] ont généralisé l'inégalité du Théorème 3.1 aux indices de Sobol' ascendants d'ordre quelconque en introduisant un nouvel indice synthétique basé sur les dérivées. Ils énoncent également un résultat similaire pour des mesures gaussiennes. Nous les réunissons dans le théorème qui suit après avoir introduit la mesure

$$\tau_{\mathrm{u}} = \sum_{i \in \mathrm{u}} \mathbb{E}\left[\frac{1 - 3X_i + 3X_i^2}{6}\left(\frac{\partial f}{\partial x_i}(X_i)\right)^2\right]. \tag{3.9}$$

Théorème 3.3. [Sobol' et Kucherenko, 2010] *Avec les notations précédentes, si* $\mathbf{X}_{\mathrm{u}}$ *suit une loi uniforme sur* $[0,1]^{|\mathrm{u}|}$*, alors*

$$\overline{S}_{\mathrm{u}} \leq \frac{24}{\pi^2 \sigma^2} \tau_{\mathrm{u}}. \tag{3.10}$$

Et si $\mathbf{X}_{\mathrm{u}}$ *est un vecteur aléatoire gaussien, alors*

$$\overline{S}_{\mathrm{u}} \leq \frac{2}{\sigma^2} \tau_{\mathrm{u}}. \tag{3.11}$$

On a également le cas particulier suivant.

Proposition 3.1. *Avec les notations précédentes, si* f *est linéaire par rapport aux variables* X_i*,* $i \in \mathrm{u}$*, alors*

$$\overline{S}_{\mathrm{u}} = \frac{\tau_{\mathrm{u}}}{\sigma^2}. \tag{3.12}$$

3.3 Notes

1) Comme nous l'avons présenté, l'incrément Δ de la méthode de Morris doit être fixé par l'utilisateur ; cela soulève la question de son impact éventuel sur les résultats. Nous exhibons le rôle de cet incrément dans un exemple élémentaire. Nous considérons un modèle additif fonction de deux paramètres A et B qui influent différemment sur la sortie d'un modèle. La variation de la sortie en fonction de chacun de ces paramètres est décrite par la Figure 3.1. En conservant les notations utilisées précédemment, on choisit un espace discrétisé Ω avec $p = 50$, et on effectue une méthode de Morris dans le cas où $\Delta = \frac{1}{49}$ — i.e. l'incrément le plus fin — et $\Delta = \frac{25}{49}$ — i.e. l'incrément imposé par la méthode économique. Dans chaque cas, on calcule $r = 25$ effets élémentaires par paramètre ; l'expérience est menée 20 fois. Les résultats présentés dans la Figure 3.2 montrent de nombreuses divergences. Dans le cas de l'incrément le plus fin, les deux paramètres peuvent être décrits comme agissant sur la sortie de manière non-linéaire et le paramètre A est légèrement plus influent que le paramètre B. Dans l'autre cas, où $\Delta = \frac{25}{49}$, le paramètre A est proche de l'influence négligeable alors que le paramètre B semble agir de manière linéaire.

2) La méthode de Morris est introduite initialement comme une méthode de criblage, i.e. un tri grossier des facteurs d'entrée afin d'indentifier ceux qui n'ont aucun effet sur la sortie, avant de mener une analyse plus fine sur les facteurs *actifs*. Par cette définition, elle doit être réalisée à faible coût en terme de nombre de simulations. Ce n'est cependant pas une chose aisée lorsque le nombre de variables d est élevé — pouvant largement dépasser la centaine — car le nombre de simulations nécessaires

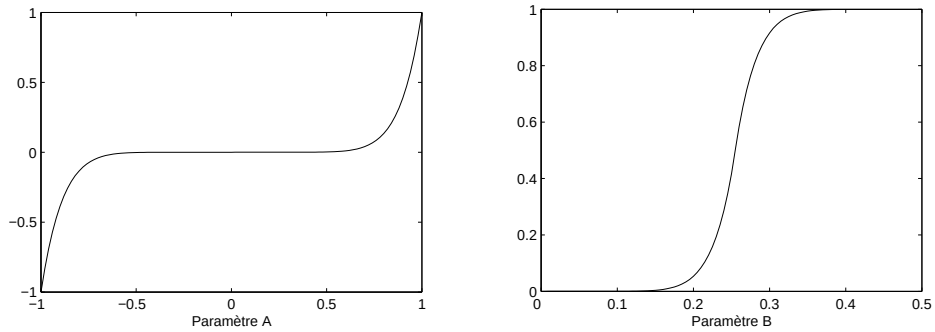


FIGURE 3.1 – Variations de la sortie en fonction des paramètres A et B

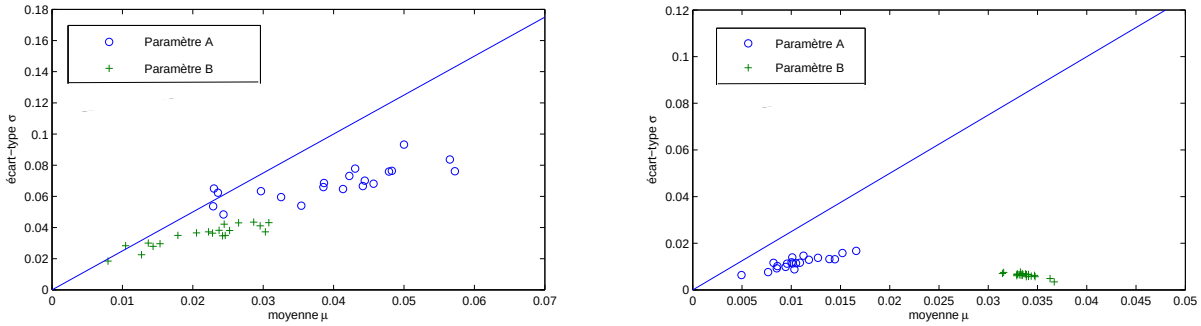


FIGURE 3.2 – Représentation graphique des statistiques issues de la méthode de Morris pour $\Delta = \frac{1}{49}$ (à gauche), et $\Delta = \frac{25}{49}$ (à droite).

dans l'analyse de Morris est en $O(d)$. Dans ce cadre, il peut être intéressant de se tourner vers une approche séquentielle dans laquelle l'évaluation des effets élémentaires d'un facteur est stoppée dès lors qu'il est identifié comme *actif* dans le modèle ; l'effort de calcul étant alors uniquement concentré sur les facteurs non correctement identifiés. Boukouvalas et al. [BGMA10] ont récemment proposé un tel algorithme séquentiel utilisant des hypercubes latins sous critère maximin, et avec un critère d'arrêt basé sur un dépassement de seuil de l'écart-type σ_i (voir Figure 1 page 9 dans [BGMA10]).

3) Comme mentionné dans [IPB+12], l'information relative à la variance des dérivées partielles lors du calcul des ISBD peut être utilisée, comme dans la méthode de Morris, pour détecter un effet linéaire et additif.

Deuxième partie

Contributions à l'estimation des indices de Sobol'

Chapitre 4

Correction de biais des estimateurs des indices de Sobol' dans les méthodes RBD et RBD-FAST

Dans ce chapitre nous revenons de façon partiellement heuristique sur le problème lié au biais des estimateurs des indices de Sobol' dans méthodes RBD et RBD-FAST. Nous proposons en particulier des méthodes de correction de biais pour les deux approches. En outre, nous introduisons une nouvelle technique basée sur des considérations combinatoires visant à estimer tous les indices de Sobol' d'ordre 1 et 2 de manière plus efficace. Ce chapitre a fait l'objet d'une publication récente dans la revue Reliability Engineering and System Safety [TP12a].

4.1 Introduction

Global sensitivity analysis of model output consists in quantifying the respective importance of input factors over their entire range of variation. Many techniques have been developed in this field (see [SCS00, SRA+08] for a review), this includes for example screening methods [Mor91], density-based methods [Bor07, BCT11] and also derivative-based methods [SK09, SK10]. But the most popular are the variance-based methods that rely on ANOVA decomposition [Sob93, Hoe48, ES81].

ANOVA decomposition and sensitivity indices Let $\mathbf{X} = (X_1, \dots, X_p)$ be a random vector and $Y = f(\mathbf{X}) \in \mathbb{R}$, where f is a square-integrable function. Under the assumption that the components of $\mathbf{X}$ are independent, the variance V of the model output Y can be decomposed as :

$$V = \sum_{k=1}^{p} \sum_{1 \le i_1 < \cdots < i_k \le p} V_{i_1 \dots i_k} \tag{4.1}$$

where

$$V_{i_1 \dots i_k} = \sum_{J \subseteq \{i_1, \dots, i_k\}} (-1)^{k - card(J)} \mathrm{Var}\big(\mathbb{E}(Y | X_j, j \in J)\big)$$

where $\mathrm{Var}(\cdot)$ and $\mathbb{E}(\cdot|\cdot)$ denote variance and conditional expectation, respectively. Thus, if $V \neq 0$ (i.e. Y is not almost surely constant), dividing both sides of (4.1) by V, yields a positive and normalized decomposition,

$$1 = \sum_{k=1}^{p} \sum_{1 \le i_1 < \cdots < i_k \le p} S_{i_1 \dots i_k} \tag{4.2}$$

where

$$S_{i_1 \dots i_k} = \frac{V_{i_1 \dots i_k}}{V} \ , \ 1 \le i_1 < \cdots < i_k \le p$$

are the so-called k^{th}-order SIs — or Sobol' indices —.

In the case of an additive model — i.e. $f(X_1, \ldots, X_p) = \sum_{k=1}^{p} f_k(X_k)$ — all terms but the first-order SI are zero and we obtain a full decomposition with only $S_1, \ldots, S_p$. On the contrary, if f is a non-additive function, it is necessary to evaluate higher-order terms to point out which interactions are significant. In practice, the first- and second-order SIs generally provide a good overview of the global variations of a model output.

FAST and its derived methods Different methods have been developed to estimate variance-based SIs, the FAST method, introduced in the 1970's, is one of the earliest. The three introduction papers [CFS+73, SS73, CSS75] describe how to compute main effects — i.e. first-order sensitivity indices — exploiting Weyl's ergodic theorem [Wey38]. Then, in a review article [CLS78], the authors precise the underlying theory considering multiple Fourier series, and suggest a decomposition of variance (see Eq. (2.29) in [CLS78]) which allows to consider higher-order SIs. But, in practice, many sources of error occur and it is generally impossible to get accurate estimates at low computational cost. As a consequence FAST has only been applied to estimate first-order and total SIs in small dimension (see the EFAST method due to Saltelli *et al.* [STC99] for total SIs).

The RBD and Hybrid FAST-RBD (HFR) methods, proposed in 2006 by Tarantola *et al.* [TGM06], partially overcome the inherent drawbacks of FAST using a new sampling technique based on Satterthwaite's random balance designs [Sat59]. These methods have been introduced to estimate first-order SIs, and as Mara [Mar09] notices, it is also possible to estimate SIs of any order or closed and total sensitivity indices, using the HFR method (renamed RBD-FAST).

Recently Plischke [Pli10] derived another FAST-like method, named Effective Algorithm for computing global Sensitivity Indices (EASI), which estimates sensitivity indices with any input sample while FAST, RBD and RBD-FAST use specific experimental designs.

In Section 4.2, we briefly recall the FAST method and discuss the different sources of error that affect the accuracy of SI estimates. In Section 4.3, we present the specific problem of interferences in RBD which leads to the positive bias of the first-order SIs and we propose a bias correction method. In Section 4.4, we extend this technique to the sensitivity indices of any order in RBD-FAST, and in Section 4.5, we describe an efficient strategy to estimate all the first and second-order SI using RBD-FAST. Numerical examples are presented in Section 4.6 to illustrate the accuracy of the proposed bias correction method. Conclusions and ideas for a future work are summarised in Section 4.7.

4.2 Sources of error in the FAST method

4.2.1 Description of the FAST method

The FAST method is based on a specific experimental design — the so-called search curve — which allows to use discrete Fourier transform. The experimental design $(x^k)_{k=1\ldots N}$ is such that

$$x_i^k = G_i\big(\sin(\omega_i s_k + \varphi_i)\big) \ , \quad i = 1, \ldots, p, \ k = 1, \ldots, N \tag{4.3}$$

where the ω_i's are integer frequencies — free of interferences up to a certain order (see Section 4.2.2), the G_i's are functions to be settled so as to impose probability density functions on the input variables X_i, φ_i are random phase-shifts and $(s_k)_{k=1\ldots N}$ is defined as

$$s_k = \frac{2\pi(k-1)}{N} \ .$$

In particular, to uniformly sample the marginal distributions over $[0, 1]$, one shall use (see for example [STC99]),

$$G_i(\cdot) = \frac{1}{\pi} \arcsin(\cdot) + \frac{1}{2} \ .$$

The Fourier spectrum of the discrete signal $\Big(f\big(x_1^j, \ldots, x_p^j\big)\Big)_{j=1\ldots N}$ can be decomposed with respect

to the frequencies $\omega_1, \ldots, \omega_p$, and the following estimators can be defined,

$$\widehat{V} = \sum_{1\leq|n|\leq N/2} |\hat{c}_n|^2 \ , \tag{4.4}$$

$$\widehat{V_i} = \sum_{1\leq|k|\leq N_1} |\hat{c}_{k\omega_i}|^2 \ , \tag{4.5}$$

$$\widehat{V_{ij}} = \sum_{2\leq|k|+|l|\leq N_2} |\hat{c}_{k\omega_i+l\omega_j}|^2 \ , \tag{4.6}$$

and so on ; where N_1 is the highest harmonic considered as non-negligible, N_2 is the order over which the linear combinations of ω_i and ω_j are considered as negligible, and

$$\hat{c}_n = \frac{1}{N}\sum_{j=1}^{N} f\big(x_1^j, \ldots, x_p^j\big) e^{-in\frac{2\pi(j-1)}{N}} \ , \quad -\frac{N}{2} \leq n \leq \frac{N}{2} \tag{4.7}$$

is the n-th complex discrete Fourier coefficient. Finally, dividing (4.5) (resp. (4.6)) by (4.4), we get the estimator of a first-order (resp. second-order) SI :

$$\widehat{S_i} = \frac{\sum_{1\leq|k|\leq N_1} |\hat{c}_{k\omega_i}|^2}{\sum_{1\leq|n|\leq N/2} |\hat{c}_n|^2} \ , \tag{4.8}$$

$$\widehat{S_{ij}} = \frac{\sum_{2\leq|k|+|l|\leq N_2} |\hat{c}_{k\omega_i+l\omega_j}|^2}{\sum_{1\leq|n|\leq N/2} |\hat{c}_n|^2} \ . \tag{4.9}$$

The accuracy of these estimates naturally depends on the sample size and we can observe an empirical convergence to the theoretical values as N tends to $+\infty$. But the dependence is intricate ; in addition to the truncation error, we distinguish two main sources of error.

4.2.2 Interferences

Whenever a linear combination of the frequencies $\omega_1, \ldots, \omega_p$ is equal to zero, some parts of variance could be attributed by error to other ones in the decomposition of the Fourier spectrum. For example, if $-2\omega_1 + \omega_2 = 0$, the discrete Fourier coefficient $\hat{c}_{2\omega_1} = \hat{c}_{\omega_2}$ contains information from both X_1 and X_2, and should not be totally attributed to $\widehat{S_1}$ and $\widehat{S_2}$. These interferences can sometimes cause a bias, and to alleviate their effect, we adopt the criterion proposed by Schaibly and Shuler [SS73] to choose frequency sets free of interferences up to a certain order M,

$$\sum_{i=1}^{p} a_i\omega_i \neq 0 \ \text{ for } \ \sum_{i=1}^{p} |a_i| \leq M+1 \ . \tag{4.10}$$

4.2.3 Aliasing

Only linear combinations such that $-N/2 < \omega = \sum_{i=1}^{p} a_i\omega_i < N/2$ are unambiguously represented by the discrete sampled signal. If ω is out of this range, its spectral component is falsely attributed to another frequency inside the Fourier spectrum. To avoid this aliasing phenomenon, which can lead to positively biased estimators, it is necessary to satisfy the Nyquist-Shannon theorem, i.e. to impose that the sampling rate is large enough. As a consequence, the sample size is bounded from below as follows :

$$N \geq 2M \max_{1\leq i\leq p} \omega_i \tag{4.11}$$

where M is defined in the previous paragraph. Practitioners generally set

$$N_1 = N_2 = \cdots = N_d = M, \tag{4.12}$$

but this constraint is not necessary, and the criterion stated in (4.10) and (4.11) can be formulated in a more general way. Indeed, for all $1 \leq q \leq p$, consider $N_q \in \mathbb{N}^*$, and for all $1 \leq i_1 < \cdots < i_q \leq p$, define

$$A_{i_1 \ldots i_q} = \left\{ (a_1, \ldots, a_p) \in \mathbb{Z}^p \;\middle|\; \forall i \notin \{i_1, \ldots, i_q\},\ a_i = 0 \text{ and } \sum_{1 \leq m \leq q} |a_{i_m}| \leq N_q \right\}$$

and

$$A = \bigcup_{q=1}^{p} \bigcup_{1 \leq i_1 < \cdots < i_q \leq p} A_{i_1 \ldots i_q}.$$

Hence we propose to replace (4.10) and (4.11) by

$$\sum_{i=1}^{p} a_i \omega_i \neq 0 \quad \text{for all } (a_1, \ldots, a_p) \in A \tag{4.13}$$

and

$$N \geq 2 \max_{(a_1, \ldots, a_p) \in A} \sum_{i=1}^{p} a_i \omega_i, \tag{4.14}$$

respectively. Note that if (4.12) is satisfied, (4.13) and (4.14) are equivalent to the classic criterion stated in (4.10) and (4.11).

4.3 Random balance design method

As we noted in the previous section, using a distinct frequency per input factor in the FAST method imposes restrictive constraints on the sample size. To overcome this drawback, an alternative sampling method is employed in RBD.

4.3.1 Sampling method

In contrary to FAST, in the RBD method, all the ω_i are equal to a unique frequency ω and input variables are distinguished by taking random permutations of the coordinates of the sample points. Let $\sigma_1, \ldots, \sigma_p$ denote random permutations on the set $\{1, \ldots, N\}$, the experimental design $(x^k)_{k=1 \ldots N}$ is such that

$$x_i^k = G_i\big(\sin(\omega s_{\sigma_i(k)})\big), \quad \forall i = 1, \ldots, p \ \text{ and } \ \forall k = 1, \ldots, N\ .$$

One shall choose an odd integer N to get a good space-filling design. In this case, RBD technique is very close to Latin hypercube sampling introduced in 1979 (see [MCB79]) ; the only difference is that the RBD design points are located at the center of the cells (see Fig. 4.1).

4.3.2 Estimator

RBD sampling method can be used to estimate first-order SI. The estimator of the total variance is defined as in FAST and the part of variance due to the factor X_i is estimated by

$$\widehat{V_i} = \sum_{1 \leq |k| \leq N_1} |\hat{c}_{k\omega}^{\sigma_i}|^2 \tag{4.15}$$

with

$$\hat{c}_{k\omega}^{\sigma_i} = \frac{1}{N} \sum_{j=1}^{N} f\big(x_1^{\sigma_i^{-1}(j)}, \ldots, x_p^{\sigma_i^{-1}(j)}\big) e^{-ik\omega \frac{2\pi(j-1)}{N}}\ , \tag{4.16}$$

where σ_i^{-1} is the inverse permutation of σ_i. Indeed, considering a fixed i, the design points $\big(x_1^{\sigma_i^{-1}(j)}, \ldots, x_p^{\sigma_i^{-1}(j)}\big)_{j=1 \ldots N}$ are such that the i^{th} coordinate is sampled with respect to the frequency ω and the other ones are

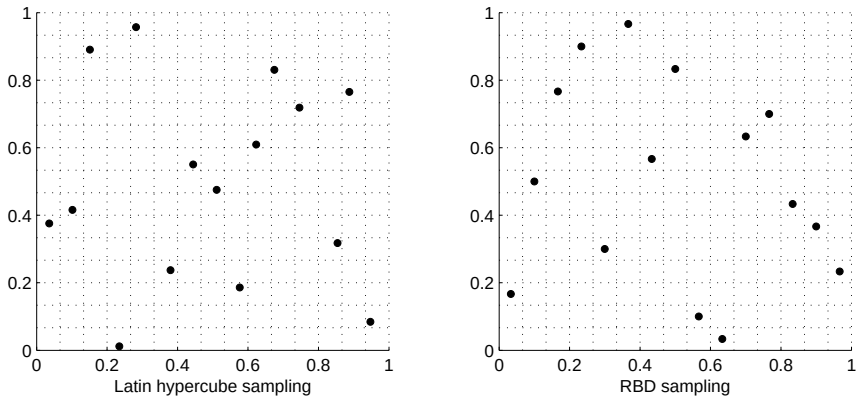


FIGURE 4.1 – Comparison between Latin hypercube and RBD samples in two-dimensional unit hypercube with sample size 15.

sampled in a random way because

$$\begin{aligned} x_k^{\sigma_i^{-1}(j)} &= G_k\Big(\sin\big(\omega s_{\sigma_k(\sigma_i^{-1}(j))}\big)\Big) \\ &= \begin{cases} G_k\big(\sin(\omega s_j)\big) & \text{if } k = i \\ G_k\big(\sin\big(\omega s_{\sigma_k^i(j)}\big)\big) & \text{if } k \neq i \end{cases} \end{aligned} \tag{4.17}$$

where $\sigma_k^i = \sigma_k \circ \sigma_i^{-1}$ is almost surely a non-trivial permutation. Therefore, in the Fourier spectrum of the signal

$$\Big(f\big(x_1^{\sigma_i^{-1}(j)}, \ldots, x_p^{\sigma_i^{-1}(j)}\big)\Big)_{j=1\ldots N}, \tag{4.18}$$

the harmonics of ω are attributed to the partial variance of X_i. Thus, using FAST estimator, we get Eqs. (4.15) and (4.16).

Remark 4.1. *The choice of the frequency ω seems to be of secondary importance. However, to avoid aliasing, the most efficient value is the smallest one, typically $\omega = 1$. In this case, the aliasing phenomenon is negligible and consequently, there is no more restriction on the sample size as in Eq.* (4.11).

4.3.3 Bias

As we explained in the last section, the RBD estimator is so defined because the harmonics of ω in the signal $\Big(f\big(x_1^{\sigma_i^{-1}(j)}, \ldots, x_p^{\sigma_i^{-1}(j)}\big)\Big)_{j=1\ldots N}$ are supposed to be only related to the part of variance V_i due to X_i. But it is essential to notice that, since the factors $(X_k)_{k\neq i}$ are randomly sampled, the remaining part of variance — denoted V_{-i} — appears in the signal $\Big(f\big(x_1^{\sigma_i^{-1}(j)}, \ldots, x_p^{\sigma_i^{-1}(j)}\big)\Big)_{j=1\ldots N}$ as a random noise. Therefore, a random fraction of each harmonic of ω is related to V_{-i} and is falsely attributed to V_i. Xu and Gertner [XG11b] quantified this interference between the harmonics of ω and the random noise, showing that for any $\hat{c}_{k\omega}^{\sigma_i}$ we have

$$E\big(|\hat{c}_{k\omega}^{\sigma_i}|^2\big) = |c_{k\omega}^{\sigma_i}|^2 + \frac{V_{-i}}{N}$$

where $c_{k\omega}^{\sigma_i}$ denotes the theoretical unbiased k^{th} harmonic of ω. Thus, following Eq. (4.15), we define the bias-corrected estimator of V_i as

$$\widehat{V_i^c} = \widehat{V_i} - \frac{2N_1}{N}\widehat{V_{-i}}$$

where $\widehat{V_{-i}}$ is an estimator of V_{-i} defined, assuming the bias correction, as

$$\widehat{V_{-i}} = \widehat{V} - \widehat{V_i^c} \ .$$

Hence

$$\widehat{V_i^c} = \widehat{V_i} - \frac{2N_1}{N}\big(\widehat{V} - \widehat{V_i^c}\big),$$

and dividing both sides of the equality by $\widehat{V}$, we obtain

$$\widehat{S_i^c} = \widehat{S_i} - \frac{2N_1}{N}\big(1 - \widehat{S_i^c}\big)$$

where $\widehat{S_i}$ et $\widehat{S_i^c}$ are the RBD estimator of the first-order sensitivity index and the corrected one, respectively. Finally, setting $\lambda = \frac{2N_1}{N}$, we get the explicit formula

$$\widehat{S_i^c} = \widehat{S_i} - \frac{\lambda}{1-\lambda}\big(1 - \widehat{S_i}\big) \ .$$

Remark 4.2. *It is important to observe that the larger N and S_i, the lower the bias.*

Remark 4.3. *In his paper, Plischke* [Pli10] *suggests to apply exactly the same bias correction to the EASI estimates (see Eq. (7) in* [Pli10]*). His approach is based on a bias correction method for correlation ratios due to Kelley* [Kel35].

4.4 Hybrid approach : RBD-FAST

The underlying idea in RBD-FAST is to combine both RBD and FAST sampling approaches. Therefore, this new method is naturally faced with the classical drawbacks of FAST, but in a lesser extent. The main interest of the hybrid approach is that estimation of higher-order SI is possible.

4.4.1 Sampling method

For this purpose, the p input variables are divided into groups of approximatively equal cardinal. Then each group is assigned a distinct random permutation and each factor of the group a distinct frequency chosen from a frequency set assumed free of interferences up to a given order (see Section 4.2.2.). For example, we can have the following configurations :

$$\begin{array}{ll}
\text{6 factors :} & \begin{array}{cccccc} X_1 & X_2 & X_3 & X_4 & X_5 & X_6 \\ \underbrace{\omega_1 \ \ \omega_2 \ \ \omega_3}_{\sigma_1} & & & \underbrace{\omega_1 \ \ \omega_2 \ \ \omega_3}_{\sigma_2} & & \end{array} \\
\text{6 factors :} & \begin{array}{cccccc} X_1 & X_2 & X_3 & X_4 & X_5 & X_6 \\ \underbrace{\omega_1 \ \ \omega_2}_{\sigma_1} & & \underbrace{\omega_1 \ \ \omega_2}_{\sigma_2} & & \underbrace{\omega_1 \ \ \omega_2}_{\sigma_3} & \end{array} \\
\text{7 factors :} & \begin{array}{ccccccc} X_1 & X_2 & X_3 & X_4 & X_5 & X_6 & X_7 \\ \underbrace{\omega_1 \ \ \omega_2 \ \ \omega_3}_{\sigma_1} & & & \underbrace{\omega_1 \ \ \omega_2}_{\sigma_2} & & \underbrace{\omega_1 \ \ \omega_2}_{\sigma_3} & \end{array}
\end{array} \ . \tag{4.19}$$

Remark 4.4. *Tarantola et al.* [TGM06] *and Mara* [Mar09] *present RBD-FAST (or HFR) in another way : the p input variables are partitioned in the same way but the permutations are applied within the groups and a different frequency is associated to each group. Actually, the methods are strictly equivalent ; these just are two different points of view.*

4.4.2 Estimators

This hybrid sampling method allows to define the estimator of SI of any order. In particular, considering two factors inside the m^{th} group associated with the frequencies ω_i and ω_j respectively, we can define the part of variance of their interaction as

$$\widehat{V_{ij}} = \sum_{2\leq|k|+|l|\leq N_2} |c^{\sigma_m}_{k\omega_i+l\omega_j}|^2 \tag{4.20}$$

where N_2 is the value over which the linear combinations of ω_i and ω_j are considered as negligible and where

$$c^{\sigma_m}_{k\omega_i+l\omega_j} = \frac{1}{N}\sum_{n=1}^{N} f\big(x_1^{\sigma_m^{-1}(n)},\dots,x_p^{\sigma_m^{-1}(n)}\big)e^{-i(k\omega_i+l\omega_j)\frac{2\pi(n-1)}{N}} \ . \tag{4.21}$$

In the same way, considering a factor inside the m^{th} group associated with the frequency ω_i, we can define its part of variance as

$$\widehat{V_i} = \sum_{1\leq|k|\leq N_1} |c^{\sigma_m}_{k\omega_i}|^2 \tag{4.22}$$

where N_1 is the highest harmonic considered as non-negligible and with

$$c^{\sigma_m}_{k\omega_i} = \frac{1}{N}\sum_{n=1}^{N} f\big(x_1^{\sigma_m^{-1}(n)},\dots,x_p^{\sigma_m^{-1}(n)}\big)e^{-ik\omega_i\frac{2\pi(n-1)}{N}} \ . \tag{4.23}$$

Indeed, considering the sample points $\big(x_1^{\sigma_m^{-1}(j)},\dots,x_p^{\sigma_m^{-1}(j)}\big)_{j=1\dots N}$ where m is fixed, for $1\leq k\leq p$, we have

(i) if X_k is associated with the couple (ω_i,σ_m) then

$$x_k^{\sigma_m^{-1}(j)} = G_k\Big(\sin\big(\omega_i s_{\sigma_m(\sigma_m^{-1}(j))}\big)\Big) = G_k\big(\sin\big(\omega_i s_j\big)\big),$$

(ii) if X_k is associated with a couple (ω_i,σ_n), for $n\neq m$, then

$$x_k^{\sigma_m^{-1}(j)} = G_k\Big(\sin\big(\omega_i s_{\sigma_n(\sigma_m^{-1}(j))}\big)\Big)$$

where $\sigma_n\circ\sigma_m^{-1}$ is almost surely a non-trivial permutation. Therefore, all input variables outside the group associated with σ_m are randomly sampled, and the other ones are sampled with respect to their frequencies. Applying FAST's estimator, Eqs. (4.20)–(4.23) follow.

4.4.3 Bias

The phenomenon leading to positive biases described for the RBD method occurs in the same way for RBD-FAST. Therefore parts of variance can be corrected with an analogous technique.

Let $X_{m_1},\dots,X_{m_d}$ be the d input factors inside the m^{th} group, and P be a nonempty subset of $\{m_1,\dots,m_d\}$. We denote V_P the part of variance due to the interaction between the input variables $(X_i)_{i\in P}$ (e.g. if $P=\{i\}$, V_P is simply V_i, and if $P=\{i,j\}$, V_P is V_{ij}). Let $\widehat{V_P}$ be the RBD-FAST classical estimator of V_P, previously described in Eqs. (4.20) and (4.22) for $card(P)=1$ and 2. Following RBD bias correction, we first define the estimator of the positive bias B_P as

$$\widehat{B_P} = \frac{n(P)}{N}\widehat{V_{-P}}$$

and the corrected estimator of V_P as

$$\widehat{V_P^c} = \widehat{V_P} - \widehat{B_P}$$

where $n(P)$ is the number of Fourier coefficients taken into account to estimate V_P. $\widehat{V_{-P}}$ is an estimate of the part of variance which is not due to any subset of factors contained in $\{m_1, \ldots, m_d\}$ defined, assuming the bias correction, as

$$\widehat{V_{-P}} = \widehat{V} - \sum_{\substack{Q \subseteq \{m_1,\ldots,m_d\} \\ Q \neq \emptyset}} \widehat{V_Q^c} \ .$$

Hence

$$\widehat{V_P^c} = \widehat{V_P} - \frac{n(P)}{N}\left(\widehat{V} - \sum_{\substack{Q \subseteq \{m_1,\ldots,m_d\} \\ Q \neq \emptyset}} \widehat{V_Q^c}\right) ,$$

and dividing both sides of the equality by $\widehat{V}$, we get

$$\widehat{S_P^c} = \widehat{S_P} - \frac{n(P)}{N}\left(1 - \sum_{\substack{Q \subseteq \{m_1,\ldots,m_d\} \\ Q \neq \emptyset}} \widehat{S_Q^c}\right) \tag{4.24}$$

where $\widehat{S_P}$ and $\widehat{S_P^c}$ are the RBD-FAST estimator of the SI S_P and the corrected one, respectively. Then setting

$$\lambda_Q = \frac{n(Q)}{N} \quad \text{for any nonempty subset } Q \in \{m_1, \ldots, m_d\} \tag{4.25}$$

and

$$\overline{\lambda} = \sum_{\substack{Q \subseteq \{m_1,\ldots,m_d\} \\ Q \neq \emptyset}} \lambda_Q \ ,$$

we conclude with the explicit formula

$$\widehat{S_P^c} = \widehat{S_P} - \frac{\lambda_P}{1-\overline{\lambda}}\left(1 - \sum_{\substack{Q \subseteq \{m_1,\ldots,m_d\} \\ Q \neq \emptyset}} \widehat{S_Q}\right). \tag{4.26}$$

(see details in Appendix 4.A).

Remark 4.5. *This bias correction formula requires the knowledge of the biased estimators $\widehat{S_Q}$ of any order relative to the input factors $(X_i)_{i\in P}$. Unfortunately, the estimation of the terms over a certain order is quite difficult ; so in practice, it is necessary to neglect SI over a certain degree δ and to consider the following bias correction*

$$\widehat{S_P^c} = \widehat{S_P} - \frac{\lambda_P}{1-\overline{\lambda}}\left(1 - \sum_{\substack{Q \subseteq \{m_1,\ldots,m_d\} \\ Q \neq \emptyset,\ card(Q) \leq \delta}} \widehat{S_Q}\right) \tag{4.27}$$

where

$$\overline{\lambda} = \sum_{\substack{Q \subseteq \{m_1,\ldots,m_d\} \\ Q \neq \emptyset,\ card(Q) \leq \delta}} \lambda_Q \ . \tag{4.28}$$

Remark 4.6. *An analagous formula for closed SI can be deduced from* (4.26)*. Keeping the same notations as previously, such indices are defined as*

$$S_P^{closed} = \sum_{Q \subseteq P,\ Q \neq \emptyset} S_Q$$

and we have

$$\widehat{S}_P^{closed,c} = \widehat{S}_P^{closed} - \frac{\lambda_P^{closed}}{1-\overline{\lambda}}\left(1 - \sum_{\substack{Q \subseteq \{m_1,\ldots,m_d\} \\ Q \neq \emptyset}} \widehat{S_Q}\right)$$

where $\widehat{S}_P^{closed}$ and $\widehat{S}_P^{closed,c}$ are the RBD-FAST estimator of the SI S_P^{closed} and the corrected one respectively, and

$$\lambda_P^{closed} = \sum_{Q \subseteq P,\ Q \neq \emptyset} \lambda_Q \ .$$

4.5 An efficient strategy to estimate both first- and second-order sensitivity indices

Throughout this section, we develop a strategy using RBD-FAST to get all the bias-corrected estimates of the first- and second-order SI of a model in which we assume that the SI over a certain order δ are negligible. In this case, we can get the first-order and second-order indices by applying Eqs. (4.27) and (4.28).

However, contrarily to the RBD method in which all the main effects of any model can be estimated using only one experimental design, the computation of all the first-order and second-order indices using RBD-FAST requires a number of sample sets increasing with the number of factors p. Through an example, Mara [Mar09] observes that 5 sample sets are necessary to estimate all the 15 second-order SI — and naturally the first-order ones — of a 6-dimensional model. In fact, in the case of 6 input factors, the number of experimental designs can be restricted to 4. More generally, we establish that the required number of experimental designs is equal to :

$$\begin{array}{ll} 1 + \min\limits_{\substack{\sqrt{p} \leq q \\ q \text{ prime}}} q & \text{for } p \geq 4, \\ 1 & \text{for } p \leq 3. \ , \end{array} \tag{4.29}$$

where p is the number of input factors. Low-dimensional models — $p \leq 3$ — can be treated using FAST method with only one design of experiments ; in the other cases we implement a strategy based on elementary combinatory considerations.

It has to be noted that, in Mara's paper [Mar09], input variables are divided into groups of 2 factors, while our configurations can contain subgroups of more than 2 factors. Thus, the constraints on the sample size that arise from FAST — see Eqs. (4.10) and (4.11) — are more restrictive in our approach. Nevertheless, as we can observe in Table 4.1 p73, at the same computational cost, our strategy provides second-order SI estimates with smaller variance.

4.5.1 Designs of experiments in the case $p = q^2$ with q prime

In this particular case, the different configurations of the designs of experiments required to estimate all the first-order and second-order SI are quite natural. First, we divide the set of input variables $\{X_1, \ldots, X_p\}$ into q groups of q factors ; for example, in the case $p = 9$, we can have,

$$\text{configuration } 0 : \quad \underbrace{X_4 X_1 X_5}_{G_1^0} \quad \underbrace{X_7 X_9 X_2}_{G_2^0} \quad \underbrace{X_3 X_8 X_6}_{G_3^0} \ . \tag{4.30}$$

Following RBD-FAST approach, each group receives a set of free of interferences frequencies and is randomly permuted. This allows to estimate the second-order indices S_{14}, S_{15}, S_{45}, S_{27}, S_{29}, S_{79}, S_{36}, S_{38} and S_{68}, and all the first-order terms.

We then obtain the other configurations applying the following rules :

(R1) each of the new configurations is a partition of the input variables into q groups of q factors,

(R2) each group in the new configurations is filled with one factor of each original group $(G_i^0)_{i=1\ldots q}$,

(R3) if a set of two distinct variables $\{X_i, X_j\}$ is already contained in a group G_n^k, then we are not allowed to define a group G_m^l, with $l \neq k$ and $m \neq n$, in a next configuration containing both X_i and X_j.

For instance, in the case $p = 9$, it is only possible to create three new configurations,

$$\begin{array}{lccc} \text{configuration } 1 : & \underbrace{X_9 X_4 X_8}_{G_1^1} & \underbrace{X_7 X_5 X_6}_{G_2^1} & \underbrace{X_3 X_2 X_1}_{G_3^1} \\ \text{configuration } 2 : & \underbrace{X_6 X_1 X_9}_{G_1^2} & \underbrace{X_3 X_7 X_4}_{G_2^2} & \underbrace{X_2 X_8 X_5}_{G_3^2} \\ \text{configuration } 3 : & \underbrace{X_7 X_1 X_8}_{G_1^3} & \underbrace{X_5 X_3 X_9}_{G_2^3} & \underbrace{X_6 X_2 X_4}_{G_3^3} . \end{array} \tag{4.31}$$

Here, it is easy to notice that these four configurations 0, 1, 2 and 3 allow to compute one estimate of all the second-order SI and four estimates of all the first-order terms.

More generally, we have the following proposition :

Proposition 4.1. *In the case $p = q^2$ with q prime, there exists an efficient strategy using $q+1$ designs of experiments and allowing to compute $q+1$ estimates of all the first-order SI and one estimate of all the second-order terms.*

Proof. See Appendix 4.B.

4.5.2 Experimental designs for any p

In the general case, we first define

$$q^* = \min_{\substack{\sqrt{p} \leq q \\ q \text{ prime}}} q \ ,$$

and

$$p^* = \left(q^*\right)^2 \ .$$

Following the strategy presented in the previous section, we can create $q+1$ designs of experiments with p^* factors, $X_1,\ldots,X_p,\ldots,X_{p^*}$. We then delete variables $X_{p+1},\ldots,X_{p^*}$ in all configurations. For example, considering an 8-dimensional model, we get $q^* = 3$ and $p^* = 9$, and we can use the designs of experiments presented in Eqs. (4.30) and (4.31), and deleting the factor X_9, we get

$$\begin{array}{lccc}
\text{configuration 0 :} & \underbrace{X_4X_1X_5}_{G_1^0} & \underbrace{X_7X_2}_{G_2^0} & \underbrace{X_3X_8X_6}_{G_3^0} \\
\text{configuration 1 :} & \underbrace{X_4X_8}_{G_1^1} & \underbrace{X_7X_5X_6}_{G_2^1} & \underbrace{X_3X_2X_1}_{G_3^1} \\
\text{configuration 2 :} & \underbrace{X_6X_1}_{G_1^2} & \underbrace{X_3X_7X_4}_{G_2^2} & \underbrace{X_2X_8X_5}_{G_3^2} \\
\text{configuration 3 :} & \underbrace{X_7X_1X_8}_{G_1^3} & \underbrace{X_5X_3}_{G_2^3} & \underbrace{X_6X_2X_4}_{G_3^3}.
\end{array} \tag{4.32}$$

Hence, for any p, we have an economical strategy for which the number of experimental designs satisfies Eq. (4.29).

Remark 4.7. *Elaborating economical strategies is also of major importance for the Sobol' method in which the curse of dimensionality is clearly problematic. In particular, one can cite the work of Saltelli* [Sal02] *who provides an economical way to estimate all the first-order, second-order and total SI using the Sobol' method.*

4.6 Numerical tests

The accuracy of the proposed bias correction method is tested on the g-function introduced by Sobol' (see e.g. [SS95]). Considering uniformly distributed independent input variables $(X_i)_{i=1\ldots,p}$ on the unit hypercube, this function is defined as

$$f(X_1,\ldots,X_p) = \prod_{i=1}^{p} g_i(X_i)$$

where $g_i(X_i)$ is given by

$$g_i(X_i) = \frac{|4X_i - 2| + a_i}{1 + a_i} \ .$$

We consider a 6-dimensional g-function where $(a_i) = (0, 0, 0, 0.5, 0.5, 0, 5)$, so that the three first parameters are important, the others are less important and interactions are quite important. We then add three dummy factors X_7, X_8 and X_9 that don't play any role in the model.

The bias correction method and the efficient strategy are tested on this 9-dimensional model.

4.6.1 Test on RBD

The correction method is tested using increasing sample sizes, $N = 501$ and $N = 2001$ (see Figs. 2 and 3). In both cases, we estimate all the first-order SI with the basic RBD method and with the corrected one. The experiment is replicated 200 times using different random permutations.

We observe that the corrected boxplots are centered on the analytical values whatever the sample size. On the contrary, in the absence of correction method, the estimates are considerably biased, even for a large sample size. For a low sample size, we can notice that the bias correction is of great importance because a factor without any effect on the output can appear as a nonnegligible one using the basic RBD method (see B_7, B_8 and B_9 in Fig. 4.2).

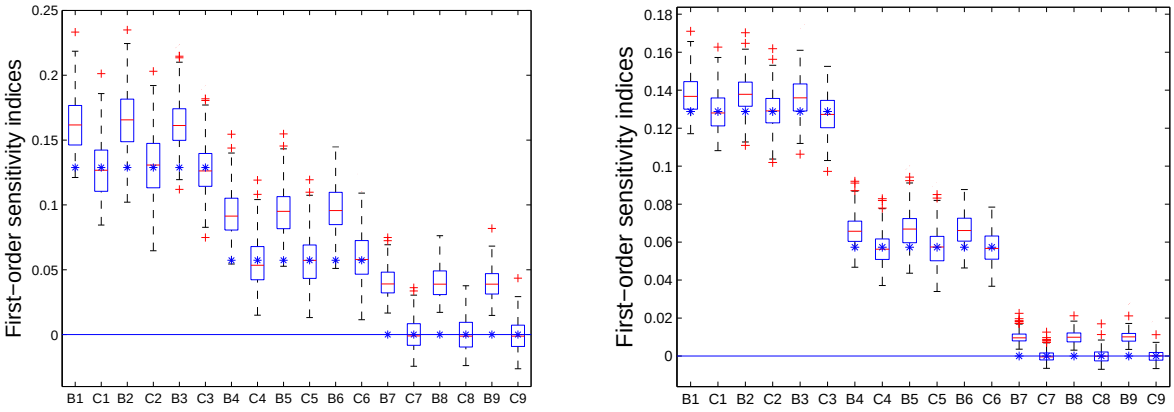


FIGURE 4.2 – Estimation of the first-order SI using RBD. We compare, for a fixed sample size $N = 501$ (on the left side) and $N = 2001$ (on the right side), the basic estimator (B1 to B9) with the bias-corrected one (C1 to C9). In each column, we mark the theoretical SI with a blue asterisk and plot several summaries of a sample of 200 estimator replicates : the red central mark is the median ; the box has its lower and upper edges at the 25^{th} percentile q and the 75^{th} percentile Q, respectively ; the whiskers extend between $q - 1.5(Q - q)$ and $Q + 1.5(Q - q)$; the red crosses are outliers.

4.6.2 Tests on RBD-FAST

Computations using the efficient strategy

In this section, we test the bias correction method on RBD-FAST. Applying the efficient strategy using RBD-FAST, we estimate all the first- and second-order SI using only 4 experimental designs — those presented in Eqs. (4.30) and (4.31) — with sample size 4001. Following Remark 5, we neglect the third-order effects — their contribution in the variance is theoretically lower than 10% —, so we apply Eqs. (4.27) and (4.28) with $\delta = 2$.

Here, designs are constructed using different random permutations and the set of frequencies free of interferences is $\{\omega_1, \omega_2, \omega_3\} = \{177, 186, 193\}$. We show on Figs. 4.3–4.4 boxplots of 200 replicates ; all first-order SI are shown in Fig. 4.3, and a representative subset of the second-order SI is shown on Fig. 4.4. As in the previous test, the corrected indices are centered on their respective theoretical value ; but some differences exist between main effects and interaction estimations. On the one hand, first-order terms are accurately evaluated, and their bias, in the absence of correction, are rather low ; on the other hand, interaction estimates suffer from a more important variance and a larger bias in absence of correction. Two main reasons justify the difference between the variances. Firstly the first-order terms are evaluated thanks to 4 estimates per indices while the second-order ones are computed with only one estimate, and secondly the complexity of SI grows with the order. In terms of bias, the lower performance of the interaction estimations without correction is essentially due to the larger number of frequencies taken into account to evaluate the second-order indices. Indeed,

considering Eq. (4.27), we can notice that the amplitude of the bias :

$$\frac{\lambda_P}{1-\overline{\lambda}}\left(1-\sum_{\substack{Q\subseteq\mathcal{G}(P),Q\neq\emptyset\\ card(Q)\leq\delta}}\widehat{S_Q}\right)$$

is proportional to $\lambda_P = n(P)/N$. In this test, we have $n(P) = 2N_1 = 2\times 10 = 20$ for the first-order SI, and $n(P) = 2N_2(N_2-1) = 2\times 7\times(7-1) = 84$ for the second-order SI. Note that the frequency set $\{177, 186, 193\}$ satisfies the criterion stated in Eqs. (4.13) and (4.14) with parameters $N_1 = 10$, $N_2 = 7$ and $N_3 = 0$.

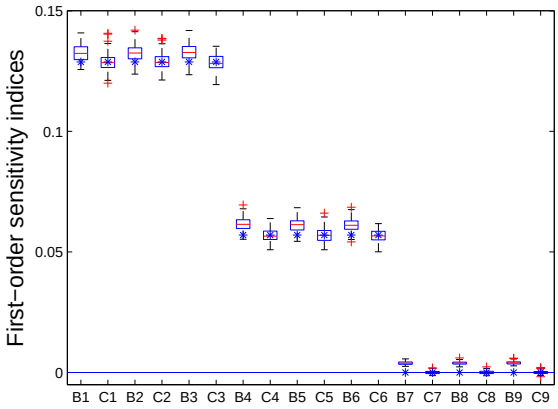


FIGURE 4.3 – Estimation of the first-order SI using RBD-FAST. We compare, for a fixed sample size $N = 4001$, the basic estimator (B1 to B9) with the bias-corrected one (C1 to C9). In each column, we mark the theoretical SI with a blue asterisk and plot several summaries of a sample of 200 estimator replicates : the red central mark is the median ; the box has its lower and upper edges at the 25^{th} percentile q and the 75^{th} percentile Q, respectively ; the whiskers extend between $q - 1.5(Q-q)$ and $Q + 1.5(Q-q)$; the red crosses are outliers.

Comparison with Mara's approach

We now estimate all the first- and second-order SI using the strategy described in Mara [Mar09]. With such an approach, input variables are divided into 4 groups of 2 factors and 1 single term. Hence, 9 experimental designs have to be employed. To keep the same computational cost as for the previous experiment in Section 4.2.1, sample size is 1791 and we use the set of frequencies $\{\omega_1, \omega_2\} = \{79, 83\}$. Note that this frequency set satisfies the criterion stated in Eqs. (4.13) and (4.14) with parameters $N_1 = 10$, and $N_2 = 7$. The experiment is replicated 200 times using different random permutations, and results (empirical mean and variance for each strategy) are reported in Table 4.1. On the one hand the accuracy of first-order SI estimates is the same, and on the other hand we observe that the efficient strategy provides second-order indices with lower variance. We conclude that the choice of strategy seems to be important in terms of variance reduction.

4.7 Conclusion

In this paper we presented a bias correction method for the estimation of SI of any order by both RBD and RBD-FAST. In particular, as we can notice through the numerical tests, this technique successfully avoids the over-estimation of the first-order and second-order indices, for any sample size.

We also introduced a strategy which, combined with the bias correction method, provides an efficient way to estimate all the first-order and second-order indices using RBD-FAST. In particular, this kind of approach allows to get a good overview of the sensitivity of a model output at a low cost.

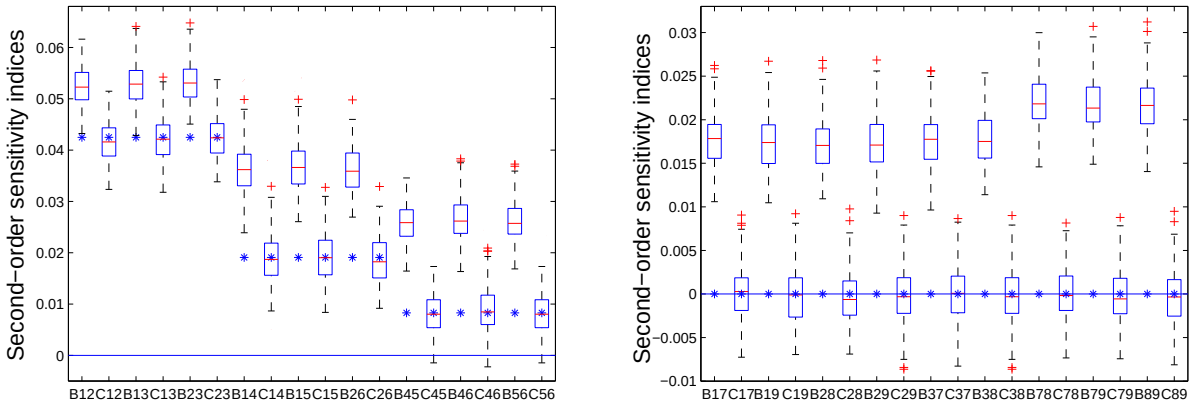


FIGURE 4.4 – Estimation of the second-order SI using RBD-FAST. We compare, for a fixed sample size $N = 4001$, the basic estimator (Bij) with the bias-corrected one (Cij). In each column, we mark the theoretical SI with a blue asterisk and plot several summaries of a sample of 200 estimator replicates : the red central mark is the median ; the box has its lower and upper edges at the 25^{th} percentile q and the 75^{th} percentile Q, respectively ; the whiskers extend between $q - 1.5(Q - q)$ and $Q + 1.5(Q - q)$; the red crosses are outliers.

	S_1	S_4	S_7	S_{14}	S_{17}	S_{47}	S_{12}	S_{45}	S_{78}
Theoretical value	0.129	0.057	0	0.019	0	0	0.043	0.008	0
Mean ES	0.129	0.057	0	0.019	0	0	0.042	0.008	0
Variance ES($\times 10^{-5}$)	1.1	0.8	0.1	2.5	0.9	1.0	1.9	1.9	1.1
Mean MS	0.129	0.057	0	0.019	0	0	0.042	0.008	0
Variance MS($\times 10^{-5}$)	1.3	0.8	0.1	10.0	6.4	6.0	9.9	9	6

TABLE 4.1 – Estimation of the first and second-order SI using the RBD-FAST method with sample size 4001 with Mara's strategy (MS) and the proposed efficient strategy (ES). We give, together with the theoretical value of the SI, the empirical means and variances of a sample of 200 estimator replicates.

Finally this efficient strategy introduces the question of variance reduction techniques (see Section 4.2.2), and a further work is to improve RBD and RBD-FAST sampling methods. In particular, optimization algorithms commonly used for Latin hypercube sampling could be adapted for RBD experimental designs which are, as we have noticed in Section 4.3, very close to Latin hypercube designs.

Acknowledgments

This work has been partially supported by French National Research Agency (ANR) through COSINUS program (project COSTA-BRAVA n° ANR-09-COSI-015).

4.A Details on formula (4.26)

We denote by $(P_i)_{i=1\dots n}$ the nonempty subsets of $\{m_1, \dots, m_d\}$ where n is given by

$$n = \sum_{k=1}^{d} \binom{d}{k} = 2^d - 1 \ ,$$

and, to simplify the notations, we denote by λ_i the coefficients λ_{P_i}. Applying Eq. (4.24) to each of the P_i, we get the linear system

$$\begin{pmatrix} \widehat{S_{P_1}} \\ \widehat{S_{P_2}} \\ \vdots \\ \widehat{S_{P_{n-1}}} \\ \widehat{S_{P_n}} \end{pmatrix} = \underbrace{\begin{pmatrix} 1-\lambda_1 & -\lambda_1 & \cdots & \cdots & -\lambda_1 \\ -\lambda_2 & 1-\lambda_2 & -\lambda_2 & \cdots & -\lambda_2 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ -\lambda_{n-1} & \cdots & \cdots & 1-\lambda_{n-1} & -\lambda_{n-1} \\ -\lambda_n & \cdots & \cdots & -\lambda_n & 1-\lambda_n \end{pmatrix}}_{A} \begin{pmatrix} \widehat{S^c_{P_1}} \\ \widehat{S^c_{P_2}} \\ \vdots \\ \widehat{S^c_{P_{n-1}}} \\ \widehat{S^c_{P_n}} \end{pmatrix} + \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_{n-1} \\ \lambda_n \end{pmatrix} \tag{4.33}$$

The determinant Δ of the matrix of the system — denoted A — is easy to compute. Subtracting the first column to all other ones, we get

$$\Delta = \begin{vmatrix} 1-\lambda_1 & -1 & \cdots & \cdots & -1 \\ -\lambda_2 & 1 & 0 & \cdots & 0 \\ \vdots & 0 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ -\lambda_n & 0 & \cdots & 0 & 1 \end{vmatrix} \tag{4.34}$$

and, using Laplace expansion,

$$\Delta = 1 - \lambda_1 - \lambda_2 \cdots - \lambda_n \ .$$

In practice, we fix N so that

$$\sum_{i=1}^{n} card(P_i) < N$$

Hence, with the definition in Eq. (4.25), we have

$$\sum_{i=1}^{n} \lambda_i < 1 \ .$$

This implies that Δ is positive; in particular A is invertible.

We get A^{-1} using the formula based on the adjugate matrix,

$$A^{-1} = \frac{{}^t\mathrm{adj}(A)}{\Delta} \ .$$

We easily obtain,

$$\mathrm{adj}(A) = \begin{pmatrix} \Delta+\lambda_1 & \lambda_2 & \cdots & \lambda_{n-1} & \lambda_n \\ \lambda_1 & \Delta+\lambda_2 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \Delta+\lambda_{n-1} & \lambda_n \\ \lambda_1 & \lambda_2 & \cdots & \lambda_{n-1} & \Delta+\lambda_n \end{pmatrix} .$$

Finally we invert the linear system (4.33). It comes

$$\begin{pmatrix} \widehat{S^c_{P_1}} \\ \widehat{S^c_{P_2}} \\ \vdots \\ \widehat{S^c_{P_{n-1}}} \\ \widehat{S^c_{P_n}} \end{pmatrix} = \begin{pmatrix} 1+\frac{\lambda_1}{\Delta} & \frac{\lambda_1}{\Delta} & \cdots & \cdots & \frac{\lambda_1}{\Delta} \\ \frac{\lambda_2}{\Delta} & 1+\frac{\lambda_2}{\Delta} & \frac{\lambda_2}{\Delta} & \cdots & \frac{\lambda_2}{\Delta} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \frac{\lambda_n}{\Delta} & \cdots & \cdots & \frac{\lambda_n}{\Delta} & 1+\frac{\lambda_n}{\Delta} \end{pmatrix} \begin{pmatrix} \widehat{S_{P_1}} \\ \widehat{S_{P_2}} \\ \vdots \\ \widehat{S_{P_{n-1}}} \\ \widehat{S_{P_n}} \end{pmatrix} - \begin{pmatrix} \frac{\lambda_1}{\Delta} \\ \frac{\lambda_2}{\Delta} \\ \vdots \\ \frac{\lambda_{n-1}}{\Delta} \\ \frac{\lambda_n}{\Delta} \end{pmatrix} \tag{4.35}$$

and we conclude that Eq. (4.26) holds.

4.B Proof of Proposition 4.1

Let $p = q^2$ with q prime. It is obvious that if there exists $q+1$ designs of experiments satisfying the rules established in Section 4.5.1, then these configurations allow to compute $q+1$ estimates of all first-order SI and one estimate of all second-order terms. So, to show that an efficient strategy exists, it is sufficient to prove the existence of such configurations under the rules (R1), (R2) and (R3) of Section 4.5.1. We give a constructive proof.

We begin by renaming the factors $(X_i)_{i=1\ldots p}$, and defining an initial configuration,

$$\text{configuration } 0: \quad \underbrace{X_1^1 \cdots X_1^q}_{G_1^0} \quad \underbrace{X_2^1 \cdots X_2^q}_{G_2^0} \quad \cdots \quad \underbrace{X_q^1 \cdots X_q^q}_{G_q^0}$$

where $X_i^j = X_{(i-1)q+j}$. We then obtain the q other experimental designs by considering for $i = 1, \ldots, q$

$$\text{configuration } i: \quad \underbrace{X_1^{\sigma_i^1(1)} \cdots X_q^{\sigma_i^q(1)}}_{G_1^i} \quad \underbrace{X_1^{\sigma_i^1(2)} \cdots X_q^{\sigma_i^q(2)}}_{G_2^i} \quad \cdots \quad \underbrace{X_1^{\sigma_i^1(q)} \cdots X_q^{\sigma_i^q(q)}}_{G_q^i} \tag{4.36}$$

where for all i and j between 1 and q, σ_i^j is a permutation on the set $\{1, \ldots, q\}$. These configurations obviously satisfy rules (R1) and (R2) since each group $(G_j^i)_{j=1..q}$ is filled with one factor of each original group $(G_k^0)_{k=1..q}$; but (R3) is not always verified. However we can observe that, letting c be a cyclic permutation of order q, the permutations

$$\sigma_i^j = c^{ij} = \underbrace{c \circ c \circ \cdots \circ c}_{ij \text{ times}} \tag{4.37}$$

allow to satisfy rule (R3). Indeed, following the formalism of Eq. (4.36), rule (R3) reads as : for all i, i', k, k', j_1 and j_2 between 1 and q, with $i < i'$ and $j_1 \neq j_2$, either the factor from $G_{j_1}^0$ inside G_k^i — i.e. $X_{j_1}^{\sigma_i^{j_1}(k)}$ — is different from the factor from $G_{j_1}^0$ inside $G_{k'}^{i'}$ — i.e. $X_{j_1}^{\sigma_{i'}^{j_1}(k')}$ — or the factor from $G_{j_2}^0$ inside G_k^i — i.e. $X_{j_2}^{\sigma_i^{j_2}(k)}$ — is different from the factor from $G_{j_2}^0$ inside $G_{k'}^{i'}$ — i.e. $X_{j_2}^{\sigma_{i'}^{j_2}(k')}$ —. That is to say

$$\forall\ 1 \leq i, i', k, k', j_1, j_2 \leq q,\ i < i',\ j_1 \neq j_2,\ \begin{cases} \sigma_i^{j_1}(k) \neq \sigma_{i'}^{j_1}(k') \\ \text{or} \\ \sigma_i^{j_2}(k) \neq \sigma_{i'}^{j_2}(k') \ . \end{cases}$$

So, assuming Eq. (4.37), let's prove that

$$\forall\ 1 \leq i, i', k, k', j_1, j_2 \leq q,\ i < i',\ j_1 \neq j_2,\ \begin{cases} c^{ij_1}(k) \neq c^{i'j_1}(k') \\ \text{or} \\ c^{ij_2}(k) \neq c^{i'j_2}(k') \ . \end{cases}$$

Suppose, by contradiction, that

$$c^{ij_1}(k) = c^{i'j_1}(k') \quad \text{and} \quad c^{ij_2}(k) = c^{i'j_2}(k')$$

for some (i, i', k, k', j_1, j_2) with $i \neq i'$ and $j_1 \neq j_2$. It follows that

$$c^{(i-i')(j_1-j_2)}(k) = k \ .$$

Then, c being a cyclic permutation of order q with q prime and i being different from i', we deduce that $c^{(i-i')}$ is a cyclic permutation of order q. Hence, $j_1 - j_2 = qr$ for a certain integer r. But, assuming $1 \leq j_1, j_2 \leq q$, we conclude that $r = 0$ and $j_1 = j_2$, a contradiction to our assumption $j_1 \neq j_2$. The conclusion follows.

Erratum dans l'article [TP12a] publié dans la revue Reliabilty Engineering and System Safety :

Dans [TP12a] à la page 207, entre les Formules (24) et (25), ainsi que dans ce chapitre à la page 63, entre les Formules (4.17) et (4.18), il faut lire : "where $\sigma_k^i = \sigma_k \circ \sigma_i^{-1}$ is a non-trivial permutation with overwhelming probability.". En effet, la probabilité en question est égale à $1 - 1/(n!)$ et non à 1 comme cela est affirmé.

Chapitre 5

Nouvelle introduction aux méthodes FAST et RBD

Dans ce chapitre, nous effectuons un travail essentiellement théorique visant à réintroduire de manière rigoureuse les méthodes FAST et RBD. Pour ce faire, nous proposons une relecture de ces techniques au regard de l'analyse harmonique sur les sous-groupes finis du tore unité, et de la notion de tableau orthogonal. Nous évitons ainsi l'écueil lié à l'approximation du théorème ergodique de Weyl [Wey38] sur lequel FAST et RBD sont construites initialement. Cela nous permet de mieux comprendre le fonctionnement de ces méthodes, de les généraliser, et également d'investiguer de manière rigoureuse la problématique du biais dans la méthode RBD. Ce chapitre a fait l'objet d'une soumission dans la revue Information and Inference (Oxford Journals) [TP12c].

5.1 Introduction

Variance-based sensitivity analysis consists in computing indices — the so-called variance-based sensitivity indices (SI) or Sobol' indices (see [Sob93]) — that are essentially multiple integrals. Many numerical techniques have been developed to estimate these quantities. This includes the crude Monte Carlo estimator (see [Sob93], and [JKL+12] for a recent work), the polynomial chaos-based estimators (see [Sud08] and [BS10]) and the FAST method (see [CLS78] and [STC99]) as well as its derived approach, RBD (see [TGM06]), and their hybrid approach, RBD-FAST (see [TGM06] and [Mar09]), and many others (see [SCS00] for a review).

The main purpose of this paper is to revisit FAST and RBD by using the discrete harmonic analysis framework, in order to carry out a theoretical error analysis. In these methods the SI estimation amounts to computing a finite number of the complex Fourier coefficients of the model of interest defined on the unit hypercube. In theory these computations could be done by performing a crude Monte Carlo integration or a cubature on a regular grid. But the rate of convergence of the Monte Carlo method is low, and cubatures are generally unfeasible in high dimension because of the exponential growth of the number of nodes, also known as the curse of dimensionality.

A first possible starting point to overcome these drawbacks is to note that the discrete complex Fourier coefficients computed by using the cubature approach are exactly the coefficients in the representation of the trigonometric interpolation polynomial of the model of interest on the regular grid. Consequently this approach consists of a trigonometric interpolation issue and can be generalized by using Smolyak algorithm on sparse grids (see [DS89]). Such interpolation schemes are quite efficient as long as the model of interest is sufficiently smooth (see [BG04]). But the matrix of the interpolation operator in such a method suffers from an increase of its condition number for both increasing refinement of the regular grid and increasing model dimension, and thus makes the interpolation scheme unstable (see [KK11]).

As a consequence, it turns out to be obvious that, in order to avoid the stability issue, one has to focus on unitary operators. Thus DFT operators on finite subgroups of the torus (see e.g. [Loo53]) — i.e. the unit hypercube view as a group — whose matrices have a perfect condition number equal

to 1 are particularly well-suited in the present framework. This leads to the use of lattice rules (see [SJ94] for a review) to which FAST, as shown in Subsection 5.4.1, is closely related. In a second time, by viewing finite subgroups of the torus as orthogonal arrays (see [HSS99] for a review), the previous method can be generalized by performing a randomization process on these arrays. This leads to the use of randomized orthogonal arrays in numerical integration (see [Owe94] and references therein) to which RBD, as shown in Subsection 5.4.2, is closely related.

The paper proceeds as follows. In Section 5.2, we set up the notation, we give background materials related to the ANOVA decomposition and to the Fourier series representation, and we introduce the class of estimators of interest. In Section 5.3, we first review both FAST and RBD, and then revisit them. Section 5.4 is devoted to the error analysis by using the revisited definition provided in Section 5.3. At last, Section 5.5 gives numerical illustrations of RBD estimates on an analytical model. Most of the proofs of the propositions are given in appendices.

5.2 Background

5.2.1 Notation

First, $\mathbb{E}[Y]$, $\mathbb{E}[Y|X]$ and $\mathrm{Var}[Y]$ denote the unconditional expectation of Y, the conditional expectation of Y given X and the variance of Y, respectively. By convention, we define $\mathbb{E}[Y|\emptyset] = \mathbb{E}[Y]$. Secondly, consider a parameter d in $\mathbb{N}^*$ — the dependence on which is omitted for convenience — and define for any $\mathfrak{u} \in \{1, \ldots, d\}$,

$$\begin{aligned} \mathbb{Z}_{\mathfrak{u}} &= \{\mathbf{k} \in \mathbb{Z}^d \mid \forall i \in \mathfrak{u},\ k_i \in \mathbb{Z} \text{ and } \forall i \notin \mathfrak{u},\ k_i = 0\} \\ \mathbb{Z}^*_{\mathfrak{u}} &= \{\mathbf{k} \in \mathbb{Z}^d \mid \forall i \in \mathfrak{u},\ k_i \in \mathbb{Z}^* \text{ and } \forall i \notin \mathfrak{u},\ k_i = 0\} \end{aligned}$$

and for all $i \in \mathbb{N}^*$,

$$\begin{aligned} \mathbb{Z}_{\mathfrak{u}}(i) &= \mathbb{Z}_{\mathfrak{u}} \cap \Big(-\frac{i}{2}, \frac{i}{2}\Big]^d \\ \mathbb{Z}^*_{\mathfrak{u}}(i) &= \mathbb{Z}^*_{\mathfrak{u}} \cap \Big(-\frac{i}{2}, \frac{i}{2}\Big]^d . \end{aligned}$$

Lastly, a design of experiments is commonly denoted by D and, for $i \in \mathbb{N}^*$, the notation $D(i)$ refers to the regular grid in $[0,1)^d$

$$D(i) = \left\{0, \frac{1}{i}, \ldots, \frac{i-1}{i}\right\}^d . \tag{5.1}$$

5.2.2 Variance-based sensitivity indices

Let $\mathbf{X} = (X_1, \ldots, X_d) \in [0,1]^d$ be a d-dimensional random vector and let us consider $Y = f(\mathbf{X})$ where $f : [0,1]^d \to \mathbb{R}$ is a measurable function such that $\mathbb{E}[Y^2] < +\infty$. Under the assumption that $\mathbf{X}$ has independent components, the Hoeffding decomposition [Hoe48, Van98] states that Y can be uniquely decomposed into summands of increasing dimensions

$$Y - \mathbb{E}[Y] = \sum_{m=1}^{d} \sum_{\substack{\mathfrak{u} \subseteq \{1,\ldots,d\} \\ |\mathfrak{u}|=m}} f_{\mathfrak{u}}(X_i, i \in \mathfrak{u}) \tag{5.2}$$

where the $2^d - 1$ random variables on the right-hand side of (5.2) should satisfy the property

$$\forall \mathfrak{v} \subsetneq \mathfrak{u},\ \ \mathbb{E}\big[f_{\mathfrak{u}}(X_i, i \in \mathfrak{u}) | X_i, i \in \mathfrak{v}\big] = 0 . \tag{5.3}$$

Note that in this case the random variables $f_{\mathfrak{u}}(X_i, i \in \mathfrak{u})$ have mean zero and are mutually uncorrelated. Therefore taking the variance of both sides in (5.2) gives the variance decomposition [ES81, Sob93] of Y

$$\mathrm{Var}[Y] = \sum_{m=1}^{d} \sum_{\substack{\mathfrak{u} \subseteq \{1,\ldots,d\} \\ |\mathfrak{u}|=m}} \mathrm{Var}\big[f_{\mathfrak{u}}(X_i, i \in \mathfrak{u})\big]. \tag{5.4}$$

Finally, if $\mathrm{Var}[Y] \neq 0$, we define the so-called variance-based sensitivity indices — or Sobol' indices — as

$$\mathrm{S}_{\mathfrak{u}}(f, \mathbf{X}) = \frac{\mathrm{Var}\big[f_{\mathfrak{u}}(X_i, i \in \mathfrak{u})\big]}{\mathrm{Var}[Y]} \ . \tag{5.5}$$

In practice, global sensitivity analysis focuses on computing the first-order ($|\mathfrak{u}| = 1$) and the second-order ($|\mathfrak{u}| = 2$) terms.

5.2.3 Fourier series representation

From here on let us assume that the X_i's are independent and uniformly distributed on $[0,1]$. Therefore the joint probability density function of $\mathbf{X}$ on $[0,1]^d$ is equal to 1 and, denoting

$$P_{\mathbf{n}}(f, \mathbf{X}) = \sum_{k_1=-n_1}^{n_1} \cdots \sum_{k_d=-n_d}^{n_d} \mathrm{c}_{\mathbf{k}}(f) \exp(2i\pi \mathbf{k} \cdot \mathbf{X}) \tag{5.6}$$

where

$$\mathrm{c}_{\mathbf{k}}(f) = \int_{[0,1]^d} f(\mathbf{X}) \exp(-2i\pi \mathbf{k} \cdot \mathbf{X}) d\mathbf{X} \ , \tag{5.7}$$

the Riesz-Fischer theorem yields

$$P_{\mathbf{n}}(f, \mathbf{X}) \quad \xrightarrow{L^2} \quad Y \ . \tag{5.8}$$

In particular, we have

$$Y = \sum_{k_1 \in \mathbb{Z}} \cdots \sum_{k_d \in \mathbb{Z}} \mathrm{c}_{\mathbf{k}}(f) \exp(2i\pi \mathbf{k} \cdot \mathbf{X}) \ \text{ a. s.} \tag{5.9}$$

and as the following proposition shows, this Fourier series representation gives an harmonic approach to handle the variance-based sensitivity indices.

Proposition 5.1. *Let $X_1, \ldots, X_d$ be independent random variables uniformly distributed on $[0,1]$ and let us consider $Y = f(\mathbf{X})$ where $f : [0,1]^d \to \mathbb{R}$ is a measurable function such that $\mathbb{E}[Y^2] < +\infty$ and $\mathrm{Var}[Y] \neq 0$. Then for any non-empty subset $\mathfrak{u}$ of $\{1, \ldots, d\}$ we have*

$$\mathrm{S}_{\mathfrak{u}}(f, \mathbf{X}) = \frac{\displaystyle\sum_{\mathbf{k} \in \mathbb{Z}_{\mathfrak{u}}^*} \big|\mathrm{c}_{\mathbf{k}}(f)\big|^2}{\displaystyle\sum_{\mathbf{k} \in (\mathbb{Z}^d)^*} \big|\mathrm{c}_{\mathbf{k}}(f)\big|^2} \ . \tag{5.10}$$

Démonstration. In view of (5.9), it is easy to notice that the components in the Hoeffding decomposition satisfy

$$f_{\mathfrak{u}}(X_i, i \in \mathfrak{u}) = \sum_{\mathbf{k} \in \mathbb{Z}_{\mathfrak{u}}^*} \mathrm{c}_{\mathbf{k}}(f) \exp(2i\pi \mathbf{k} \cdot \mathbf{X}) \ \text{ a. s.} \tag{5.11}$$

and the conclusion follows from Parseval's identity. □

As in (5.10) the index $\mathrm{S}_{\mathfrak{u}}(f, \mathbf{X})$ does no more depend on $\mathbf{X}$ we now simply denote the sensitivity indices by $\mathrm{S}_{\mathfrak{u}}(f)$. In the same way, we now denote $\mathrm{V}_{\mathfrak{u}}(f)$ and $\mathrm{V}(f)$ the parts of variance $\mathrm{Var}\big[f_{\mathfrak{u}}(X_i, i \in \mathfrak{u})\big]$ and the total variance $\mathrm{Var}[Y]$, respectively. Lastly, when $\mathfrak{u} = \{i_1, \ldots, i_s\}$ is explicitely given, we use the more common notation $\mathrm{V}_{i_1 \ldots i_s}(f)$ and $\mathrm{S}_{i_1 \ldots i_s}(f)$.

5.2.4 Estimation

We now define basic estimators based on Proposition 5.1. For any non-empty subset $\mathfrak{u}$ of $\{1, \ldots, d\}$, let $K_{\mathfrak{u}}$ be a finite subset of $\mathbb{Z}_{\mathfrak{u}}^*$ and D a finite subset of $[0,1)^d$ with $|D| = n$. Denoting

$$\widehat{\mathrm{c}}_{\mathbf{k}}(f, D) = \frac{1}{n} \sum_{\mathbf{x} \in D} f(\mathbf{x}) \exp(-2i\pi \mathbf{k} \cdot \mathbf{x}), \tag{5.12}$$

we define the estimator of $\mathrm{V}_{\mathrm{u}}(f)$ as the truncated series

$$\widehat{\mathrm{V}}_{\mathrm{u}}(f, K_{\mathrm{u}}, D) = \sum_{\mathbf{k} \in K_{\mathrm{u}}} |\widehat{\mathrm{c}}_{\mathbf{k}}(f, D)|^2 \ , \tag{5.13}$$

the estimator of $\mathrm{V}(f)$ as the empirical variance

$$\widehat{\mathrm{V}}(f, D) = \frac{1}{n} \sum_{\mathbf{x} \in D} \Big(f(\mathbf{x}) - \frac{1}{n} \sum_{\mathbf{y} \in D} f(\mathbf{y}) \Big)^2 \tag{5.14}$$

and the estimator of $\mathrm{S}_{\mathrm{u}}(f)$ naturally as

$$\widehat{\mathrm{S}}_{\mathrm{u}}(f, K_{\mathrm{u}}, D) = \frac{\widehat{\mathrm{V}}_{\mathrm{u}}(f, K_{\mathrm{u}}, D)}{\widehat{\mathrm{V}}(f, D)}. \tag{5.15}$$

Example 5.1. *If the design of experiments D is a set of independent random points uniformly distributed on $[0,1]^d$ and*

$$K = \bigsqcup_{\substack{\mathrm{u} \subseteq \{1, \dots, d\} \\ \mathrm{u} \neq \emptyset}} K_{\mathrm{u}}, \tag{5.16}$$

we have

$$\widehat{\mathrm{V}}_{\mathrm{u}}(f, K_{\mathrm{u}}, D) = \mathrm{V}_{\mathrm{u}}(\tilde{f}) \tag{5.17}$$

where

$$\tilde{f}(\mathbf{X}) = \sum_{\mathbf{k} \in K \cup \{\mathbf{0}\}} \widehat{\mathrm{c}}_{\mathbf{k}}(f, D) \mathrm{e}^{2i\pi \mathbf{k} \cdot \mathbf{X}} \tag{5.18}$$

is the approximation of $f(\mathbf{X})$ using the quasi-regression approach [AO01] *based on the random sample D. Note that $|\widehat{\mathrm{c}}_{\mathbf{k}}(f, D)|^2$ is a biased estimator of $|\mathrm{c}_{\mathbf{k}}(f, D)|^2$ and it is recommended to use the unbiased estimator*

$$\frac{n}{n-1} \left(|\widehat{\mathrm{c}}_{\mathbf{k}}(f, D)|^2 - \frac{1}{n^2} \sum_{\mathbf{x} \in D} f^2(\mathbf{x}) \right) \tag{5.19}$$

(see e.g. [LO02]*). In the same way, the empirical variance $\widehat{\mathrm{V}}(f, D)$ should be replaced by the unbiased sample variance $\frac{n}{n-1} \widehat{\mathrm{V}}(f, D)$.*

Example 5.2. *If the design of experiments D is the regular grid $D(q)$ — with $n = q^d$, $q \in \mathbb{N}^*$ — and if for all non-empty subsets u of $\{1, \dots, d\}$, $K_{\mathrm{u}} = \mathbb{Z}^*_{\mathrm{u}}(q)$ and*

$$K = \bigsqcup_{\substack{\mathrm{u} \subseteq \{1, \dots, d\} \\ \mathrm{u} \neq \emptyset}} K_{\mathrm{u}} \tag{5.20}$$

then by Parseval's identity, it can be easily shown that

$$\widehat{\mathrm{S}}_{\mathrm{u}}\big(f, K_{\mathrm{u}}, D(q)\big) = \mathrm{S}_{\mathrm{u}}(\tilde{f}) \tag{5.21}$$

where

$$\tilde{f}(\mathbf{x}) = \sum_{\mathbf{k} \in K} \widehat{\mathrm{c}}_{\mathbf{k}}\big(f, D(q)\big) \mathrm{e}^{2i\pi \mathbf{k} \cdot \mathbf{x}} \tag{5.22}$$

is the trigonometric interpolation polynomial of $f(\mathbf{x})$ (see e.g. [DR84]*) at the $n = q^d$ equally spaced nodes $\mathbf{x} \in D(q)$.*

5.3 New introduction to FAST and RBD

In the sequel, since the X_i's are independent and uniformly distributed on $[0,1]$, we have

$$\mathbb{E}\big[f(\mathbf{X})\big] = \int_{[0,1]^d} f(\mathbf{x}) d\mathbf{x} \tag{5.23}$$

so we use no more probabilistic notation. Moreover, the integrability assumption on f now reads $f \in L^2([0,1]^d)$.

5.3.1 Review of FAST

Numerical integration

FAST is essentially an application of the following result due to Weyl [Wey38] (see also the Weyl's ergodic theorem [Wey16] in German or [Sin77])

Theorem 5.1. [Weyl] *Let g be a bounded Riemann integrable function on $[0,1]^d$ and for all $i = 1,\dots,d$, $x_i(t) = \{\omega_i t\}$ where the ω_i's are real numbers linearly independent over $\mathbb{Q}$ and $\{\cdot\}$ denotes the fractional part, then*

$$\int_{[0,1]^d} g(\mathbf{x})d\mathbf{x} = \lim_{T\to\infty} \frac{1}{2T}\int_{-T}^{T} g\big(x_1(t),\dots,x_d(t)\big)dt. \tag{5.24}$$

In particular, for any $\mathbf{k}\in\mathbb{Z}^d$ and $g : \mathbf{x} \mapsto f(\mathbf{x})\exp\big(-2i\pi\mathbf{k}\cdot\mathbf{x}\big)$, (5.24) reads

$$\mathrm{c}_{\mathbf{k}}(f) = \lim_{T\to\infty} \frac{1}{2T}\int_{-T}^{T} f\circ\mathbf{x}(t)\exp\big(-2i\pi(\mathbf{k}\cdot\boldsymbol{\omega})t\big)dt. \tag{5.25}$$

Then FAST consists in replacing $x_i(t) = \{\omega_i t\}$ with semiparametric functions $x_i(t) = G_i\big(\sin(\omega_i t)\big)$ (see [CFS+73]) where the ω_i's are positive integers and the transformations G_i are chosen to preserve the marginal distributions of the X_i's. If the latter are uniformly distributed — as in the present paper —, it can be shown (see [CLS78] and [STC99]) that $G_i(\cdot) = \frac{1}{\pi}\arcsin(\cdot) + \frac{1}{2}$. Saltelli et al. [STC99] also propose to add a random phase-shift $\varphi_i \in [0, 2\pi)$, getting the semiparametric functions $x_i^*(t) = \frac{1}{\pi}\arcsin\big(\sin(2\pi\omega_i t + \varphi_i)\big) + \frac{1}{2}$. Hence, replacing $\mathbf{x}$ with $\mathbf{x}^*$ in (5.25) gives

$$\mathrm{c}_{\mathbf{k}}(f) \approx \lim_{T\to\infty} \frac{1}{2T}\int_{-T}^{T} f\circ\mathbf{x}^*(t)\exp\big(-2i\pi(\mathbf{k}\cdot\boldsymbol{\omega})t\big)dt.$$

Thus, since the functions x_i^* are 1-periodic, it comes

$$\mathrm{c}_{\mathbf{k}}(f) \approx \int_{0}^{1} f\circ\mathbf{x}^*(t)\exp\big(-2i\pi(\mathbf{k}\cdot\boldsymbol{\omega})t\big)dt$$

and applying the rectangle rule to the right-hand side integral gives

$$\mathrm{c}_{\mathbf{k}}(f) \approx \widehat{\mathrm{c}}_{\mathbf{k}\cdot\boldsymbol{\omega}}(f\circ\mathbf{x}^*). \tag{5.26}$$

where

$$\widehat{\mathrm{c}}_{\mathbf{k}\cdot\boldsymbol{\omega}}(f\circ\mathbf{x}^*) = \frac{1}{n}\sum_{j=0}^{n-1} f\circ\mathbf{x}^*\Big(\frac{j}{n}\Big)\exp\Big(-2i\pi j\frac{\mathbf{k}\cdot\boldsymbol{\omega}}{n}\Big)$$

is the complex discrete Fourier coefficient of the one-dimensional function $f\circ\mathbf{x}^*$. In the sequel, the dependence on n, $\boldsymbol{\omega}$ and $\boldsymbol{\varphi}$ is generally omitted for convenience.

Estimation

The estimators of $\mathrm{V}_{\mathrm{u}}(f)$, $\mathrm{V}(f)$ and consequently of $\mathrm{S}_{\mathrm{u}}(f)$ were introduced by using the approximation in (5.26) (see [CFS+73] and Appendix C in [CLS78]). On the one hand, for any non-empty subset $\mathrm{u}\subseteq\{1,\dots,d\}$ and any finite subset $K_{\mathrm{u}}\subseteq\mathbb{Z}^*_{\mathrm{u}}$, (5.26) leads to the definition of the estimator of $\mathrm{V}_{\mathrm{u}}(f)$

$$\widehat{\mathrm{V}}_{\mathrm{u}}^{\mathrm{FAST}}(f, K_{\mathrm{u}}, \mathbf{x}^*) = \sum_{\mathbf{k}\in K_{\mathrm{u}}} \big|\widehat{\mathrm{c}}_{\mathbf{k}\cdot\boldsymbol{\omega}}(f\circ\mathbf{x}^*)\big|^2. \tag{5.27}$$

On the other hand, (5.26) gives

$$\begin{aligned}\mathrm{V}(f) &= \mathrm{c}_{\mathbf{0}}(f^2) - \mathrm{c}_{\mathbf{0}}(f)^2\\ &\approx \widehat{\mathrm{c}}_0(f^2\circ\mathbf{x}^*) - \widehat{\mathrm{c}}_0(f\circ\mathbf{x}^*)^2\end{aligned}$$

and Parseval's identity leads to the definition of the estimator of $\mathrm{V}(f)$

$$\widehat{\mathrm{V}}^{\mathrm{FAST}}(f,\mathbf{x}^*) = \sum_{k=1}^{n-1} \left|\widehat{\mathrm{c}}_k(f\circ\mathbf{x}^*)\right|^2.$$

This naturally leads to the estimator of the variance-based sensitivity indices $\mathrm{S}_\mathrm{u}(f)$

$$\widehat{\mathrm{S}}_\mathrm{u}^{\mathrm{FAST}}(f,K_\mathrm{u},\mathbf{x}^*) = \frac{\displaystyle\sum_{\mathbf{k}\in K_\mathrm{u}} \left|\widehat{\mathrm{c}}_{\mathbf{k}\cdot\boldsymbol{\omega}}(f\circ\mathbf{x}^*)\right|^2}{\displaystyle\sum_{k=1}^{n-1} \left|\widehat{\mathrm{c}}_k(f\circ\mathbf{x}^*)\right|^2}.$$

As in Example 5.2, note that by Parseval's identity $\widehat{\mathrm{V}}^{\mathrm{FAST}}(f,\mathbf{x}^*)$ is equal to the empirical variance $\widehat{\mathrm{V}}(f,\{\mathbf{x}^*(\frac{j}{n})\}_{j=0..n-1})$.

Choice of parameters $\boldsymbol{\omega}$ and n

As discussed by Schaibly and Shuler [SS73] and Cukier et al. [CSS75], $\boldsymbol{\omega}$ and n should be correctly chosen so as to minimize the cubature error in the approximation in (5.26). In order to avoid interferences i.e.

$$\mathbf{k}\cdot\boldsymbol{\omega} - \mathbf{k}'\cdot\boldsymbol{\omega} = 0 \quad \text{for } \mathbf{k},\,\mathbf{k}'\in\mathbb{Z}^d,\,\mathbf{k}\neq\mathbf{k}'$$

and aliasing i.e.

$$\mathbf{k}\cdot\boldsymbol{\omega} - \mathbf{k}'\cdot\boldsymbol{\omega} = jn \quad \text{for } \mathbf{k},\,\mathbf{k}'\in\mathbb{Z}^d,\,\mathbf{k}\neq\mathbf{k}' \text{ and } j\in\mathbb{Z}^*$$

— that both lead to $\widehat{\mathrm{c}}_{\mathbf{k}\cdot\boldsymbol{\omega}}(f\circ\mathbf{x}^*) = \widehat{\mathrm{c}}_{\mathbf{k}'\cdot\boldsymbol{\omega}}(f\circ\mathbf{x}^*)$ — Schaibly and Shuler [SS73] propose to choose ω_1, $\ldots$, ω_d free of interferences up to order $N\in\mathbb{N}^*$:

$$(\mathbf{k}-\mathbf{k}')\cdot\boldsymbol{\omega}\neq 0 \quad \text{for all } \mathbf{k},\ \mathbf{k}'\in\mathbb{Z}^d,\ \mathbf{k}\neq\mathbf{k}', \text{ s.t. } \sum_{i=1}^{d}|k_i-k_i'|\leq N+1 \tag{5.28}$$

and n sufficiently large

$$n\approx N\max(\omega_1,\ldots,\omega_d). \tag{5.29}$$

More recently, referring to the classical information theory, Saltelli et al. [STC99] suggest to replace (5.29) with Nyquist-Shannon sampling theorem (see e.g. [Mar09])

$$n > 2N\max(\omega_1,\ldots,\omega_d). \tag{5.30}$$

In our opinion, the criterion stated in (13) should be written

$$(\mathbf{k}-\mathbf{k}')\cdot\boldsymbol{\omega}\neq 0 \quad \text{for all } \mathbf{k},\ \mathbf{k}'\in\mathbb{Z}^d,\ \mathbf{k}\neq\mathbf{k}', \text{ s.t. } \sum_{i=1}^{d}|k_i|\leq N' \text{ and } \sum_{i=1}^{d}|k_i'|\leq N' \tag{5.31}$$

since the main objective is to avoid interferences within a finite subset of $\mathbb{Z}^d$ out of which the Fourier coefficients of f are a priori negligible — in (5.31), this subset is the closed l^1-norm ball of radius N'. Thus we may reformulate the whole criterion stated in (5.28) and (5.30) with respect to the set $K=\sqcup_\mathrm{u}K_\mathrm{u}$ where the K_u's are the truncation sets in the FAST estimator of $\mathrm{V}_\mathrm{u}(f)$ given in (5.27). We propose to choose ω_1, $\ldots$, ω_d free of interferences within K i.e.

$$(\mathbf{k}-\mathbf{k}')\cdot\boldsymbol{\omega}\neq 0 \ \text{ for all } \mathbf{k},\ \mathbf{k}'\in K,\ \mathbf{k}\neq\mathbf{k}' \quad \text{and} \quad n > \max_{\mathbf{k},\mathbf{k}'\in K}\big((\mathbf{k}-\mathbf{k}')\cdot\boldsymbol{\omega}\big). \tag{5.32}$$

In the sequel, we refer to the latter as the "classic" criterion of FAST.

5.3.2 Review of RBD

RBD makes use of the previous framework setting $\boldsymbol{\varphi} = \mathbf{0}$, $\omega_1 = \cdots = \omega_d = \omega \in \mathbb{N}^*$ — usually set to 1 — and applying random permutations on the coordinates of the resulting points $\mathbf{x}^*(\frac{j}{n})$. More precisely, let $\sigma_1, \ldots, \sigma_d$ be random permutations on $\{0, \ldots, n-1\}$ and $\mathfrak{S}$ denote the set of all possible $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_d)$. Given $\boldsymbol{\sigma} \in \mathfrak{S}$, consider the function $\mathbf{x}^\times = (x_1^\times, \ldots, x_d^\times)$ defined on $\{0, \frac{1}{n}, \ldots, \frac{n-1}{n}\}$ such that for all $i \in \{1, \ldots, d\}$ and $j \in \{0, \ldots, n-1\}$,

$$x_i^\times\Big(\frac{j}{n}\Big) = \frac{1}{\pi}\arcsin\Big(\sin\Big(2\pi\omega\frac{\sigma_i(j)}{n}\Big)\Big) + \frac{1}{2}.$$

Thus denoting σ_i^{-1} the inverse permutation of σ_i, define

$$\mathbf{x}^{\times,i}\Big(\frac{j}{n}\Big) = \mathbf{x}^\times\Big(\frac{\sigma_i^{-1}(j)}{n}\Big).$$

Finally through a heuristic argument Tarantola et al. [TGM06] introduce the RBD estimators of $\mathrm{V}_{\mathfrak{u}}(f)$, $\mathrm{V}(f)$ and $\mathrm{S}_{\mathfrak{u}}(f)$ for first-order terms — i.e. $\mathfrak{u} = \{i\}$, $i \in \{1, \ldots, d\}$ —. For any finite subset $K_{\{i\}} \subseteq \mathbb{Z}^*_{\{i\}}$, we have

$$\widehat{\mathrm{V}}_i^{\mathrm{RBD}}(f, K_{\{i\}}, \mathbf{x}^\times) = \sum_{\mathbf{k}\in K_{\{i\}}} \big|\widehat{\mathrm{c}}_{k_i\omega}(f\circ\mathbf{x}^{\times,i})\big|^2,$$

$$\widehat{\mathrm{V}}^{\mathrm{RBD}}(f, \mathbf{x}^\times) = \sum_{k=1}^{n-1} \big|\widehat{\mathrm{c}}_k(f\circ\mathbf{x}^\times)\big|^2$$

and

$$\widehat{\mathrm{S}}_i^{\mathrm{RBD}}(f, K_{\{i\}}, \mathbf{x}^\times) = \frac{\displaystyle\sum_{\mathbf{k}\in K_{\{i\}}} \big|\widehat{\mathrm{c}}_{k_i\omega}(f\circ\mathbf{x}^{\times,i})\big|^2}{\displaystyle\sum_{k=1}^{n-1} \big|\widehat{\mathrm{c}}_k(f\circ\mathbf{x}^\times)\big|^2}.$$

As in FAST note that by Parseval's identity, the estimator $\widehat{\mathrm{V}}^{\mathrm{RBD}}(f, \mathbf{x}^\times)$ is equal to the empirical variance $\widehat{\mathrm{V}}(f, \{\mathbf{x}^\times(\frac{j}{n})\}_{j=0..n-1})$. In the sequel, the dependence on ω and $\boldsymbol{\sigma}$ is generally omitted for convenience.

5.3.3 FAST and RBD revisited

Main result

First we introduce more notation. For any $p \in \mathbb{N}^*$, let

$$r_p : \begin{array}{ccc} [0,1] & \longrightarrow & [0,1] \\ x & \longmapsto & \begin{cases} 2\{px\} & \text{if } 0 \le \{px\} < \frac{1}{2} \\ 2 - 2\{px\} & \text{if } \frac{1}{2} \le \{px\} \le 1 \end{cases} \end{array}$$

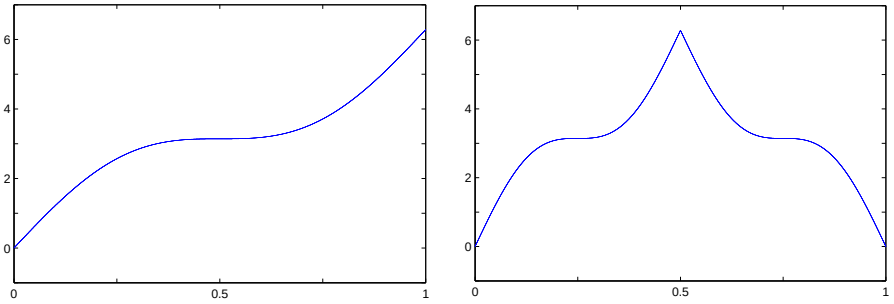
and for any $\varphi \in [0, 2\pi)$

$$t_\varphi : \begin{array}{ccc} [0,1] & \longrightarrow & [0,1] \\ x & \longmapsto & \{x + \tilde{\varphi}\} \quad \text{with } \tilde{\varphi} = \frac{1}{4} + \frac{\varphi}{2\pi}\ . \end{array}$$

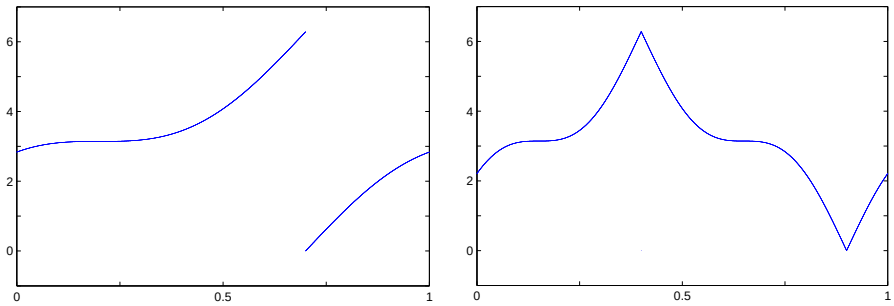
Then we define the linear operators $\mathcal{R}_p$ and $\mathcal{T}_{\boldsymbol{\varphi}}$ (see Figure 5.1) on $L^2([0,1]^d)$ such that for all $\mathbf{x} \in [0,1]^d$,

$$\mathcal{R}_p f(\mathbf{x}) = f\big(r_p(x_1)\ldots, r_p(x_d)\big) \quad \text{et} \quad \mathcal{T}_{\boldsymbol{\varphi}} f(\mathbf{x}) = f\big(t_{\varphi_1}(x_1), \ldots, t_{\varphi_d}(x_d)\big).$$

and note that $\mathcal{R}_p = \underbrace{\mathcal{R}_1 \circ \cdots \circ \mathcal{R}_1}_{p \text{ times}}$. We also introduce two classical designs of experiments. For any



(a) Plot of $f : x \mapsto x + \sin(x)$ (b) Plot of $\mathcal{R}_1 f$



(c) Plot of $\mathcal{T}_{\frac{\pi}{10}} f$ (d) Plot of $(\mathcal{T}_{\frac{\pi}{30}} \circ \mathcal{R}_1) f$

FIGURE 5.1 – Examples of operators $\mathcal{R}_p$ and $\mathcal{T}_{\boldsymbol{\varphi}}$ in dimension 1.

$\boldsymbol{\omega} \in (\mathbb{N}^*)^d$, we denote

$$G(\boldsymbol{\omega}) = \left\{ \left(\left\{ \frac{j}{n} \omega_1 \right\}, \ldots, \left\{ \frac{j}{n} \omega_d \right\} \right),\ j \in \{0, \ldots, n-1\} \right\}$$

the cyclic subgroup — of order $n/gcd(\omega_1, \ldots, \omega_d, n)$ — of the torus $\mathbb{T}^d = (\mathbb{R}/\mathbb{Z})^d \simeq [0,1)^d$ generated by $(\{\frac{\omega_1}{n}\}, \ldots, \{\frac{\omega_d}{n}\})$ (see e.g. [Hal59]). For any $\boldsymbol{\sigma} \in \mathfrak{S}$ we also denote

$$A(\boldsymbol{\sigma}) = \left\{ \left(\frac{\sigma_1(j)}{n}, \ldots, \frac{\sigma_d(j)}{n} \right),\ j \in \{0, \ldots, n-1\} \right\}$$

the orthogonal array of strength 1 and index unity with elements taken from $\{0, \frac{1}{n}, \ldots, \frac{n-1}{n}\}$ and based on the permutation σ (see e.g. [HSS99]). FAST and RBD methods are now introduced in a new way by using the basic estimator in (5.15).

Proposition 5.2. *Let $f : [0,1]^d \to \mathbb{R}$ be a square-integrable function. For any non-empty subset $\mathfrak{u} \subseteq \{1, \ldots, d\}$, any finite subset $K_{\mathfrak{u}} \subseteq \mathbb{Z}^*_{\mathfrak{u}}$, $\boldsymbol{\varphi} \in [0, 2\pi)^d$ and $\boldsymbol{\omega} \in (\mathbb{N}^*)^d$, we have*

$$\widehat{\mathrm{S}}^{\mathrm{FAST}}_{\mathfrak{u}}(f, K_{\mathfrak{u}}, \mathbf{x}^*) = \widehat{\mathrm{S}}_{\mathfrak{u}}\big((\mathcal{T}_{\boldsymbol{\varphi}} \circ \mathcal{R}_1) f, K_{\mathfrak{u}}, G(\boldsymbol{\omega})\big). \tag{5.33}$$

For any $i \in \{1, \ldots, d\}$, any finite subset $K_{\{i\}} \subseteq \mathbb{Z}^*_{\{i\}}$, $\boldsymbol{\sigma} \in \mathfrak{S}$ and $\omega \in \mathbb{N}^*$, we have

$$\widehat{\mathrm{S}}^{\mathrm{RBD}}_{i}(f, K_{\{i\}}, \mathbf{x}^{\times}) = \widehat{\mathrm{S}}_{i}\big((\mathcal{T}_{\tilde{\boldsymbol{\omega}}} \circ \mathcal{R}_{\omega}) f, \omega K_{\{i\}}, A(\boldsymbol{\sigma})\big) \tag{5.34}$$

where $\tilde{\boldsymbol{\omega}} = \left(\frac{(1-\omega)\pi}{2\omega}, \cdots, \frac{(1-\omega)\pi}{2\omega} \right)$ and $\omega K_{\{i\}} = \{(\omega k_1, \ldots, \omega k_d),\ \mathbf{k} \in K_{\{i\}}\}$.

Démonstration. It essentially consists in showing that for all $j \in \{0, \ldots n-1\}$

$$f \circ \mathbf{x}^*\left(\frac{j}{n}\right) = (\mathcal{T}_{\boldsymbol{\varphi}} \circ \mathcal{R}_1) f\left(\left\{ \frac{j}{n} \omega_1 \right\}, \ldots, \left\{ \frac{j}{n} \omega_d \right\} \right)$$

and

$$f \circ \mathbf{x}^{\times}\Big(\frac{j}{n}\Big) = (\mathcal{T}_{\tilde{\boldsymbol{\omega}}} \circ \mathcal{R}_{\omega}) f\Big(\frac{\sigma_1(j)}{n}, \ldots, \frac{\sigma_d(j)}{n}\Big).$$

See details in Appendix 5.A. □

Remark 5.1. *In the RBD method, the parameter ω is usually set to 1 but its role is not well understood up to now. In our opinion there is no reason to set $\omega \neq 1$ since if $gcd(\omega, n) = 1$ then it leads to the case $\omega = 1$, and otherwise the estimator in (5.34) is potentially less efficient than in the case $\omega = 1$ (see details in Appendix 5.B).*

What FAST and RBD are

It is clear from Proposition 5.2 that FAST and RBD only consist in applying the basic estimator introduced in (5.15) to a particular transform $(\mathcal{T}_{\boldsymbol{\varphi}} \circ \mathcal{R}_p)f$ of the function f and a particular design of experiments $G(\boldsymbol{\omega})$ or $A(\boldsymbol{\sigma})$. Now it is also clear that the basic estimator generates an error term due to truncations — in (5.13) — and an other one due to numerical integrations — in (5.12) and (5.14). Moreover, the use of $(\mathcal{T}_{\boldsymbol{\varphi}} \circ \mathcal{R}_p)f$ instead of f could also have an impact on the sensitivity indices estimation error. We now investigate this latter issue by introducing the notion of invariance of the variance decomposition.

Definition 5.1. *Let $\mathcal{L}$ be a linear operator on $L^2([0,1]^d)$. The variance decomposition is said to be $\mathcal{L}$-invariant on $L^2([0,1]^d)$ if for any non-empty set $\mathfrak{u} \subseteq \{1, \ldots, d\}$ and any function $f \in L^2([0,1]^d)$ we have*

$$\mathrm{V}_{\mathfrak{u}}(\mathcal{L}f) = \mathrm{V}_{\mathfrak{u}}(f).$$

This leads to the following result

Lemma 5.1. *For any $p \in \mathbb{N}^*$ and any $\boldsymbol{\varphi} \in [0, 2\pi)^d$, the variance decomposition is $\mathcal{R}_p$ and $\mathcal{T}_{\boldsymbol{\varphi}}$-invariant on $L^2([0,1]^d)$.*

Démonstration. See Appendix 5.C □

As a consequence, for any non-empty subset $\mathfrak{u} \subseteq \{1, \ldots, d\}$, we have

$$\mathrm{S}_{\mathfrak{u}}\big((\mathcal{T}_{\boldsymbol{\varphi}} \circ \mathcal{R}_p)f\big) = \mathrm{S}_{\mathfrak{u}}(f)$$

and this asserts the validity of FAST and RBD methods. Note that the linear operator $\mathcal{R}_p$ "regularizes" the function f in the sense that if $\mathbf{x} \mapsto f(\mathbf{x})$ is continuous on $[0,1]^d$ and $\mathbf{x} \to f(\{x_1\}, \ldots, \{x_d\})$ is discontinuous on $\mathbb{R}^d$ then $\mathbf{x} \to \mathcal{R}_p f(\{x_1\}, \ldots, \{x_d\})$ is continuous on $\mathbb{R}^d$. This is an important property since by Riemann-Lebesgue lemma $|\mathrm{c}_{\mathbf{k}}(f)|$ converges to 0 as $||\mathbf{k}||$ tends to ∞, and the smoother the function f, the faster the convergence (see e.g. [Zar68]). The other operator $\mathcal{T}_{\boldsymbol{\varphi}}$ essentially allows to define randomized estimators in FAST.

Potential generalizations

To end with, we list three natural generalizations that are further discussed in the next section :
- the estimator $\widehat{\mathrm{S}}_{\mathfrak{u}}\big((\mathcal{T}_{\boldsymbol{\varphi}} \circ \mathcal{R}_1)f, K_{\mathfrak{u}}, G(\boldsymbol{\omega})\big)$ can also be defined for a group G of any rank $r \leq d$
- the estimator $\widehat{\mathrm{S}}_i\big((\mathcal{T}_{\tilde{\boldsymbol{\omega}}} \circ \mathcal{R}_{\omega})f, \omega K_{\{i\}}, A(\boldsymbol{\sigma})\big)$ can also be defined for a sensitivity index of any order : $\widehat{\mathrm{S}}_{\mathfrak{u}}\big((\mathcal{T}_{\tilde{\boldsymbol{\omega}}} \circ \mathcal{R}_{\omega})f, \omega K_{\mathfrak{u}}, A(\boldsymbol{\sigma})\big)$, note that it has been already applied in [XG11b]
- the latter estimator $\widehat{\mathrm{S}}_{\mathfrak{u}}\big((\mathcal{T}_{\tilde{\boldsymbol{\omega}}} \circ \mathcal{R}_{\omega})f, \omega K_{\mathfrak{u}}, A(\boldsymbol{\sigma})\big)$ can also be defined for an orthogonal array A having any parameters.

5.4 Error analysis

For convenience, operators $\mathcal{T}_{\boldsymbol{\varphi}}$ and $\mathcal{R}_p$ are now omitted. Moreover, we assume that the function f has an absolutely convergent Fourier representation, i.e. $\sum_{\mathbf{k} \in \mathbb{Z}^d} |\mathrm{c}_{\mathbf{k}}(f)| < +\infty$.

5.4.1 Cubature error in FAST

Two points of view

In this section we mainly focus on the error term

$$e_{\mathbf{k}}(f,G) = \widehat{\mathrm{c}}_{\mathbf{k}}(f,G) - \mathrm{c}_{\mathbf{k}}(f) \tag{5.35}$$

where G is a subgroup of $\mathbb{T}^d$ of order n and $\mathbf{k} \in \mathbb{Z}^d$. By its definition, the term $\widehat{\mathrm{c}}_{\mathbf{k}}(f,G)$ consists of an equal weight cubature rule at the n nodes of the group G, also known as a lattice rule (see [SJ94] for a survey). Moreover by the generalized Poisson summation formula (see e.g. [Loo53]), the error term in (5.35) is precisely

$$e_{\mathbf{k}}(f,G) = \sum_{\mathbf{h}\in G^\perp\setminus\{\mathbf{0}\}} \mathrm{c}_{\mathbf{k}+\mathbf{h}}(f) \tag{5.36}$$

where $G^\perp = \{\mathbf{h} \in \mathbb{Z}^d \mid \forall \mathbf{x} \in G,\ \mathbf{h}\cdot\mathbf{x} \equiv 0 \pmod 1\}$ is the subgroup of $\mathbb{Z}^d$ orthogonal to G, also known as the dual lattice of G.
In the lattice rules field, $e_{\mathbf{0}}(f,G)$ is the only term of interest, and there exist two main points of view to control it. One consists in looking for "good" groups G such that the cubature rule is exact for a set of trigonometric polynomials, i.e. for a finite subset K of $\mathbb{Z}^d$,

$$e_{\mathbf{0}}(f,G) = 0 \ \text{ for all } \ f \text{ such that } \forall \mathbf{k} \notin K,\ \mathrm{c}_{\mathbf{k}}(f) = 0\ .$$

The other point of view aims to find "good" groups G such that the cubature rule has an absolute error $|e_{\mathbf{0}}(f,G)|$ dominated by an explicit bound for all f in a particular space of smooth functions. Note that these approaches are compatible to each other (see e.g. [CKN10] and the references therein).

Now concerning the study of error in FAST, the first point of view, which essentially corresponds to the classic FAST, consists of a trigonometric interpolation issue and leads to a metamodel approach of the estimation of the sensitivity indices. The second one, which is more original, allows to derive error bounds for $\widehat{\mathrm{V}}_{\mathrm{u}}(f,K_{\mathrm{u}},G)$ and $\widehat{\mathrm{V}}(f,G)$ in spaces of smooth functions. Both these methods are discussed below.

Metamodel approach

Let K be a finite subset of $\mathbb{Z}^d$. Then an immediate consequence of (5.36) is that a group G satisfies the property

$$e_{\mathbf{k}}(f,G) = 0 \ \text{ for all } \mathbf{k}\in K \ \text{ and for all } f \text{ such that } \forall \mathbf{k} \notin K,\ \mathrm{c}_{\mathbf{k}}(f) = 0$$

if and only if

$$\forall \mathbf{k},\mathbf{k}' \in K, \mathbf{k}\neq\mathbf{k}',\ \exists \mathbf{x}\in G,\ (\mathbf{k}-\mathbf{k}')\cdot\mathbf{x} \not\equiv 0 \pmod 1\ . \tag{5.37}$$

More fundamentally, for any $E \subseteq \mathbb{Z}^d$, consider the trigonometric polynomial

$$\tilde{f}_E(\mathbf{x}) = \sum_{\mathbf{k}\in E} \widehat{\mathrm{c}}_{\mathbf{k}}(f,G)\exp(2i\pi\mathbf{k}\cdot\mathbf{x})\ , \tag{5.38}$$

then the equivalence above leads to the following result

Proposition 5.3. *Let G be a subgroup of the torus $\mathbb{T}^d$ of order $|G| = n$ and $K = \cup_{\mathrm{u}\neq\emptyset} K_{\mathrm{u}}$ satisfying the criterion (5.37) where for all non-empty subsets u of $\{1,\dots,d\}$, $K_{\mathrm{u}} \subseteq \mathbb{Z}^*_{\mathrm{u}}$*

i) *if $|K| = n$, then $\tilde{f}_K$ is a trigonometric interpolation polynomial of f at the n nodes $\mathbf{x}\in G$ and we have*

$$\widehat{\mathrm{S}}_{\mathrm{u}}(f,K_{\mathrm{u}},G) = \mathrm{S}_{\mathrm{u}}(\tilde{f}_K).$$

ii) *if $|K| < n$, let H be any subset of $\mathbb{Z}^d$ such that $K \subseteq H$, H satisfies the criterion (5.37) and $|H| = n$. Then $\tilde{f}_H$ is a trigonometric interpolation polynomial of f at the n nodes $\mathbf{x}\in G$ and we have*

$$\widehat{\mathrm{V}}_{\mathrm{u}}(f,K_{\mathrm{u}},G) = \mathrm{V}_{\mathrm{u}}(\tilde{f}_K) \text{ and } \widehat{\mathrm{V}}(f,G) = \mathrm{V}(\tilde{f}_H).$$

Démonstration. The only difficulty is to prove that the trigonometric polynomials $\tilde{f}_K$ in the assertion i) and $\tilde{f}_H$ in the assertion ii) are interpolation polynomials at the points $\mathbf{x} \in G$. We demonstrate it for $\tilde{f}_K$, the proof for $\tilde{f}_H$ is exactly the same.
Since the function f has absolutely convergent Fourier representation, we can write

$$f(\mathbf{x}) = \sum_{\mathbf{k}\in\mathbb{Z}^d} \mathrm{c}_{\mathbf{k}}(f)\exp(2i\pi\mathbf{k}\cdot\mathbf{x}) = \sum_{\mathbf{k}\in K}\sum_{\mathbf{h}\in G^\perp} \mathrm{c}_{\mathbf{k}+\mathbf{h}}(f)\exp\big(2i\pi(\mathbf{k}+\mathbf{h})\cdot\mathbf{x}\big) \tag{5.39}$$

(see details in Appendix 5.D) and by definition of $G^\perp$, we have that for any $\mathbf{x} \in G$,

$$f(\mathbf{x}) = \sum_{\mathbf{k}\in K}\sum_{\mathbf{h}\in G^\perp} \mathrm{c}_{\mathbf{k}+\mathbf{h}}(f)\exp(2i\pi\mathbf{k}\cdot\mathbf{x}).$$

The conclusion follows from the definition in (5.38) since (5.35) and (5.36) give

$$\sum_{\mathbf{h}\in G^\perp} \mathrm{c}_{\mathbf{k}+\mathbf{h}}(f) = \widehat{\mathrm{c}}_{\mathbf{k}}(f, G).$$

□

From this point of view, FAST returns analytical values from trigonometric metamodels of the function $(\mathcal{T}_{\boldsymbol{\varphi}} \circ \mathcal{R}_1)f$ and the error analysis should be performed on the metamodel itself.

In practice, a set of a priori non-negligible frequencies $K = \cup_{\mathrm{u}\neq\emptyset} K_{\mathrm{u}}$ is given and a group G satisfying the criterion (5.37) and with the smallest order $|G| = n$ has to be found. Searching for this group G is computationaly expensive and may rapidly become unfeasible. One of the cheapest way is to look for cyclic groups $G = G(\boldsymbol{\omega})$, coming back to the classic FAST. In this case, the criterion (5.37) simply reads

$$\forall \mathbf{k}, \mathbf{k}' \in K, \mathbf{k} \neq \mathbf{k}',\ (\mathbf{k}-\mathbf{k}')\cdot\boldsymbol{\omega} \not\equiv 0 \pmod{n}\ . \tag{5.40}$$

Note that this new criterion plays the same role as the classic criterion of FAST given in (5.32). The main difference between these two approaches is that optimization on n is performed in (5.40), consequently this new criterion allows to find group G with smaller order n. We illustrate the efficiency of both criterions by using basic exhaustive algorithms with computational complexity $O(n^d)$. The results are gathered in Table 5.1 and show that the new criterion leads to a non-negligible improvement.

Remark 5.2. *Even if cyclic groups seem to be suitable in the previous issue, the computational cost of the research of a generator $\boldsymbol{\omega}$ can become prohibitive in high-dimensional problems. In this case, alternative algorithms can be used instead of a systematic research technique (for a recent reference, see e.g.* [KKP11]).

Error bounds

Searching for a finite subgroup G of the torus $\mathbb{T}^d$ such that $e_{\mathbf{0}}(f, G)$ has an explicit bound in a particular function space is a problem known as the construction of good lattice rules (for a survey see [SJ94] or more recently [Nuy07]). Most of the results in this field are established in Korobov spaces which are suitable to handle lattice methods ; so we derive error bounds for sensitivity indices in these spaces. For $\alpha > 1$ and $\boldsymbol{\gamma} = (\gamma_{\mathrm{u}})_{\mathrm{u}\subseteq\{1,\dots,d\}}$ with non-negative γ_{u}'s, define the weighted Korobov space $\mathcal{H}_{\alpha,\boldsymbol{\gamma}}$ to be the Hilbert space with reproducing kernel

$$RK_{\alpha,\boldsymbol{\gamma}}(\mathbf{x},\mathbf{y}) = 1 + \sum_{\mathbf{k}\in(\mathbb{Z}^d)^*} r(\mathbf{k},\alpha,\boldsymbol{\gamma})^{-1}\exp\big(2i\pi\mathbf{k}\cdot(\mathbf{x}-\mathbf{y})\big)$$

where for any $\mathbf{k} \neq \mathbf{0}$, $r(\mathbf{k},\alpha,\boldsymbol{\gamma}) = \gamma_{\mathrm{u}_{\mathbf{k}}}^{-1}\prod_{i\in\mathrm{u}_{\mathbf{k}}}|k_i|^\alpha$, where $\mathrm{u}_{\mathbf{k}}$ is such that $\mathbf{k} \in \mathbb{Z}^*_{\mathrm{u}_{\mathbf{k}}}$. For $\mathbf{k}$ such that $\gamma_{\mathrm{u}_{\mathbf{k}}} = 0$, we set by convention $r(\mathbf{k},\alpha,\boldsymbol{\gamma}) = \infty$. Thus the kernel can be rewritten

$$RK_{\alpha,\boldsymbol{\gamma}}(\mathbf{x},\mathbf{y}) = 1 + \sum_{\substack{\mathbf{k}\in(\mathbb{Z}^d)^* \\ \gamma_{\mathrm{u}_{\mathbf{k}}}\neq 0}} r(\mathbf{k},\alpha,\boldsymbol{\gamma})^{-1}\exp\big(2i\pi\mathbf{k}\cdot(\mathbf{x}-\mathbf{y})\big)$$

		$d=2$			$d=3$			$d=4$			$d=5$		
$N1$	N_2	$\lvert K\rvert$	n_{old}	n_{new}	$\lvert K\rvert$	n_{old}	n_{new}	$\lvert K\rvert$	n_{old}	n_{new}	$\lvert K\rvert$	n_{old}	n_{new}
4	2	20	41	29	36	65	50	56	105	63	80	177	111
5	3	32	61	48	66	141	102	112	241	173	170	471	302
6	4	48	85	65	108	241	155	192	541	323	300	997	613
7	5	68	113	89	162	421	284	296	1177	586	470	1891	1279
8	6	92	145	120	228	625	429	424	1985	1033	680	3457	2222
9	7	120	181	149	306	937	645	576	3007	1706	930	–	–
10	8	152	221	185	396	1281	933	752	4501	2529	1220	–	–
11	9	188	265	228	498	1805	1284	952	7261	3684	1550	–	–

TABLE 5.1 – Comparison in dimension $d = 2$, 3, 4 and 5 between the minimum sample size n given by the classic criterion of FAST (denoted n_{old}) and the new one proposed in (5.40) (denoted n_{new}). Here, the $K_{\{i\}}$'s are equal to $\mathbb{Z}^*_{\{i\}} \cap \{|k_i| \leq N_1\}$, the $K_{\{i,j\}}$'s are equal to $\mathbb{Z}^*_{\{i,j\}} \cap \{|k_i| + |k_j| \leq N_2\}$ and for all $\mathfrak{u}$ such that $|\mathfrak{u}| > 2$, $K_{\mathfrak{u}} = \emptyset$. Such sets K are particularly well-suited to analyse functions whose effective dimension is less than 2 — see Definition 5.4 in Section 5.4.2.

and we deduce that the norm of $f \in \mathcal{H}_{\alpha,\boldsymbol{\gamma}}$ satisfies

$$||f||^2_{\mathcal{H}_{\alpha,\boldsymbol{\gamma}}} = \mathrm{c}_{\mathbf{0}}(f)^2 + \sum_{\substack{\mathbf{k}\in(\mathbb{Z}^d)^* \\ \gamma_{\mathfrak{u}_\mathbf{k}}\neq 0}} r(\mathbf{k},\alpha,\boldsymbol{\gamma})|\mathrm{c}_{\mathbf{k}}(f)|^2 < +\infty$$

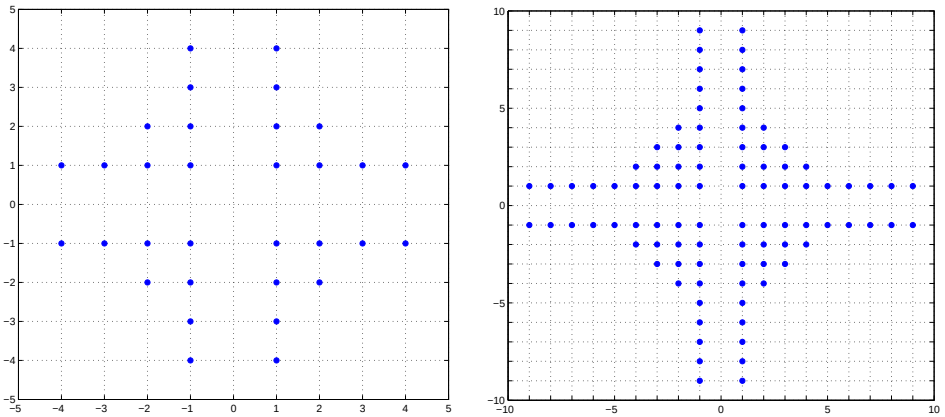
and consequently

$$\forall \mathbf{k} \in (\mathbb{Z}^d)^* \textit{ such that } \gamma_{\mathfrak{u}_\mathbf{k}} \neq 0,\ |\mathrm{c}_{\mathbf{k}}(f)|^2 \leq \frac{\gamma_{\mathfrak{u}_\mathbf{k}}||f||^2_{\mathcal{H}_{\alpha,\boldsymbol{\gamma}}}}{\prod\limits_{i\in\mathfrak{u}_\mathbf{k}} |k_i|^\alpha}.$$

Note that for any $\mathbf{k} \in (\mathbb{Z}^d)^*$ such that $\gamma_{\mathfrak{u}_\mathbf{k}} = 0$, $f \in \mathcal{H}_{\alpha,\boldsymbol{\gamma}}$ implies $\mathrm{c}_{\mathbf{k}}(f) = 0$. We also make a restriction on the sets of frequencies $K_{\mathfrak{u}}$'s. Here we assume that for any non-empty set $\mathfrak{u} \subseteq \{1,\dots,d\}$, $K_{\mathfrak{u}}$ is of Zaremba cross-type (see Figure 5.2)

$$K_{\mathfrak{u}} = \mathcal{Z}_{\mathfrak{u},\beta_{\mathfrak{u}}} = \left\{ \mathbf{k} \in \mathbb{Z}^*_{\mathfrak{u}},\ \prod_{i\in\mathfrak{u}} |k_i| \leq \beta_{\mathfrak{u}} \right\} \tag{5.41}$$

where $\beta_{\mathfrak{u}} \geq 1$. This kind of sparse grids is particularly well-suited for the analysis of high-dimensional smooth functions. We now give the result on error bounds for $\widehat{\mathrm{V}}_{\mathfrak{u}}(f, K_{\mathfrak{u}}, G)$ and $\widehat{\mathrm{V}}(f, G)$ in $\mathcal{H}_\alpha$.



(a) Plot of $\mathcal{Z}_{\{1,2\},4}$ (b) Plot of $\mathcal{Z}_{\{1,3\},9}$

FIGURE 5.2 – Illustration of crosses $\mathcal{Z}_{\mathfrak{u},\beta_{\mathfrak{u}}}$.

Proposition 5.4. *Let $f \in \mathcal{H}_{\alpha,\boldsymbol{\gamma}}$ with $\alpha > 2$ and $\boldsymbol{\gamma} = (\gamma_{\mathfrak{u}})_{\mathfrak{u}\subseteq\{1,\dots,d\}}$ with non-negative components. Let G be a subgroup of $\mathbb{T}^d$ of order n such that the cubature error related to G is dominated by the explicit bound $B(\alpha,n,d,\boldsymbol{\gamma})$ on the unit ball of $\mathcal{H}_{\alpha,\boldsymbol{\gamma}}$ i.e. for all f in $\mathcal{H}_{\alpha,\boldsymbol{\gamma}}$, $|\widehat{\mathrm{c}}_{\mathbf{0}}(f,G) - \mathrm{c}_{\mathbf{0}}(f)| \le B(\alpha,n,d,\boldsymbol{\gamma})||f||_{\mathcal{H}_{\alpha,\boldsymbol{\gamma}}}$. Then*
i) if there exists $\alpha' > 2$ and $\boldsymbol{\gamma}' = (\gamma'_{\mathfrak{u}})_{\mathfrak{u}\subseteq\{1,\dots,d\}}$ with non-negative components such that $f^2 \in \mathcal{H}_{\alpha',\boldsymbol{\gamma}'}$, we have

$$\big|\widehat{\mathrm{V}}(f,G) - \mathrm{V}(f)\big| \le ||f||^2_{\mathcal{H}_\alpha} B(\alpha,n,d,\boldsymbol{\gamma})\big(2 + B(\alpha,n,d,\boldsymbol{\gamma})\big) + ||f^2||_{\mathcal{H}_{\alpha'}} B(\alpha',n,d,\boldsymbol{\gamma}')$$

ii) for any non-empty set $\mathfrak{u} \subseteq \{1,\dots,d\}$ and $K_{\mathfrak{u}} = \mathcal{Z}_{\mathfrak{u},\beta_{\mathfrak{u}}}$, we have

$$\big|\widehat{\mathrm{V}}_{\mathfrak{u}}(f,K_{\mathfrak{u}},G) - \mathrm{V}_{\mathfrak{u}}(f)\big| \le ||f||^2_{\mathcal{H}_{\alpha,\boldsymbol{\gamma}}}\Big[C(\alpha,\boldsymbol{\gamma},\beta_{\mathfrak{u}},|\mathfrak{u}|) + B(\alpha,n,d,\boldsymbol{\gamma})^2 S_1(\alpha,\boldsymbol{\gamma},\beta_{\mathfrak{u}},\mathfrak{u}) + B(\alpha,n,d,\boldsymbol{\gamma}) S_2(\alpha,\boldsymbol{\gamma},\beta_{\mathfrak{u}},\mathfrak{u})\Big]$$

where

$$\begin{aligned} S_1(\alpha,\boldsymbol{\gamma},\beta_{\mathfrak{u}},\mathfrak{u}) &= \gamma_{frac} \sum_{\mathbf{k}\in K_{\mathfrak{u}}} \prod_{i\in\mathfrak{u}} \big(|k_i|+1\big)^{\alpha} \quad , \ \gamma_{frac} = \max_{\substack{\mathfrak{u},\mathfrak{v}\subset\{1,\dots,d\}\\ \gamma_{\mathfrak{v}}\neq 0}} \gamma_{\mathfrak{u}}/\gamma_{\mathfrak{v}} \\ S_2(\alpha,\boldsymbol{\gamma},\beta_{\mathfrak{u}},\mathfrak{u}) &= \gamma_{frac}\gamma_{\mathfrak{u}}^{1/2} 2^{\alpha|\mathfrak{u}|/2} |K_{\mathfrak{u}}| \end{aligned}$$

and for $|\mathfrak{u}| \le 2$, the truncation error term $C(\alpha,\beta_{\mathfrak{u}},|\mathfrak{u}|)$ are

$$C(\alpha,\boldsymbol{\gamma},\beta_{\mathfrak{u}},1) = \frac{2\gamma_{max}\zeta(\alpha)}{\beta_{\mathfrak{u}}^{\alpha-1}} \quad , \ \gamma_{max} = \max_{\mathfrak{u}\subset\{1,\dots,d\}} \gamma_{\mathfrak{u}} \tag{5.42}$$

$$C(\alpha,\boldsymbol{\gamma},\beta_{\mathfrak{u}},2) = \frac{4\gamma_{max}\big[\zeta(\alpha)^2 + \zeta(\alpha)\big(\log(\beta_{\mathfrak{u}}) + 2\big)\big]}{\beta_{\mathfrak{u}}^{\alpha-1}} \ . \tag{5.43}$$

Démonstration. See Appendix 5.E □

It is also possible to derive explicit formulas of the truncation error term for $|\mathfrak{u}| > 2$, but this is more complicated and of second interest. Secondly, it has to be noted that, in the second item of Proposition 5.4, the functions S_1 and S_2 are increasing with respect to the parameter $\beta_{\mathfrak{u}}$ while the function C is decreasing. As a consequence, efficient bounds consist of a trade-off between $\beta_{\mathfrak{u}}$ and n such that $B(\alpha,n,d,\boldsymbol{\gamma})^2 S_1(\alpha,\boldsymbol{\gamma},\beta_{\mathfrak{u}},\mathfrak{u})$, $B(\alpha,n,d,\boldsymbol{\gamma}) S_2(\alpha,\boldsymbol{\gamma},\beta_{\mathfrak{u}},\mathfrak{u})$ and $C(\alpha,\boldsymbol{\gamma},\beta_{\mathfrak{u}},|\mathfrak{u}|)$ have the same order. For example,

i) if $|\mathfrak{u}| = 1$ and $\alpha > 2$, note that $|K_{\mathfrak{u}}| = 2\beta_{\mathfrak{u}}$ and deduce $S_1(\alpha,\boldsymbol{\gamma},\beta_{\mathfrak{u}},\mathfrak{u}) \le 2^{\alpha|\mathfrak{u}|+1}\beta_{\mathfrak{u}}^{1+\alpha}$, and recall that $C(\alpha,\boldsymbol{\gamma},\beta_{\mathfrak{u}},1) = O(\beta_{\mathfrak{u}}^{1-\alpha})$. Thus the trade-off gives

$$\big|\widehat{\mathrm{V}}_{\mathfrak{u}}(f,K_{\mathfrak{u}},G) - \mathrm{V}_{\mathfrak{u}}(f)\big| = O\Big(B(\alpha,n,d,\boldsymbol{\gamma})^{1-\frac{1}{\alpha}}\Big),$$

ii) if $|\mathfrak{u}| = 2$ and $\alpha > 2$, note that $|K_{\mathfrak{u}}| \le 4\beta_{\mathfrak{u}}(\log(\beta_{\mathfrak{u}})+1)$ — see argument for (5.65) in Appendix 5.E — and deduce $S_1(\alpha,\boldsymbol{\gamma},\beta_{\mathfrak{u}},\mathfrak{u}) \le 2^{\alpha|\mathfrak{u}|+2}\beta_{\mathfrak{u}}^{1+\alpha}(\log(\beta_{\mathfrak{u}})+1)$ and recall that $C(\alpha,\boldsymbol{\gamma},\beta_{\mathfrak{u}},1) = O\big(\beta_{\mathfrak{u}}^{1-\alpha}\log(\beta_{\mathfrak{u}})\big)$. Thus the trade-off gives

$$\big|\widehat{\mathrm{V}}_{\mathfrak{u}}(f,K_{\mathfrak{u}},G) - \mathrm{V}_{\mathfrak{u}}(f)\big| = O\Big(\log\big(B(\alpha,n,d,\boldsymbol{\gamma})^{-1/\alpha}\big)B(\alpha,n,d,\boldsymbol{\gamma})^{1-\frac{1}{\alpha}}\Big).$$

Remark 5.3. *In unweighted Korobov spaces i.e. $\boldsymbol{\gamma} = \mathbf{1}$, it is known that the optimal rate of convergence of a rank-1 lattice rule is*

$$B(\alpha,n,d,\boldsymbol{\gamma}) = O\left(\frac{(\log n)^{d\alpha/2}}{n^{\alpha/2}}\right)$$

(see e.g. [SJ94]*). For weighted Korobov spaces, there exist better rates of convergence for product weights i.e. $\gamma_{\mathfrak{u}} = \prod_{i\in\mathfrak{u}}\gamma_i$ (see* [Kuo03]*) or for finite-order weights i.e $\forall\mathfrak{u}$ with $|\mathfrak{u}| > d^*$ ($d^* \le d$), $\gamma_{\mathfrak{u}} = 0$ (see* [DSWW06]*). The latter are essentially related to an assumption on the effective dimension of f in the truncation sense and in the superposition sense, respectively (see* [CMO97] *for the definition of effective dimension).*

5.4.2 Bias in RBD

We now give some results on the well-known issue related to the bias of the estimates in RBD.

Preliminaries

We begin with the definitions of an orthogonal array and the "coincidence defect" (see, e.g., [Owe94])

Definition 5.2. *An orthogonal array in dimension d, with q levels, strength $t \leq d$ and index λ is a matrix with $n = \lambda q^t$ rows and d columns such that in every n-by-t submatrix each of the q^t possible rows — i.e. the distinct t-uples $(l_1, \ldots, l_t)$ where the l_i's take their values in the set of the q levels — occurs exactly the same number λ of times.*

Definition 5.3. *Let A be an orthogonal array in dimension d, with q levels, strength t and index λ. We say that A has the coincidence defect when there exist two rows of A that do agree in $t+1$ columns; otherwise we say that A is defect-free.*

Let $\Pi(q)$ be the set of permutations on $\{0, \frac{1}{q}, \ldots, \frac{q-1}{q}\}$, $\Pi = \Pi(q,d)$ the Cartesian product $(\Pi(q))^d$ and $\mu = \mu(q,d)$ the normalized counting measure on $\Pi(q,d)$. Let A be an orthogonal array in dimension d, with q levels $\{0, \frac{1}{q} \ldots, \frac{q-1}{q}\}$, strength t and index λ, and denote $n = \lambda q^t$ its number of rows. For any permutation $\boldsymbol{\pi} = (\pi_1, \ldots \pi_d) \in \Pi$, denote $A(\boldsymbol{\pi})$ the orthogonal array obtained from A after applying each permutation π_j on the levels of the corresponding j-th factor i.e.

$$\text{for all } \; 1 \leq i \leq n \;\text{ and }\; 1 \leq j \leq d, \quad \big(A(\boldsymbol{\pi})\big)_{ij} = \pi_j(A_{ij}) \; .$$

Note that the $A(\boldsymbol{\pi})$'s and A are orthogonal arrays with the same parameters (see [HSS99]). Conversely, it is also easy to show that if A has strength and index equal to 1 — i.e. as in the classic RBD with an odd integer [1] n —; any other orthogonal array A' with the same paramaters as A is of the form $A(\boldsymbol{\pi})$ for a permutation $\boldsymbol{\pi} \in \Pi$. We are now interested in the quantities

$$\mathbb{E}_\mu\Big[\widehat{\mathrm{V}}(f, A(\boldsymbol{\pi}))\Big] \;\text{ and }\; \mathbb{E}_\mu\Big[\widehat{\mathrm{V}}_{\mathrm{u}}(f, K_{\mathrm{u}}, A(\boldsymbol{\pi}))\Big]$$

where K_{u} is a finite subset of $\mathbb{Z}^*_{\mathrm{u}}$.

Bias of the estimator in RBD

Let $\widehat{\mathrm{c}}_{\mathbf{k}}(f) = \widehat{\mathrm{c}}_{\mathbf{k}}(f, D(q))$ denote the $\mathbf{k}$-th complex discrete Fourier coefficient; we begin with the following important lemma

Theorem 5.2. [Owen] *Following the previous notation, we have*

$$\mathrm{Var}_\mu\Big[\widehat{\mathrm{c}}_{\mathbf{0}}\big(f, A(\boldsymbol{\pi})\big)\Big] = \frac{1}{n^2} \sum_{|\mathrm{u}|>t} \left(\sum_{r=0}^{|\mathrm{u}|} B(\mathrm{u}, r)(1-q)^{r-|\mathrm{u}|} \right) \left(\sum_{\mathbf{k} \in \mathbb{Z}^*_{\mathrm{u}}(q)} |\widehat{\mathrm{c}}_{\mathbf{k}}(f)|^2 \right)$$

where

$$B(\mathrm{u}, r) = \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbf{1}_{|\{l \in \mathrm{u},\; A_{il} = A_{jl}\}| = r}$$

consists of the number of pairs of rows (A_i, A_j) that match on exactly r of the axes in u.

Démonstration. This is exactly Theorem 1 given by Owen in [Owe94]. Just note that, the embedded ANOVA terms on a q^d regular grid — denoted β_{u} by Owen — are

$$\beta_{\mathrm{u}}(\mathbf{x}) = \sum_{\mathbf{k} \in \mathbb{Z}^*_{\mathrm{u}}(q)} \widehat{\mathrm{c}}_{\mathbf{k}}(f) \exp(2i\pi \mathbf{k} \cdot \mathbf{x}).$$

1. If n is even, the design of experiments in RBD consists of an orthogonal array with $n/2$ levels, strength 1 and index 2, and may be faced with the coincidence defect.

Indeed, for all $\mathbf{x}$ in the regular grid $\{0, \frac{1}{q}, \dots, \frac{q-1}{q}\}^d$,

$$f(\mathbf{x}) = \sum_{\mathfrak{u} \subseteq \{1,\dots,d\}} \beta_{\mathfrak{u}}(\mathbf{x})$$

by a trigonometric interpolation argument, and it is also easy to show that the random variables $\beta_{\mathfrak{u}}(X_i,\ i \in \mathfrak{u})$ satisfy the property (5.3) for independent random variables X_i uniformly distributed on $\{0, \frac{1}{q} \dots, \frac{q-1}{q}\}$. □

Then we have the following proposition in which the bias of the variance estimate is investigated in unweighted Korobov spaces $\mathcal{H}_\alpha = \mathcal{H}_{\alpha,\mathbf{1}}$ (see Section 5.4.1.)

Proposition 5.5. *Let A be a defect-free orthogonal array in dimension d with parameters q, t and λ in $\mathbb{N}^*$ with $t < d$. If there exists $\alpha > 2t+1$ such that f and f^2 are in $\mathcal{H}_\alpha$, we have*

$$\mathbb{E}_\mu\Big[\widehat{\mathrm{V}}(f, A(\boldsymbol{\pi}))\Big] = \mathrm{V}(f) - \frac{1}{n} \sum_{1 \le |\mathfrak{u}| < t} \mathrm{V}_{\mathfrak{u}}(f) + O\big(n^{-(1+\frac{1}{t})}\big)\ .$$

Démonstration. See Appendix 5.F □

As a consequence, considering the classic definition of effective dimension in the superposition sense (see e.g. [CMO97])

Definition 5.4. *The effective dimension of f, in the superposition sense, is the smallest $d_S(f)$ such that*

$$\sum_{1 \le |\mathfrak{u}| \le d_S(f)} \mathrm{V}_{\mathfrak{u}}(f) \ge l_S(f) \mathrm{V}(f)$$

where $l_S(f)$ is an arbitrary constant generally set at 0.99.

we have the corollary

Corollary 5.1. *Under the assumptions of Proposition 5.5, let $d_S(f)$ and $l_S(f)$ be defined as in Definition 5.4. If $t \ge d_S$, we have*

$$\mathbb{E}_\mu\Big[\widehat{\mathrm{V}}(f, A(\boldsymbol{\pi}))\Big] = \left(1 - \frac{\varepsilon}{n}\right) \mathrm{V}(f) + O\big(n^{-(1+\frac{1}{t})}\big)$$

where $0 \le \varepsilon \le 1 - l_S(f)$.

Démonstration. Straightforward from Proposition 5.5. □

In a second time, since

$$\mathbb{E}_\mu\Big[\widehat{\mathrm{V}}_{\mathfrak{u}}(f, K_{\mathfrak{u}}, A(\boldsymbol{\pi}))\Big] = \sum_{\mathbf{k} \in K_{\mathfrak{u}}} \mathbb{E}_\mu\Big[\big|\widehat{\mathrm{c}}_{\mathbf{k}}\big(f, A(\boldsymbol{\pi})\big)\big|^2\Big],$$

the analysis of the bias of the parts of variance estimates rests on the following result

Proposition 5.6. *Let A be a defect-free orthogonal array in dimension d with parameters q, t and λ in $\mathbb{N}^*$ with $t < d$. Let $\mathfrak{u}$ be a non-empty subset of $\{1, \dots, d\}$ and $\mathbf{k} \in \mathbb{Z}^*_{\mathfrak{u}}$. If there exists $\alpha > 2t+1$ such that f and f^2 are in $\mathcal{H}_\alpha$, we have*

$$\mathbb{E}_\mu\Big[\big|\widehat{\mathrm{c}}_{\mathbf{k}}(f, A(\boldsymbol{\pi}))\big|^2\Big] = \frac{n-1}{n}|\mathrm{c}_{\mathbf{k}}(f)|^2 + \frac{1}{n}\big(\mathrm{V}(f) + \mathrm{c}_{\mathbf{0}}(f)^2 - R_1 - R_2\big) + O\big(n^{-(1+\frac{1}{t})}\big)$$

where

$$R_1(q, t, \lambda, \mathbf{k}) = \sum_{\substack{1 \le |\mathfrak{v}| \le t \\ \mathfrak{u} \cap \mathfrak{v} = \emptyset}} \sum_{\mathbf{h} \in \mathbb{Z}^*_{\mathfrak{v}}(q)} \big|\mathrm{c}_{\mathbf{k}+\mathbf{h}}(f)\big|^2$$

consists of terms of order strictly higher than $|\mathfrak{u}|$, and

$$R_2(q, t, \lambda, \mathbf{k}) = \sum_{\substack{1 \le |\mathfrak{v}| \le t \\ \mathfrak{u} \cap \mathfrak{v} \ne \emptyset}} \sum_{\mathfrak{v}' \subseteq \mathfrak{v}} (-1)^{|\mathfrak{v}| - |\mathfrak{v}'|} \sum_{\mathfrak{v}'' \subseteq \mathfrak{v}'} \sum_{\mathbf{h} \in \mathbb{Z}^*_{\mathfrak{v}''}(q)} \big|\mathrm{c}_{\mathbf{k}_{\mathfrak{v}'}+\mathbf{h}}(f)\big|^2$$

where $(\mathbf{k}_{\mathfrak{v}'})_i = 0$ if $i \in \mathfrak{v}'$, and $(\mathbf{k}_{\mathfrak{v}'})_i = k_i$ otherwise.

Démonstration. See Appendix 5.G □

We conclude that estimators in RBD are asymptotically unbiased in unweighted Korobov spaces since

$$\begin{aligned}\mathbb{E}_\mu\Big[\widehat{\mathrm{V}}(f,A(\boldsymbol{\pi}))\Big] &= V(f)+\frac{B_1}{n}+o(n^{-1})\\ \mathbb{E}_\mu\Big[\big|\widehat{\mathrm{c}}_{\mathbf{k}}(f,A(\boldsymbol{\pi}))\big|^2\Big] &= |\mathrm{c}_{\mathbf{k}}|^2+\frac{B_2}{n}+o(n^{-1})\end{aligned}$$

where $B_1 \leq \mathrm{V}(f)$ and $B_2 \leq \mathrm{V}(f)+\mathrm{c}_{\mathbf{0}}(f)^2$, and more generally

$$\mathbb{E}_\mu\Big[\widehat{\mathrm{V}}_{\mathrm{u}}(f,K_{\mathrm{u}},A(\boldsymbol{\pi}))\Big]=V_{\mathrm{u}}(f)+\frac{B_3}{n}+\varepsilon_{trunc}(K_{\mathrm{u}})+o(n^{-1})$$

where $B_3 \leq |K_{\mathrm{u}}|(\mathrm{V}(f)+\mathrm{c}_{\mathbf{0}}(f)^2)$ and

$$\varepsilon_{trunc}(K_{\mathrm{u}})=\sum_{\mathbf{k}\in\mathbb{Z}^*_{\mathrm{u}}\setminus K_{\mathrm{u}}}|\mathrm{c}_{\mathbf{k}}(f)|^2$$

is for instance of order $O(M^{|\mathrm{u}|-\alpha})$ if $K_{\mathrm{u}}=\mathbb{Z}^*_{\mathrm{u}}(M)$. Nevertheless, we propose a correction method to reduce a part of these biases.

Application to bias correction

We do not propose any bias correction for the variance estimates since in practice the bias of the latter is generally negligible. So, we are only interested in the bias of the parts of variance estimates

$$\begin{aligned}\widehat{\mathrm{V}}_{\mathrm{u}}(M) &= \widehat{\mathrm{V}}_{\mathrm{u}}\big(f,\mathbb{Z}^*_{\mathrm{u}}(M),A(\boldsymbol{\pi})\big)\ ,\ 1\leq M\leq q\\ \widehat{\mathrm{V}}_{\mathrm{u}}(K_{\mathrm{u}}) &= \widehat{\mathrm{V}}_{\mathrm{u}}\big(f,K_{\mathrm{u}},A(\boldsymbol{\pi})\big)\ ,\ K_{\mathrm{u}}\subseteq\mathbb{Z}^*_{\mathrm{u}}(q)\end{aligned}$$

under the assumptions of Proposition 5.6. In practice, the truncation parameter M, as well as the term $|K_{\mathrm{u}}|^{1/|\mathrm{u}|}$, is of order 5 or higher, and is generally less than 15. For convenience, we now simply denote $R_1(\mathbf{k})=R_1(q,t,\lambda,\mathbf{k})$ and $R_1(K)=\sum_{\mathbf{k}\in K}R_1(q,t,\lambda,\mathbf{k})$.

Example 1 $(t=1,\ |\mathrm{u}|=1)$ Let $1\leq i\leq d$ and $\mathbf{k}\in\mathbb{Z}^*_{\{i\}}$, we have

$$\mathbb{E}_\mu\Big[\big|\widehat{\mathrm{c}}_{\mathbf{k}}(f,A(\boldsymbol{\pi}))\big|^2\Big]=|\mathrm{c}_{\mathbf{k}}(f)|^2+\frac{1}{n}\mathrm{V}_{\sim i}(f)-\frac{1}{n}R_1(\mathbf{k})+O\big(n^{-2}\big) \tag{5.44}$$

where $\mathrm{V}_{\sim i}(f)=\mathrm{V}(f)-\mathrm{V}_i(f)$. Consequently, for any integer $M\leq q$, the estimator $\widehat{\mathrm{V}}_i(M)$ satisfies

$$\mathbb{E}_\mu\Big[\widehat{\mathrm{V}}_i(M)\Big]=\frac{n-(M-1)}{n}\mathrm{V}_i(f)+\frac{M-1}{n}\mathrm{V}(f)-\frac{1}{n}R_1\big(\mathbb{Z}^*_{\{i\}}(M)\big)+O(M^{1-\alpha})+(M-1)O\big(n^{-2}\big)$$

and should be corrected as follows

$$\widehat{\mathrm{V}}^c_i(M)=\frac{n}{n-(M-1)}\widehat{\mathrm{V}}_i(M)-\frac{M-1}{n-(M-1)}\widehat{\mathrm{V}}\big(f,A(\boldsymbol{\pi})\big).$$

Proceeding in this way, the remaining bias is

$$\frac{1}{n-(M-1)}\Big[nO\big(M^{1-\alpha}\big)+(M-1)O\big(n^{-1}\big)-R_1\big(\mathbb{Z}^*_i(M)\big)\Big]$$

where $R_1(\mathbb{Z}^*_i(M))\leq\sum_{j\neq i}\mathrm{V}_{ij}(f)$. Note that (5.44) was partially guessed by Xu & Gertner in [XG11b] (see (44) in their paper) and the bias correction is the same as suggested by Plischke in [Pli10] and proposed by Tissot & Prieur in [TP12a]. More generally, let $K_{\{i\}}$ be a finite subset of $\mathbb{Z}^*_{\{i\}}(q)$; the estimator $\widehat{\mathrm{V}}_i(K_{\{i\}})$ should be corrected as follows

$$\widehat{\mathrm{V}}^c_i(K_{\{i\}})=\frac{n}{n-|K_{\{i\}}|}\widehat{\mathrm{V}}_i(K_{\{i\}})-\frac{|K_{\{i\}}|}{n-|K_{\{i\}}|}\widehat{\mathrm{V}}\big(f,A(\boldsymbol{\pi})\big).$$

Example 2 ($t = 1$, $|\mathfrak{u}| = 2$) This example may be considered as a problematic case since $|\mathfrak{u}| > t$. Let $1 \le i < j \le d$ and $\mathbf{k} \in \mathbb{Z}^*_{\{i,j\}}$, we have

$$\mathbb{E}_\mu\Big[\big|\widehat{\mathrm{c}}_{\mathbf{k}}(f, A(\boldsymbol{\pi}))\big|^2\Big] = \frac{n+1}{n}|\mathrm{c}_{\mathbf{k}}(f)|^2 + \frac{1}{n}\big(\mathrm{V}(f) + \mathrm{c}_{\mathbf{0}}(f)^2\big) + O\big(n^{-2}\big) - \frac{1}{n}\big(R_1(\mathbf{k}) + R_3(\mathbf{k})\big)$$

where

$$R_3(\mathbf{k}) = \frac{1}{n}\Big(|\mathrm{c}_{\mathbf{k}_{\{i\}}}(f)|^2 + |\mathrm{c}_{\mathbf{k}_{\{j\}}}(f)|^2 + \sum_{\mathbf{h}\in\mathbb{Z}^*_{\{i\}}(q)} |\mathrm{c}_{\mathbf{k}_{\{i\}}+\mathbf{h}}(f)|^2 + \sum_{\mathbf{h}\in\mathbb{Z}^*_{\{j\}}(q)} |\mathrm{c}_{\mathbf{k}_{\{j\}}+\mathbf{h}}(f)|^2\Big).$$

Then for any integer $M \le q$, the estimator $\widehat{\mathrm{V}}_{ij}(M)$ satisfies

$$\begin{aligned}\mathbb{E}_\mu\Big[\widehat{\mathrm{V}}_{ij}(M)\Big] &= \frac{n+1}{n}\mathrm{V}_{ij}(f) + \frac{(M-1)^2}{n}\big(\mathrm{V}(f) + \mathrm{c}_{\mathbf{0}}(f)^2\big) + O(M^{2-\alpha}) + (M-1)^2 O\big(n^{-2}\big)\\ &- \frac{1}{n}\Big(R_1\big(\mathbb{Z}^*_{\{i,j\}}(M)\big) + R_3\big(\mathbb{Z}^*_{\{i,j\}}(M)\big)\Big)\end{aligned}$$

and should be corrected as follows

$$\widehat{\mathrm{V}}^c_{ij}(M) = \frac{n}{n+1}\widehat{\mathrm{V}}_{ij}(M) - \frac{(M-1)^2}{n+1}\Big(\widehat{\mathrm{V}}\big(f, A(\boldsymbol{\pi})\big) + \widehat{\mathrm{c}}_{\mathbf{0}}\big(f, A(\boldsymbol{\pi})\big)^2\Big)\Big).$$

Proceeding in this way, the remaining bias is

$$\frac{1}{n+1}\Big[nO\big(M^{2-\alpha}\big) + (M-1)^2 O\big(n^{-1}\big) - R_1\big(\mathbb{Z}^*_{\{i,j\}}(M)\big) - R_3\big(\mathbb{Z}^*_{\{i,j\}}(M)\big)\Big]$$

where $R_1(\mathbb{Z}^*_{\{i,j\}}(M)) \le \sum_{l\ne i,j} \mathrm{V}_{ijl}(f)$ and $R_3(\mathbb{Z}^*_{\{i,j\}}(M)) \le (M-1)(\mathrm{V}_i(f) + \mathrm{V}_j(f) + 2\mathrm{V}_{ij}(f))$. More generally, let $K_{\{i,j\}}$ be a finite subset of $\mathbb{Z}^*_{\{i,j\}}(q)$; the estimator $\widehat{\mathrm{V}}_{ij}(K_{\{i,j\}})$ should be corrected as follows

$$\widehat{\mathrm{V}}^c_{ij}(K_{\{i,j\}}) = \frac{n}{n+1}\widehat{\mathrm{V}}_{ij}(K_{\{i,j\}}) - \frac{|K_{\{i,j\}}|}{n+1}\Big(\widehat{\mathrm{V}}\big(f, A(\boldsymbol{\pi})\big) + \widehat{\mathrm{c}}_{\mathbf{0}}\big(f, A(\boldsymbol{\pi})\big)^2\Big)\Big).$$

Example 3 ($t = 2$, $|\mathfrak{u}| = 1$) Let $1 \le i \le d$ and $\mathbf{k} \in \mathbb{Z}^*_{\{i\}}$, we have

$$\mathbb{E}_\mu\Big[\big|\widehat{\mathrm{c}}_{\mathbf{k}}(f, A(\boldsymbol{\pi}))\big|^2\Big] = |\mathrm{c}_{\mathbf{k}}(f)|^2 + \frac{1}{n}\mathrm{V}_{\sim II}(f) - \frac{d-1}{n}\mathrm{V}_i(f) - \frac{1}{n}R'_1(\mathbf{k}) + O\big(n^{-3/2}\big)$$

where $\mathrm{V}_{\sim II}(f) = \mathrm{V}(f) - \sum_{j=1}^d \mathrm{V}_j(f) - \sum_{\substack{j=1\\ j\ne i}}^d \mathrm{V}_{ij}(f)$ and

$$R'_1(\mathbf{k}) = \sum_{\substack{|\mathfrak{v}|=2\\ \mathfrak{u}\cap\mathfrak{v}=\emptyset}} \sum_{\mathbf{h}\in\mathbb{Z}^*_{\mathfrak{v}}(q)} \big|\mathrm{c}_{\mathbf{k}+\mathbf{h}}(f)\big|^2.$$

Consequently, for any integer $M \le q$, the estimator $\widehat{\mathrm{V}}_i(M)$ satisfies

$$\mathbb{E}_\mu\Big[\widehat{\mathrm{V}}_i(M)\Big] = \frac{n-(d-1)(M-1)}{n}\mathrm{V}_i(f) + \frac{M-1}{n}\mathrm{V}_{\sim II}(f) - \frac{1}{n}R_1\big(\mathbb{Z}^*_{\{i\}}(M)\big) + O\big(M^{1-\alpha}\big) + (M-1)O\big(n^{-3/2}\big)$$

where

$$R'_1(\mathbb{Z}^*_{\{i\}}(M)) \le \sum_{\substack{j<k\\ j,k\ne i}} \mathrm{V}_{ijk}(f).$$

In this case a bias correction could be performed on the term $\mathrm{V}_{\sim II}(f)$, but this is quite intricate — a linear system inversion is needed and the variance of the corrected estimator could significantly increase — and we prefer to keep the basic estimator without bias correction. Proceeding in this way, the bias is

$$B_i = \lambda \mathrm{V}_i(f) + \frac{\lambda}{d-1}\mathrm{V}_{\sim II}(f) - \frac{\lambda}{(d-1)(M-1)}R_1\big(\mathbb{Z}^*_{\{i\}}(M)\big) + O\big(M^{1-\alpha}\big) + (M-1)O\big(n^{-3/2}\big).$$

where $\lambda = (d-1)(M-1)/n$ should be small in practice. More generally, let $K_{\{i\}}$ be a finite subset of $\mathbb{Z}^*_{\{i\}}(q)$; the estimator $\widehat{\mathrm{V}}_i(K_{\{i\}})$ should be kept without bias correction.

Example 4 ($t = 2$, $|\mathrm{u}| = 2$) Let $1 \leq i < j \leq d$ and $\mathbf{k} \in \mathbb{Z}^*_{\{i,j\}}$, we have

$$\mathbb{E}_\mu\Big[\big|\widehat{\mathrm{c}}_{\mathbf{k}}(f, A(\boldsymbol{\pi}))\big|^2\Big] = |\mathrm{c}_{\mathbf{k}}(f)|^2 + \frac{1}{n}\mathrm{V}_{\sim ij}(f) - \frac{1}{n}R_1(\mathbf{k}) - \frac{1}{n}R_3(\mathbf{k}) + O\big(n^{-3/2}\big)$$

where $\mathrm{V}_{\sim ij}(f) = \mathrm{V}(f) - \mathrm{V}_i(f) - \mathrm{V}_j(f) - \mathrm{V}_{ij}(f)$, and

$$\begin{aligned} R_3(\mathbf{k}) &= \sum_{\substack{l=1 \\ l\notin\{i,j\}}}^{d} \sum_{\mathbf{h}\in\mathbb{Z}^*_{\{l\}}(q)} \Big(|\mathrm{c}_{\mathbf{k}_{\{i\}}+\mathbf{h}}(f)|^2 + |\mathrm{c}_{\mathbf{k}_{\{j\}}+\mathbf{h}}(f)|^2 - 2|\mathrm{c}_{\mathbf{k}+\mathbf{h}}(f)|^2 \\ &+ \sum_{\mathbf{h}'\in\mathbb{Z}^*_{\{i\}}(q)} |\mathrm{c}_{\mathbf{k}_{\{i\}}+\mathbf{h}+\mathbf{h}'}(f)|^2 + \sum_{\mathbf{h}'\in\mathbb{Z}^*_{\{j\}}(q)} |\mathrm{c}_{\mathbf{k}_{\{j\}}+\mathbf{h}+\mathbf{h}'}(f)|^2\Big) \,. \end{aligned}$$

Then for any integer $M \leq q$, the estimator $\widehat{\mathrm{V}}_{ij}(M)$ satisfies

$$\begin{aligned} \mathbb{E}_\mu\Big[\widehat{\mathrm{V}}_{ij}(M)\Big] &= \frac{n-(M-1)^2}{n}\mathrm{V}_{ij}(f) + \frac{(M-1)^2}{n}\big(\mathrm{V}(f) - \mathrm{V}_i(f) - \mathrm{V}_j(f)\big) \\ &- \frac{1}{n}R_1\big(\mathbb{Z}^*_{\{i,j\}}(M)\big) - \frac{1}{n}R_3\big(\mathbb{Z}^*_{\{i,j\}}(M)\big) + (M-1)^2 O\big(n^{-3/2}\big) + O(M^{2-\alpha}). \end{aligned}$$

and should be corrected as follows

$$\widehat{\mathrm{V}}^c_{ij}(M) = \frac{1}{n-(M-1)^2}\Big(n\widehat{\mathrm{V}}_{ij}(M) - (M-1)^2\Big(\widehat{\mathrm{V}}\big(f, A(\boldsymbol{\pi})\big) - \widehat{\mathrm{V}}_i(M) - \widehat{\mathrm{V}}_j(M)\Big)\Big).$$

Proceeding in this way, the remaining bias is

$$\frac{1}{n-(M-1)^2}\Big[-R_1\big(\mathbb{Z}^*_{\{i,j\}}(M)\big) - R_3\big(\mathbb{Z}^*_{\{i,j\}}(M)\big) + (M-1)^2 O\big(n^{-1/2}\big) + nO\big(M^{2-\alpha}\big) + (M-1)^2\big(B_i + B_j\big)\Big]$$

where

$$\begin{aligned} R_1(\mathbb{Z}^*_{\{i\}}(M)) &\leq \sum_{k\notin\{i,j\}} \mathrm{V}_{ijk}(f) + \sum_{\substack{k<l \\ \{k,l\}\cap\{i,j\}\neq\emptyset}} \mathrm{V}_{ijkl}(f) \\ R_3(\mathbb{Z}^*_{\{i\}}(M)) &\leq \sum_{k\notin\{i,j\}} \Big(2(M-2)\mathrm{V}_{ijk}(f) + (M-1)\mathrm{V}_{ik}(f) + (M-1)\mathrm{V}_{jk}(f)\Big) \end{aligned}$$

and where the B_i's are the remaining bias in Example 3. More generally, let $K_{\{i,j\}}$ be a finite subset of $\mathbb{Z}^*_{\{i,j\}}(q)$; the estimators $\widehat{\mathrm{V}}_{ij}(K_{\{i,j\}})$ should be corrected as follows

$$\widehat{\mathrm{V}}^c_{ij}(K_{\{i,j\}}) = \frac{1}{n-|K_{\{i,j\}}|}\Big(n\widehat{\mathrm{V}}_{ij}(K_{\{i,j\}}) - |K_{\{i,j\}}|\Big(\widehat{\mathrm{V}}\big(f, A(\boldsymbol{\pi})\big) - \widehat{\mathrm{V}}_i(K_{\{i\}}) - \widehat{\mathrm{V}}_j(K_{\{j\}})\Big)\Big).$$

In the sequel, we denote $\widehat{\mathrm{S}}^c_{\mathrm{u}}\big(f, K, A(\boldsymbol{\pi})\big)$ the index $\widehat{\mathrm{V}}^c_{\mathrm{u}}\big(f, K, A(\boldsymbol{\pi})\big)/\widehat{\mathrm{V}}\big(f, A(\boldsymbol{\pi})\big)$.

5.5 Numerical illustrations

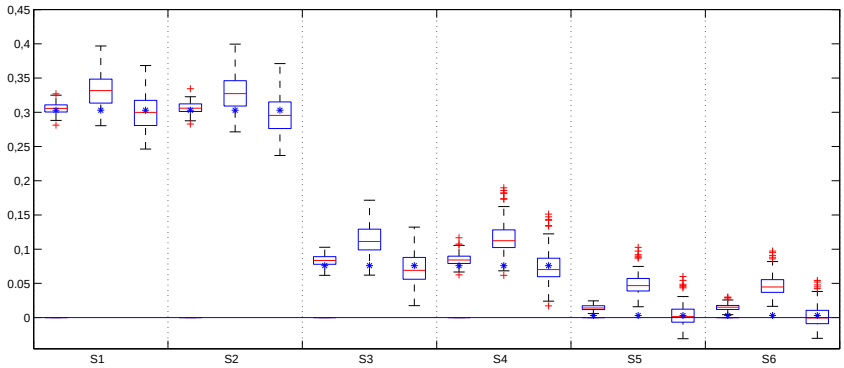
In this section, we apply the bias correction method of Section 5.4.2. on the first and the second-order sensitivity indices computed with RBD when the model is the Sobol' g-function (see [Sob03])

$$f(X_1, \ldots, X_d) = \prod_{i=1}^{d} \frac{|4X_i - 2| + a_i}{1 + a_i}$$

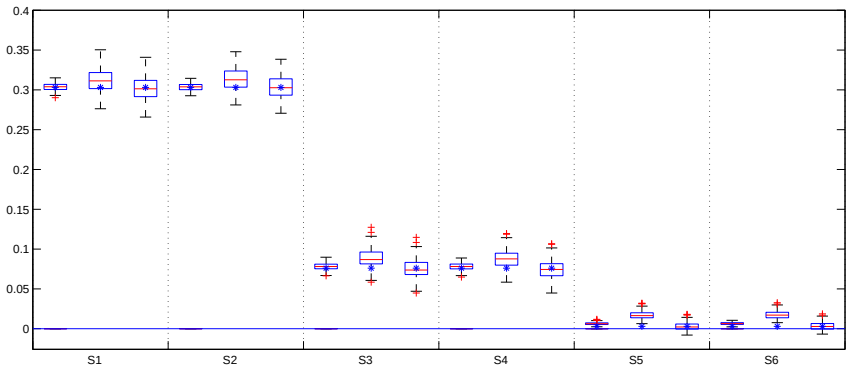
where the a_i's are non-negative parameters and the X_i's are independent random variables uniformly distributed in $[0,1]$. Note that for any $\mathbf{k} \in \mathbb{Z}^d$

$$\mathrm{c}_{\mathbf{k}}(f) = \begin{cases} 0 & \text{if } \exists i \in \{1,\dots,d\} \mid k_i \neq 0 \text{ and } k_i \text{ is even} \\ \displaystyle\frac{\displaystyle\prod_{i \mid k_i \neq 0} 4\pi^{-2}(1+a_i)^{-1}}{\displaystyle\prod_{i \mid k_i \neq 0} k_i^2} & \text{otherwise.} \end{cases}$$

We consider a test-case with $d = 6$ and $\mathbf{a} = (0,0,1,1,9,9)$. Exact values of the sensitivity indices are known; we have $\mathrm{S}_1(f) = \mathrm{S}_2(f) = 0.303$, $\mathrm{S}_3(f) = \mathrm{S}_4(f) = 0.076$, $\mathrm{S}_{12} = 0.101$,$\mathrm{S}_{13}(f) = \mathrm{S}_{14}(f) = \mathrm{S}_{23}(f) = \mathrm{S}_{24}(f) = 0.025$, $\mathrm{S}_{34} = 0.006$ and the other indices are less than 5.10^{-3}. In each illustration, we show boxplots of 100 estimates computed on a randomized array $A(\boldsymbol{\pi})$ — see Section 5.4.2. — of a certain orthogonal array A. In these boxplots, the red central mark is the median; the box has its lower and upper edges at the 25^{th} percentile q and the 75^{th} percentile Q, respectively; the whiskers extend between $q - 1.5(Q-q)$ and $Q + 1.5(Q-q)$; the red crosses are outliers and blue asterisks are exact values. Two arrays A are tested. The first one, denoted $A_{1,n}$, is an orthogonal array with index unity, strength 1 and q levels — and then $n = q$ —; it corresponds with the classic RBD method and its construction is obvious. The second one, denoted $A_{2,n}$ is an orthogonal array with index unity, strength 2 and q levels, where q is a prime — and then $n = q^2$. This array is obtained by using Bose's construction (see [Bos38]).



(a) $n = 529$

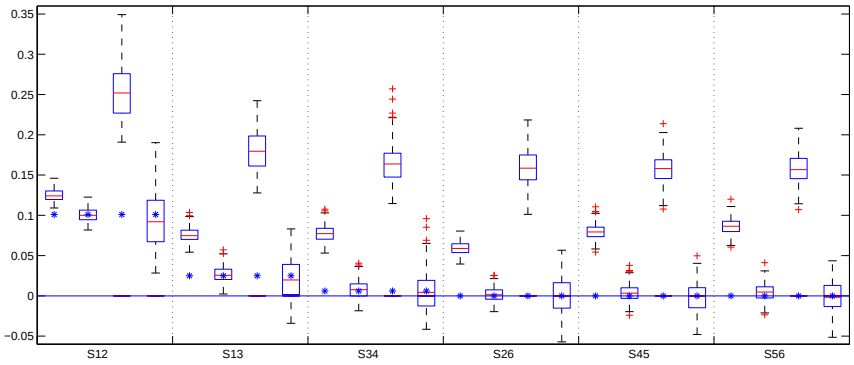


(b) $n = 1681$

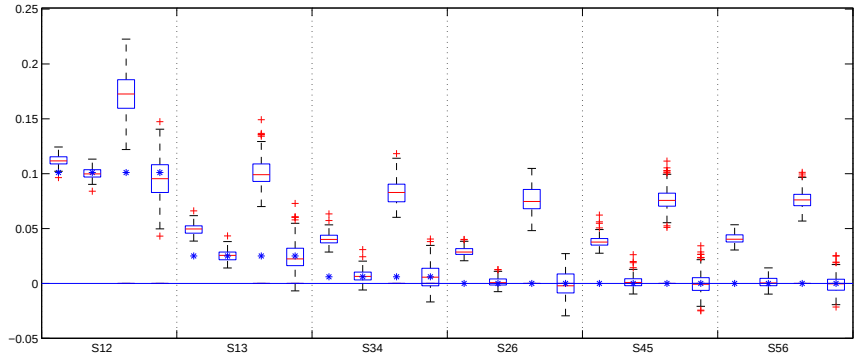
FIGURE 5.3 – Boxplots of the first-order sensitivity indices estimates. For each sensitivity index, S_1, S_2,..., S_6, from the left to the right are $\widehat{\mathrm{S}}_i\big(\mathcal{R}_1 f, \mathcal{Z}_{\{i\},12}, A_2(\boldsymbol{\pi})\big)$ (strength 2 orthogonal array, without bias correction), $\widehat{\mathrm{S}}_i\big(\mathcal{R}_1 f, \mathcal{Z}_{\{i\},12}, A_1(\boldsymbol{\pi})\big)$ (strength 1 orthogonal array, without bias correction), $\widehat{\mathrm{S}}_i^c\big(\mathcal{R}_1 f, \mathcal{Z}_{\{i\},12}, A_1(\boldsymbol{\pi})\big)$ (strength 1 orthogonal array, with bias correction), respectively.

Figure 5.3 shows boxplots of the first-order sensitivity indices estimates when the orthogonal array A is $A_{1,529}$, $A_{2,529}$, $A_{1,1681}$ and $A_{2,1681}$, with and without bias correction. We see obviously that A_2 leads to better estimates than A_1 in term of variance. We also notice that the bias correction performed, when A_1 is used, is efficient ; and the estimates, when A_2 is used, are almost without any bias.

Figure 5.4 shows boxplots of six of the fifteen second-order sensitivity estimates when the orthogonal array A is $A_{1,1681}$, $A_{2,1681}$, $A_{1,3481}$ and $A_{2,3481}$, with and without bias correction. One more time, A_2 leads to better estimates than A_1 in term of variance, and the bias correction methods perform well.



(a) $n = 1681$



(b) $n = 3481$

FIGURE 5.4 – Boxplots of the second-order sensitivity indices estimates. For each sensitivity index, S_{12}, S_{13},..., S_{56}, from the left to the right are $\widehat{\mathrm{S}}_{ij}\big(\mathcal{R}_1 f, \mathcal{Z}_{\{i,j\},12}, A_{2,n}(\boldsymbol{\pi})\big)$ (strength 2 orthogonal array, without bias correction), $\widehat{\mathrm{S}}^c_{ij}\big(\mathcal{R}_1 f, \mathcal{Z}_{\{i,j\},12}, A_{2,n}(\boldsymbol{\pi})\big)$ (strength 2 orthogonal array, with bias correction), $\widehat{\mathrm{S}}_{ij}\big(\mathcal{R}_1 f, \mathcal{Z}_{\{i,j\},12}, A_{1,n}(\boldsymbol{\pi})\big)$ (strength 1 orthogonal array, without bias correction), $\widehat{\mathrm{S}}^c_{ij}\big(\mathcal{R}_1 f, \mathcal{Z}_{\{i,j\},12}, A_{1,n}(\boldsymbol{\pi})\big)$ (strength 1 orthogonal array, with bias correction), respectively.

5.6 Conclusions

In this paper we revisited the variance-based sensitivity methods, FAST and RBD, by linking them to commonly used methods in numerical integration field. They are introduced in light of the DFT on finite subgroups of the torus and the use of randomized orthogonal arrays for integration. First we explained the classic FAST in terms of trigonometric interpolation and we introduced a new criterion to choose the set of frequencies free of interferences. We also derived, from the lattice rules theory, explicit rates of convergence for the estimators of the first and second-order partial variances, and the total variance. In a second time, we explained the classic RBD in terms of integration on a randomized orthogonal array with strength 1, and naturally generalized this method to any orthogonal array. We

then studied the well-known issue due to the bias and proposed a correction method in the most common cases. Further work will consist in investigating the variance of the estimators in RBD in order to propose a bias-variance trade-off. As far as we know, apart from the application of shrinkage due to Tarantola & Koda [TK10], this issue related to the variance is not studied much. It will also consist in applying the FAST method by following Proposition 5.4 and employing embedded lattice rules (see [CKN06]).

Acknowledgments

The authors thank Hervé Monod for valuable discussions. This work has been partially supported by French National Research Agency (ANR) through COSINUS program (project COSTA-BRAVA n° ANR-09-COSI-015).

5.A Proof of Proposition 5.2

On the one hand, noting that for all $x \in \mathbb{R}$,

$$\arcsin\big(\sin(x)\big) = \arcsin\Big(\sin\Big(2\pi\Big\{\frac{x}{2\pi}\Big\}\Big)\Big) = \begin{cases} 2\pi\big\{\frac{x}{2\pi}\big\} & \text{if } 0 \le \big\{\frac{x}{2\pi}\big\} < \frac{1}{4} \\ \pi - 2\pi\big\{\frac{x}{2\pi}\big\} & \text{if } \frac{1}{4} \le \big\{\frac{x}{2\pi}\big\} < \frac{3}{4} \\ 2\pi\big\{\frac{x}{2\pi}\big\} - 2\pi & \text{otherwise,} \end{cases} \tag{5.45}$$

we get that for any $i \in \{1, \dots d\}$ and $j \in \{0, \dots, n-1\}$,

$$x_i^*\Big(\frac{j}{n}\Big) = \frac{1}{\pi}\arcsin\Big(\sin\Big(2\pi\omega_i\frac{j}{n} + \varphi_i\Big)\Big) + \frac{1}{2} = r_1 \circ t_{\varphi_i}\Big(\Big\{\frac{j}{n}\omega_i\Big\}\Big). \tag{5.46}$$

Thus we have

$$f \circ \mathbf{x}^*\Big(\frac{j}{n}\Big) = (\mathcal{T}_{\boldsymbol{\varphi}} \circ \mathcal{R}_1)f\Big(\Big\{\frac{j}{n}\omega_1\Big\}, \dots, \Big\{\frac{j}{n}\omega_d\Big\}\Big), \tag{5.47}$$

and we easily deduce that for all $\mathbf{k} \in \mathbb{Z}^d$,

$$\big|\widehat{\mathrm{c}}_{\mathbf{k}\cdot\boldsymbol{\omega}}(f \circ \mathbf{x}^*)\big| = \big|\widehat{\mathrm{c}}_{\mathbf{k}}\big((\mathcal{T}_{\boldsymbol{\varphi}} \circ \mathcal{R}_1)f, G(\boldsymbol{\omega})\big)\big|. \tag{5.48}$$

Finally we obtain that for any non-empty set $\mathfrak{u} \subseteq \{1, \dots, d\}$ and any finite set $K_{\mathfrak{u}} \subseteq \mathbb{Z}^*_{\mathfrak{u}}$

$$\widehat{\mathrm{V}}^{\mathrm{FAST}}_{\mathfrak{u}}(f, K_{\mathfrak{u}}, \mathbf{x}^*) = \widehat{\mathrm{V}}_{\mathfrak{u}}\big((\mathcal{T}_{\boldsymbol{\varphi}} \circ \mathcal{R}_1)f, K_{\mathfrak{u}}, G(\boldsymbol{\omega})\big). \tag{5.49}$$

Recalling that $\widehat{\mathrm{V}}^{\mathrm{FAST}}(f, \mathbf{x}^*) = \widehat{\mathrm{V}}(f, \{\mathbf{x}^*(\frac{j}{n})\}_{j=0..n-1})$, (5.47) obviously leads to

$$\widehat{\mathrm{V}}^{\mathrm{FAST}}(f, \mathbf{x}^*) = \widehat{\mathrm{V}}\big((\mathcal{T}_{\boldsymbol{\varphi}} \circ \mathcal{R}_1)f, G(\boldsymbol{\omega})\big). \tag{5.50}$$

We conclude to (5.33) by combining (5.49) and (5.50).

On the other hand, we also deduce from (5.45) that for any $i \in \{1, \dots d\}$ and $j \in \{0, \dots, n-1\}$,

$$x_i^\times\Big(\frac{j}{n}\Big) = \frac{1}{\pi}\arcsin\Big(\sin\Big(2\pi\omega\frac{\sigma_i(j)}{n}\Big)\Big) + \frac{1}{2} = r_\omega \circ t_{\frac{(1-\omega)\pi}{2\omega}}\Big(\frac{\sigma_i(j)}{n}\Big).$$

Thus we have

$$f \circ \mathbf{x}^\times\Big(\frac{j}{n}\Big) = (\mathcal{T}_{\tilde{\boldsymbol{\omega}}} \circ \mathcal{R}_\omega)f\Big(\frac{\sigma_1(j)}{n}, \dots, \frac{\sigma_d(j)}{n}\Big), \tag{5.51}$$

and we easily deduce that for all $i \in \{1, \dots, d\}$ and $k_i \in \mathbb{Z}$,

$$\widehat{\mathrm{c}}_{k_i\omega}(f \circ \mathbf{x}^{\times,i}) = \widehat{\mathrm{c}}_{(0,\dots,0,k_i\omega,0,\dots,0)}\big((\mathcal{T}_{\tilde{\boldsymbol{\omega}}} \circ \mathcal{R}_\omega)f, A(\boldsymbol{\sigma})\big).$$

Finally we obtain that for any non-empty $i \in \{1, \dots, d\}$ and any finite set $K_{\{i\}} \subseteq \mathbb{Z}^*_{\{i\}}$

$$\widehat{\mathrm{V}}^{\mathrm{RBD}}_i(f, K_{\{i\}}, \mathbf{x}^\times) = \widehat{\mathrm{V}}_i\big((\mathcal{T}_{\tilde{\boldsymbol{\omega}}} \circ \mathcal{R}_\omega)f, \omega K_{\{i\}}, A(\boldsymbol{\sigma})\big). \tag{5.52}$$

Recalling that $\widehat{\mathrm{V}}^{\mathrm{RBD}}(f, \mathbf{x}^\times) = \widehat{\mathrm{V}}(f, \{\mathbf{x}^\times(\frac{j}{n})\}_{j=0..n-1})$, (5.51) obviously leads to

$$\widehat{\mathrm{V}}^{\mathrm{RBD}}(f, \mathbf{x}^\times) = \widehat{\mathrm{V}}\big((\mathcal{T}_{\tilde{\boldsymbol{\omega}}} \circ \mathcal{R}_\omega)f, A(\boldsymbol{\sigma})\big). \tag{5.53}$$

We conclude to (5.34) by combining (5.52) and (5.53).

5.B Further issue : influence of the parameter ω in the classic RBD

In the proof of Proposition 5.2, it is easy to show that Eqs. (5.46) to (5.50) can be successively replaced by

$$\begin{aligned} x_i^{\times}\Big(\frac{j}{n}\Big) &= r_1\Big(\Big\{\omega\frac{\sigma_i(j)}{n}\Big\}\Big), \\ f\circ\mathbf{x}^{\times}\Big(\frac{j}{n}\Big) &= \mathcal{R}_1 f\Big(\Big\{\omega\frac{\sigma_1(j)}{n}\Big\},\ldots,\Big\{\omega\frac{\sigma_d(j)}{n}\Big\}\Big), \\ \widehat{\mathrm{c}}_{k_i\omega}(f\circ\mathbf{x}^{\times,i}) &= \widehat{\mathrm{c}}_{(0,\ldots,0,k_i,0,\ldots,0)}\big(\mathcal{R}_1 f,\{\omega A(\boldsymbol{\sigma})\}\big), \\ \widehat{\mathrm{V}}_i^{\mathrm{RBD}}(f,K_{\{i\}},\mathbf{x}^{\times}) &= \widehat{\mathrm{V}}_i\big(\mathcal{R}_1 f,K_{\{i\}},\{\omega A(\boldsymbol{\sigma})\}\big) \end{aligned}$$

and

$$\widehat{\mathrm{V}}^{\mathrm{RBD}}(f,\mathbf{x}^{\times}) = \widehat{\mathrm{V}}\big(\mathcal{R}_1 f,\{\omega A(\boldsymbol{\sigma})\}\big)$$

where

$$\{\omega A(\boldsymbol{\sigma})\} = \left\{\Big(\Big\{\omega\frac{\sigma_1(j)}{n}\Big\},\ldots,\Big\{\omega\frac{\sigma_d(j)}{n}\Big\}\Big),\ j\in\{0,\ldots,n-1\}\right\}.$$

Consequently, (5.34) can be replaced by

$$\widehat{\mathrm{S}}_i^{\mathrm{RBD}}(f,K_{\{i\}},\mathbf{x}^{\times}) = \widehat{\mathrm{S}}_i\big(\mathcal{R}_1 f,K_{\{i\}},\{\omega A(\boldsymbol{\sigma})\}\big), \tag{5.54}$$

and it means that ω has an influence on the estimator through the orthogonal array on which the function $\mathcal{R}_1 f$ is evaluated.

Now following the Definition 5.2 in Section 5.4.2., note that if A is an orthogonal array with q levels $\{0,\frac{1}{q},\ldots,\frac{q-1}{q}\}$, strength t and index λ — and denote $n=\lambda q^t$ its cardinal —, then for any $p\in\mathbb{N}^*$, $\{pA\}$ is an orthogonal array with $q'=q/gcd(p,q)$ levels $\{0,\frac{1}{q'},\ldots,\frac{q'-1}{q'}\}$, strength t' larger or equal to t, and index $\lambda'=n/(q't')$. Indeed, consider $\{0,\frac{1}{q},\ldots,\frac{q-1}{q}\}$ as the cyclic group $\mathbb{Z}/q\mathbb{Z}$ and note that the homomorphism

$$\begin{array}{rccc} \Phi: & \mathbb{Z}/q\mathbb{Z} & \longrightarrow & \mathbb{Z}/q\mathbb{Z} \\ & \overline{z} & \longmapsto & \overline{pz} \end{array}$$

is surjective on $\mathbb{Z}/q'\mathbb{Z}$, where $q'=q/gcd(p,q)$. Consequently, it is easy to deduce that $\{pA\}$ has q' levels and has at least strength t.

As a consequence, in the classic RBD, if ω is relatively prime with the number of levels of the orthogonal array $A(\boldsymbol{\sigma})$ — recall that it is $|A(\boldsymbol{\sigma})|/2$ if $A(\boldsymbol{\sigma})$ is even and $|A(\boldsymbol{\sigma})|$ otherwise —, then the method is exactly equivalent to the basic one with $\omega=1$. On the contrary, if they are not relatively prime, the orthogonal array on which $\mathcal{R}_1 f$ is evaluated has fewer levels and at least the same strength. Moreover in this case, the orthogonal array could be not simple, i.e. its points are not distinct. Thus the estimator (5.54) has potentially a larger bias and a larger variance.

5.C Proof of Lemma 5.1

Let $X_1,\ldots,X_d$ be d independent random variables uniformly distributed on $[0,1]$ and denote $f_{\mathfrak{u}}(X_i,i\in\mathfrak{u})$, $\mathfrak{u}\subseteq\{1,\ldots d\}$ the Hoeffding decomposition of $f(\mathbf{X})$. We first prove the result for the linear operator $\mathcal{R}_1$. Let s be a positive integer and $\mathcal{Q}^s$ be the set of the subset Q of $[0,1[^s$ of the form $\mathrm{Q}=[q_1,q_1+\frac{1}{2}[\times\cdots\times[q_s,q_s+\frac{1}{2}[$ where $q_i\in\{0,\frac{1}{2}\}$. Note that, since the Lebesgue measure is isometry-invariant, we have for any $\mathrm{Q}\in\mathcal{Q}^s$ and any function $g\in L^2([0,1]^s)$,

$$\int_{\mathrm{Q}}\mathcal{R}_1 g(\mathbf{x})d\mathbf{x} = \int_{[0,\frac{1}{2}[^s}\mathcal{R}_1 g(\mathbf{x})d\mathbf{x}\ .$$

Thus it comes

$$\begin{aligned}\int_{[0,1[^s} \mathcal{R}_1 g(\mathbf{x}) d\mathbf{x} &= \sum_{\mathbb{Q}\in\mathcal{Q}^s} \int_{\mathbb{Q}} \mathcal{R}_1 g(\mathbf{x}) d\mathbf{x} \\ &= 2^s \int_{[0,\frac{1}{2}[^s} \mathcal{R}_1 g(\mathbf{x}) d\mathbf{x}\end{aligned}$$

and the definition of $\mathcal{R}_1$ gives

$$\int_{[0,1[^s} \mathcal{R}_1 g(\mathbf{x}) d\mathbf{x} = \int_{[0,1[^s} g(\mathbf{x}) d\mathbf{x} \ . \tag{5.55}$$

Then noting that for all $\mathbf{x} \in [0,1[^d$, $(\mathcal{R}_1 g(\mathbf{x}))^2 = \mathcal{R}_1(g(\mathbf{x}))^2$, we deduce that for all set $\mathfrak{u} \subseteq \{1,\dots,d\}$,

$$\mathrm{Var}\big[\mathcal{R}_1 f_{\mathfrak{u}}(X_i, i\in\mathfrak{u})\big] = \mathrm{Var}\big[f_{\mathfrak{u}}(X_i, i\in\mathfrak{u})\big]. \tag{5.56}$$

We also deduce from (5.55) that for all set $\mathfrak{u} \subseteq \{1,\dots,d\}$,

$$\forall \beta \subsetneq \mathfrak{u}, \quad \mathbb{E}\big[\mathcal{R}_1 f_{\mathfrak{u}}(X_i, i\in\mathfrak{u}) | X_i, i\in\beta\big] = \mathbb{E}\big[f_{\mathfrak{u}}(X_i, i\in\mathfrak{u}) | X_i, i\in\beta\big] \ ,$$

and then, by the uniqueness of the Hoeffding decomposition and the criterion in (5.3),

$$\forall \mathfrak{u} \subseteq \{1,\dots,d\}, \quad (\mathcal{R}_1 f)_{\mathfrak{u}} = \mathcal{R}_1 f_{\mathfrak{u}} \ . \tag{5.57}$$

Finally (5.56) and (5.57) lead to the conclusion of Lemma 5.1 for the linear operator $\mathcal{R}_1$. The proof of Lemma 5.1 for any $\mathcal{R}_p$ with $p \in \mathbb{N}^*$ and for the $\mathcal{T}_{\boldsymbol{\varphi}}$'s is exactly the same as the previous one. It only sufficies to prove that the property in (5.55) hold for any $\mathcal{R}_p$ and $\mathcal{T}_{\boldsymbol{\varphi}}$. This property for the $\mathcal{T}_{\boldsymbol{\varphi}}$'s is a consequence of the translation-invariance of the Lebesgue measure and is omitted here. For the $\mathcal{R}_p$'s, note that for all $x \in [0,1]$, $r_p(x) = r_1(\{px\})$ and deduce that for all $\mathbf{x} \in [0,1]^s$, $\mathcal{R}_p g(\mathbf{x}) = \mathcal{R}_1 g(\{px_1\},\dots,\{px_s\})$. Hence, noting that $\mathcal{R}_p g$ is $\frac{1}{p}$-periodic in each direction, it comes

$$\begin{aligned}\int_{[0,1[^s} \mathcal{R}_p g(\mathbf{x}) d\mathbf{x} &= p^s \int_{[0,\frac{1}{p}[^s} \mathcal{R}_p g(\mathbf{x}) d\mathbf{x} \\ &= p^s \int_{[0,\frac{1}{p}[^s} \mathcal{R}_1 g(px_1,\dots,px_s) d\mathbf{x} \\ &= \int_{[0,1[^s} g(\mathbf{x}) d\mathbf{x} \ .\end{aligned}$$

5.D Proof of (5.39) in Proposition 5.3

Let $\sim$ denote the relation such that for all $\mathbf{k}$, and $\mathbf{k}'$ in $\mathbb{Z}^d$,

$$\mathbf{k} \sim \mathbf{k}' \Longleftrightarrow \mathbf{k} - \mathbf{k}' \in G^\perp \ .$$

This is obviously an equivalence relation and its classes are of the form

$$G^\perp_{\mathbf{k}} = \{\mathbf{k} + \mathbf{h}, \ \mathbf{h} \in G^\perp\} \ .$$

Hence we have

$$\sum_{\mathbf{k}\in K} \sum_{\mathbf{h}\in G^\perp} \mathrm{c}_{\mathbf{k}+\mathbf{h}}(f) \exp\big(2i\pi(\mathbf{k}+\mathbf{h})\cdot\mathbf{x}\big) = \sum_{\mathbf{k}\in K} \sum_{\mathbf{h}\in G^\perp_{\mathbf{k}}} \mathrm{c}_{\mathbf{h}}(f) \exp(2i\pi\mathbf{h}\cdot\mathbf{x})$$

Now, under the assumption that G satisfies the criterion (5.37), for all $\mathbf{k} \in K$ the classes $G^\perp_{\mathbf{k}}$ are distinct. Moreover, it can be shown that

$$\mathbb{Z}^d / G^\perp \simeq G^*$$

where G^* is the dual group of G (see e.g. Paragraph 2.1.2. in [Rud62]) and as a consequence, the number of classes — which is equal to the cardinal of the quotient $\mathbb{Z}^d/G^\perp$ — is equal to $|G^*| = |G| = n$. Thus we have

$$\bigsqcup_{\mathbf{k}\in K} G^\perp_{\mathbf{k}} = \mathbb{Z}^d$$

and we conclude that

$$\sum_{\mathbf{k}\in K}\sum_{\mathbf{h}\in G^\perp} \mathrm{c}_{\mathbf{k}+\mathbf{h}}(f)\exp\big(2i\pi(\mathbf{k}+\mathbf{h})\cdot\mathbf{x}\big) = \sum_{\mathbf{k}\in\mathbb{Z}^d} \mathrm{c}_{\mathbf{k}}(f)\exp(2i\pi\mathbf{k}\cdot\mathbf{x}) \ .$$

5.E Proof of Proposition 5.4

For convenience we now denote $B(\alpha) = B(\alpha, n, d, \boldsymbol{\gamma})$.

First for any $\mathbf{k}\in\mathbb{Z}^d$ and $f\in\mathcal{H}_{\alpha,\boldsymbol{\gamma}}$, denote $f_{\mathbf{k}} : \mathbf{x}\mapsto f(\mathbf{x})\exp(-2i\pi\mathbf{k}\cdot\mathbf{x})$ and note that $f_{\mathbf{k}}\in\mathcal{H}_{\alpha,\boldsymbol{\gamma}}$, $\mathrm{c}_{\mathbf{0}}(f_{\mathbf{k}}) = \mathrm{c}_{\mathbf{k}}(f)$ and $\widehat{\mathrm{c}}_{\mathbf{0}}(f_{\mathbf{k}}, G) = \widehat{\mathrm{c}}_{\mathbf{k}}(f, G)$. Now we have

$$\begin{aligned}
\big||\widehat{\mathrm{c}}_{\mathbf{k}}(f,G)|^2 - |\mathrm{c}_{\mathbf{k}}(f)|^2\big| &= \big|\big(\widehat{\mathrm{c}}_{\mathbf{k}}(f,G) - \mathrm{c}_{\mathbf{k}}(f)\big)\overline{\widehat{\mathrm{c}}_{\mathbf{k}}(f,G)} - \mathrm{c}_{\mathbf{k}}(f)\big(\overline{\mathrm{c}_{\mathbf{k}}(f)} - \overline{\widehat{\mathrm{c}}_{\mathbf{k}}(f,G)}\big)\big| \\
&\le \big|\widehat{\mathrm{c}}_{\mathbf{k}}(f,G) - \mathrm{c}_{\mathbf{k}}(f)\big|\cdot\big|\overline{\widehat{\mathrm{c}}_{\mathbf{k}}(f,G)}\big| + \big|\mathrm{c}_{\mathbf{k}}(f)\big|\cdot\big|\overline{\mathrm{c}_{\mathbf{k}}(f)} - \overline{\widehat{\mathrm{c}}_{\mathbf{k}}(f,G)}\big| \\
&\le ||f_{\mathbf{k}}||_{\mathcal{H}_{\alpha,\boldsymbol{\gamma}}} B(\alpha)\big(2|\mathrm{c}_{\mathbf{k}}(f)| + ||f_{\mathbf{k}}||_{\mathcal{H}_{\alpha,\boldsymbol{\gamma}}} B(\alpha)\big) \ . \qquad (5.58)
\end{aligned}$$

In particular, for $\mathbf{k} = \mathbf{0}$, it comes

$$\big||\widehat{\mathrm{c}}_{\mathbf{0}}(f,G)|^2 - |\mathrm{c}_{\mathbf{0}}(f)|^2\big| \le ||f||^2_{\mathcal{H}_{\alpha,\boldsymbol{\gamma}}} B(\alpha)\big(2 + B(\alpha)\big) \ . \qquad (5.59)$$

We now prove the two items of Proposition 5.4. For the first one, Note that

$$\begin{aligned}
\big|\widehat{\mathrm{V}}(f,G) - \mathrm{V}(f)\big| &= \left|\frac{1}{n}\sum_{g\in G} f^2(g) - |\widehat{\mathrm{c}}_{\mathbf{0}}(f,G)|^2 - \int_{[0,1]^d} f^2(\mathbf{x})d\mathbf{x} + |\mathrm{c}_{\mathbf{0}}(f)|^2\right| \\
&\le \big||\widehat{\mathrm{c}}_{\mathbf{0}}(f^2,G)| - |\mathrm{c}_{\mathbf{0}}(f^2)|\big| + \big||\widehat{\mathrm{c}}_{\mathbf{0}}(f,G)|^2 - |\mathrm{c}_{\mathbf{0}}(f)|^2\big|
\end{aligned}$$

and the conclusion follows from (5.59). For the second item, (5.58) gives

$$\big|\widehat{\mathrm{V}}_{\mathrm{u}}(f,K_{\mathrm{u}},G) - \mathrm{V}_{\mathrm{u}}(f)\big| = \left|\sum_{\mathbf{k}\in\mathbb{Z}^*_{\mathrm{u}}\setminus K_{\mathrm{u}}} |\mathrm{c}_{\mathbf{k}}(f)|^2 - \sum_{\mathbf{k}\in K_{\mathrm{u}}}\Big(|\mathrm{c}_{\mathbf{k}}(f)|^2 - |\widehat{\mathrm{c}}_{\mathbf{k}}(f,G)|^2\Big)\right|$$

$$\le \sum_{\mathbf{k}\in\mathbb{Z}^*_{\mathrm{u}}\setminus K_{\mathrm{u}}} \frac{||f||^2_{\mathcal{H}_{\alpha,\boldsymbol{\gamma}}}}{r(\mathbf{k},\alpha,\boldsymbol{\gamma})} + B(\alpha)^2\sum_{\mathbf{k}\in K_{\mathrm{u}}} ||f_{\mathbf{k}}||^2_{\mathcal{H}_{\alpha,\boldsymbol{\gamma}}} + 2B(\alpha)\sum_{\mathbf{k}\in K_{\mathrm{u}}} |\mathrm{c}_{\mathbf{k}}(f)|\ ||f_{\mathbf{k}}||_{\mathcal{H}_{\alpha,\boldsymbol{\gamma}}}, \qquad (5.60)$$

and the proof is then divided into two parts :

First part. In the second term in the right-hand side of (5.60), let $r(\mathbf{0},\alpha,\boldsymbol{\gamma}) = 1$ and note that

$$||f_{\mathbf{k}}||^2_{\mathcal{H}_{\alpha,\boldsymbol{\gamma}}} = \sum_{\substack{\mathbf{h}\in\mathbb{Z}^d \\ \gamma_{\mathrm{u}_{\mathbf{h}}}\neq 0}} r(\mathbf{h},\alpha,\boldsymbol{\gamma})|\mathrm{c}_{\mathbf{h}}(f_{\mathbf{k}})|^2 = \sum_{\substack{\mathbf{h}\in\mathbb{Z}^d \\ \gamma_{\mathrm{u}_{\mathbf{h}}}\neq 0}} \frac{r(\mathbf{h},\alpha,\boldsymbol{\gamma})}{r(\mathbf{h}+\mathbf{k},\alpha,\boldsymbol{\gamma})} r(\mathbf{h}+\mathbf{k},\alpha,\boldsymbol{\gamma})|\mathrm{c}_{\mathbf{h}+\mathbf{k}}(f)|^2 \ .$$

Then denoting $\gamma_{frac} = \max_{\mathrm{u},\mathfrak{v}\subset\{1,\dots,d\},\gamma_{\mathfrak{v}}\neq 0} \gamma_{\mathrm{u}}/\gamma_{\mathfrak{v}}$, for any $\mathbf{k}\in K_{\mathrm{u}}$,

$$\frac{r(\mathbf{h},\alpha,\boldsymbol{\gamma})}{r(\mathbf{h}+\mathbf{k},\alpha,\boldsymbol{\gamma})} \le \gamma_{frac}\prod_{i\in\mathrm{u}}(|k_i|+1)^\alpha \qquad (5.61)$$

and thus

$$||f_{\mathbf{k}}||_{\mathcal{H}_{\alpha,\boldsymbol{\gamma}}} \le \gamma_{frac}\prod_{i\in\mathrm{u}}(|k_i|+1)^{\alpha/2}||f||_{\mathcal{H}_{\alpha,\boldsymbol{\gamma}}} \ .$$

To prove (5.61), note that

$$\frac{r(\mathbf{h},\alpha,\boldsymbol{\gamma})}{r(\mathbf{h}+\mathbf{k},\alpha,\boldsymbol{\gamma})} = \gamma_{frac}\prod_{i\in\mathfrak{u}}\left(\frac{\max(1,|h_i|)}{\max(1,|h_i+k_i|)}\right)^{\alpha}$$

and prove that for any $h,\ k\in\mathbb{Z}$, we have

$$\frac{\max(1,|h|)}{\max(1,|h+k|)} \le |k|+1\ . \tag{5.62}$$

Indeed, it is obvious if $h=0$ or $h=-k$, otherwise

$$\frac{\max(1,|h|)}{\max(1,|h+k|)} = \frac{|h|}{|h+k|}\ .$$

At last (5.62) is still obvious if h and k have same sign and otherwise,
if $|h|>|k|$ then $|h/(k+h)| = |h|/(|h|-|k|)$ decreases with respect to $|h|$, so $|h/(k+h)|\le|k|+1$
if if $|h|<[k|$ then $|h/(k+h)| = |h|/(|k|-|h|)$ increases with respect to $|h|$, so $|h/(k+h)|\le|k|-1$.

Second part. In the first term in the right-hand side of (5.60), denote $K^c_{\mathfrak{u}+} = (\mathbb{Z}^*_{\mathfrak{u}}\setminus K_{\mathfrak{u}})\cap\mathbb{Z}^d_+$, $I_{\mathfrak{u}} = [1,\beta_{\mathfrak{u}}^{1/|\mathfrak{u}|}]\cap\mathbb{Z}$. Then for any set $\mathfrak{v}\subsetneq\mathfrak{u}$, define

$$Q_{\mathfrak{u},\mathfrak{v}} = \Big\{\mathbf{k}\in K^c_{\mathfrak{u}+},\ \forall i\in\mathfrak{v},\ k_i\in I_{\mathfrak{u}}\ \text{ and }\ \forall i\in\mathfrak{u}\setminus\mathfrak{v},\ k_i\notin I_{\mathfrak{u}}\Big\}$$

and note that

$$K^c_{\mathfrak{u}+} = \bigsqcup_{\mathfrak{v}\subsetneq\mathfrak{u}} Q_{\mathfrak{u},\mathfrak{v}}.$$

Hence denoting $\gamma_{max} = \max_{\mathfrak{u}\subset\{1,\dots,d\}}\gamma_{\mathfrak{u}}$, it comes

$$\begin{aligned}\sum_{\mathbf{k}\in\mathbb{Z}^*_{\mathfrak{u}}\setminus K_{\mathfrak{u}}}\frac{1}{r(\mathbf{k},\alpha,\boldsymbol{\gamma})} &\le 2^{|\mathfrak{u}|}\gamma_{max}\sum_{\mathbf{k}\in K^c_{\mathfrak{u}+}}\prod_{i\in\mathfrak{u}}k_i^{-\alpha}\\ &\le 2^{|\mathfrak{u}|}\gamma_{max}\sum_{\mathfrak{v}\subsetneq\mathfrak{u}}\Big(\sum_{\mathbf{k}\in Q_{\mathfrak{u},\mathfrak{v}}}\prod_{i\in\mathfrak{u}}k_i^{-\alpha}\Big)\end{aligned}$$

and it leads to the proof of (5.42) and (5.43). If $\mathfrak{u}=\{i\}$,the proof is easy since we have

$$\begin{aligned}\sum_{\mathbf{k}\in Q_{\{i\},\emptyset}}k_i^{-\alpha} &= \sum_{k=\lfloor\beta_{\{i\}}+1\rfloor}^{+\infty}k^{-\alpha}\\ &= \sum_{j=0}^{\lfloor\beta_{\{i\}}\rfloor}\sum_{k=1}^{+\infty}(k\lfloor\beta_{\{i\}}+1\rfloor+j)^{-\alpha}\\ &\le \sum_{j=0}^{\lfloor\beta_{\{i\}}\rfloor}\sum_{k=1}^{+\infty}(k\lfloor\beta_{\{i\}}+1\rfloor)^{-\alpha}\\ &\le \zeta(\alpha)\beta_{\{i\}}^{1-\alpha}\end{aligned}\tag{5.63}$$

and the conclusion for (5.42) follows. If $\mathfrak{u}=\{i,j\}$, as in (5.63) it is easy to obtain

$$\sum_{\mathbf{k}\in Q_{\{i,j\},\emptyset}}k_i^{-\alpha}k_j^{-\alpha}\le\frac{\zeta(\alpha)^2}{\beta_{\{i,j\}}^{\alpha-1}}.\tag{5.64}$$

And if $\mathfrak{v}=\{i\}$ or $\{j\}$, in view of (5.63) we have

$$\begin{aligned}\sum_{\mathbf{k}\in Q_{\{i,j\},\mathfrak{v}}}k_i^{-\alpha}k_j^{-\alpha} &= \sum_{k_i=1}^{\lfloor\beta_{\{i,j\}}^{1/2}\rfloor}\sum_{k_j=\beta_{\{i,j\}}/k_i}^{+\infty}k_i^{-\alpha}k_j^{-\alpha}\\ &\le \sum_{k_i=1}^{\lfloor\beta_{\{i,j\}}^{1/2}\rfloor}\frac{\zeta(\alpha)}{\beta_{\{i,j\}}^{\alpha-1}}k_i^{-1}\ .\end{aligned}$$

Then note that the harmonic number $\sum_{k=1}^{M} k^{-1}$ is bounded by $\log(M)+1$ and deduce

$$\sum_{\mathbf{k}\in Q_{\{i,j\},\mathfrak{v}}} k_i^{-\alpha} k_j^{-\alpha} \leq \frac{\zeta(\alpha)}{\beta_{\{i,j\}}^{\alpha-1}} \big(\log(\beta_{\{i,j\}}^{1/2})+1\big) \ . \tag{5.65}$$

Finally, (5.64) and (5.65) gives the conclusion for (5.43)

$$\sum_{\mathbf{k}\in\mathbb{Z}^*_{\{i,j\}}\setminus K_{\{i,j\}}} \frac{1}{r(\mathbf{k},\alpha,\boldsymbol{\gamma})} \leq \frac{4\gamma_{max}\big[\zeta(\alpha)^2 + 2\zeta(\alpha)\big(\log(\beta_{\{i,j\}}^{1/2})+1\big)\big]}{\beta_{\{i,j\}}^{\alpha-1}} \ .$$

5.F Proof of Proposition 5.5

The proof is divided into three parts.

First part. If $f\in\mathcal{H}_\alpha$ then for any $\mathbf{k}\in\mathbb{Z}^d\cap(-\frac{q}{2},\frac{q}{2}]^d$,

$$|\widehat{\mathrm{c}}_{\mathbf{k}}(f)| = |\mathrm{c}_{\mathbf{k}}(f)| + O(q^{-\alpha/2}) \tag{5.66}$$

and consequently

$$|\widehat{\mathrm{c}}_{\mathbf{k}}(f)|^2 = |\mathrm{c}_{\mathbf{k}}(f)|^2 + O(q^{-\alpha/2}) \ . \tag{5.67}$$

Indeed, Poisson summation formula gives

$$|\widehat{\mathrm{c}}_{\mathbf{k}}(f)| - |\mathrm{c}_{\mathbf{k}}(f)| \leq \sum_{\substack{\mathfrak{u}\subseteq\{1,\dots,d\}\\ \mathfrak{u}\neq\emptyset}} \sum_{\mathbf{h}\in\mathbb{Z}^*_{\mathfrak{u}}} |\mathrm{c}_{\mathbf{k}+q\mathbf{h}}(f)|$$

and for any non-empty subset $\mathfrak{u}\subseteq\{1,\dots,d\}$, we have

$$\begin{aligned}
\sum_{\mathbf{h}\in\mathbb{Z}^*_{\mathfrak{u}}} |\mathrm{c}_{\mathbf{k}+q\mathbf{h}}(f)| &\leq ||f||_{\mathcal{H}_\alpha} \sum_{\mathbf{h}\in\mathbb{Z}^*_{\mathfrak{u}}} \prod_{i\in\mathfrak{u}} |k_i+qh_i|^{-\alpha/2} \\
&\leq 2^{|\mathfrak{u}|}||f||_{\mathcal{H}_\alpha} \sum_{h_1=1}^{+\infty}\cdots\sum_{h_{|\mathfrak{u}|}=1}^{+\infty} \prod_{i\in\mathfrak{u}} \Big|qh_i-\frac{q}{2}\Big|^{-\alpha/2} \\
&\leq q^{-|\mathfrak{u}|\alpha/2} 2^{|\mathfrak{u}|(1+\alpha/2)} ||f||_{\mathcal{H}_\alpha} \sum_{h_1=1}^{+\infty}\cdots\sum_{h_{|\mathfrak{u}|}=1}^{+\infty} \prod_{i\in\mathfrak{u}} |2h_i-1|^{-\alpha/2} \\
&\leq q^{-|\mathfrak{u}|\alpha/2} 2^{|\mathfrak{u}|(1+\alpha/2)} \zeta\Big(\frac{\alpha}{2}\Big)^{|\mathfrak{u}|} ||f||_{\mathcal{H}_\alpha} \ .
\end{aligned}$$

Second part. Recall that $\{0,\frac{1}{q},\dots\frac{q-1}{q}\}^d$ is denoted by $D(q)$. First we have

$$\begin{aligned}
\mathbb{E}_\mu\Big[\widehat{\mathrm{c}}_{\mathbf{0}}\big(f,A(\boldsymbol{\pi})\big)\Big] &= \frac{1}{|\Pi|}\sum_{\boldsymbol{\pi}\in\Pi}\bigg(\frac{1}{n}\sum_{i=1}^{n} f\big((A(\boldsymbol{\pi}))_{i1},\dots,(A(\boldsymbol{\pi}))_{id}\big)\bigg) \\
&= \frac{1}{n}\sum_{i=1}^{n}\bigg(\frac{1}{|\Pi|}\sum_{\pi\in\Pi} f\big((A(\boldsymbol{\pi}))_{i1},\dots,(A(\boldsymbol{\pi}))_{id}\big)\bigg) \\
&= \frac{1}{n}\sum_{i=1}^{n}\bigg(\frac{1}{q^d}\sum_{\mathbf{x}\in D(q)} f(\mathbf{x})\bigg) \\
&= \frac{1}{q^d}\sum_{\mathbf{x}\in D(q)} f(\mathbf{x}) \ .
\end{aligned}$$

Thus, we deduce

$$\begin{aligned}
\mathbb{E}_\mu\Big[\widehat{\mathrm{V}}\big(f,A(\boldsymbol{\pi})\big)\Big] &= \mathbb{E}_\mu\Big[\widehat{\mathrm{c}}_{\mathbf{0}}\big(f^2,A(\boldsymbol{\pi})\big)-\widehat{\mathrm{c}}_{\mathbf{0}}\big(f,A(\boldsymbol{\pi})\big)^2\Big]\\
&= \mathbb{E}_\mu\Big[\widehat{\mathrm{c}}_{\mathbf{0}}\big(f^2,A(\boldsymbol{\pi})\big)\Big]-\mathbb{E}_\mu\Big[\widehat{\mathrm{c}}_{\mathbf{0}}\big(f,A(\boldsymbol{\pi})\big)\Big]^2-\mathrm{Var}_\mu\Big[\widehat{\mathrm{c}}_{\mathbf{0}}\big(f,A(\boldsymbol{\pi})\big)\Big]\\
&= \frac{1}{q^d}\sum_{\mathbf{x}\in D(q)}f(\mathbf{x})^2-\Big(\frac{1}{q^d}\sum_{\mathbf{x}\in D(q)}f(\mathbf{x})\Big)^2-\mathrm{Var}_\mu\Big[\widehat{\mathrm{c}}_{\mathbf{0}}\big(f,A(\boldsymbol{\pi})\big)\Big]\\
&= \mathrm{V}(f)+\widehat{\mathrm{c}}_{\mathbf{0}}(f^2)-\mathrm{c}_{\mathbf{0}}(f^2)+\mathrm{c}_{\mathbf{0}}(f)^2-\widehat{\mathrm{c}}_{\mathbf{0}}(f)^2-\mathrm{Var}_\mu\Big[\widehat{\mathrm{c}}_{\mathbf{0}}\big(f,A(\boldsymbol{\pi})\big)\Big]. \qquad (5.68)
\end{aligned}$$

We conclude from (5.66) and (5.67)

$$\mathbb{E}_\mu\Big[\widehat{\mathrm{V}}\big(f,A(\boldsymbol{\pi})\big)\Big]=\mathrm{V}(f)-\mathrm{Var}_\mu\Big[\widehat{\mathrm{c}}_{\mathbf{0}}\big(f,A(\boldsymbol{\pi})\big)\Big]+O\big(q^{-\alpha/2}\big)\ . \qquad (5.69)$$

Third part. From Theorem 5.2, we have

$$\begin{aligned}
\mathrm{Var}_\mu\Big[\widehat{\mathrm{c}}_{\mathbf{0}}\big(f,A(\boldsymbol{\pi})\big)\Big] &= \frac{1}{n}\sum_{|\mathfrak{u}|\geq 1}\sum_{\mathbf{k}\in\mathbb{Z}^*_{\mathfrak{u}}(q)}\big|\widehat{\mathrm{c}}_{\mathbf{k}}(f)\big|^2-\frac{1}{n}\sum_{1\leq|\mathfrak{u}|\leq t}\sum_{\mathbf{k}\in\mathbb{Z}^*_{\mathfrak{u}}(q)}\big|\widehat{\mathrm{c}}_{\mathbf{k}}(f)\big|^2\\
&\quad+\frac{1}{n^2}\sum_{|\mathfrak{v}|>t}\Big(-n+\sum_{r=0}^{|\mathfrak{v}|}B(\mathfrak{v},r)(1-q)^{r-|\mathfrak{v}|}\Big)\sum_{\mathbf{k}\in\mathbb{Z}^*_{\mathfrak{v}}(q)}\big|\widehat{\mathrm{c}}_{\mathbf{k}}(f)\big|^2\ . \qquad (5.70)
\end{aligned}$$

And we now detail the three terms on the right-hand side of (5.70) :
i) the first term is

$$\frac{1}{n}\widehat{\mathrm{V}}(f,D(q))=\frac{1}{n}\Big(\frac{1}{q^d}\sum_{\mathbf{x}\in D(q)}f(\mathbf{x})^2-\Big(\frac{1}{q^d}\sum_{\mathbf{x}\in D(q)}f(\mathbf{x})\Big)^2\Big)$$

and is equal to $\frac{1}{n}\big(\mathrm{V}(f)+O(q^{-\alpha/2})\big)$ (see (5.68) and (5.69)).
ii) the second term can be rewritten

$$-\frac{1}{n}\sum_{1\leq|\mathfrak{u}|\leq t}\Big(\mathrm{V}_{\mathfrak{u}}(f)+\varepsilon_{integ}(\mathfrak{u})+\varepsilon_{trunc}(\mathfrak{u})\Big)$$

where, from (5.67), we have

$$\begin{aligned}
\frac{1}{n}\varepsilon_{integ}(\mathfrak{u}) &= \frac{1}{n}\sum_{\mathbf{k}\in\mathbb{Z}^*_{\mathfrak{u}}(q)}\Big(\big|\widehat{\mathrm{c}}_{\mathbf{k}}(f)\big|^2-\big|\mathrm{c}_{\mathbf{k}}(f)\big|^2\Big)\\
&\leq \frac{1}{n}(q-1)^{|\mathfrak{u}|}O(q^{-\alpha/2})\\
&\leq q^{-t}q^tO(q^{-\alpha/2})\\
&= O(q^{-\alpha/2})
\end{aligned}$$

and letting for any $\mathfrak{v}\subseteq\mathfrak{u}$,

$$Q'_{\mathfrak{u},\mathfrak{v}}=\{\mathbf{k}\in\mathbb{Z}^*_{\mathfrak{u}},\ \forall i\in\mathfrak{v},\ 1\leq k_i\leq\frac{q}{2},\ \forall i\in\mathfrak{u}\setminus\mathfrak{v},\ k_i\geq\frac{q}{2}\}$$

we have from (5.63)

$$\begin{aligned}
\frac{1}{n}\varepsilon_{trunc}(\mathfrak{u}) &= \frac{1}{n}\sum_{\mathbf{k}\in\mathbb{Z}^*_{\mathfrak{u}}\setminus\mathbb{Z}^*_{\mathfrak{u}}(q)}\big|\mathrm{c}_{\mathbf{k}}(f)\big|^2 \\
&\le \frac{2^{|\mathfrak{u}|}}{n}||f||^2_{\mathcal{H}_\alpha}\sum_{\mathfrak{v}\subsetneq\mathfrak{u}}\sum_{\mathbf{k}\in Q'_{\mathfrak{u},\mathfrak{v}}}\prod_{i\in\mathfrak{u}}k_i^{-\alpha} \\
&\le \frac{2^{|\mathfrak{u}|}}{n}||f||^2_{\mathcal{H}_\alpha}\sum_{\mathfrak{v}\subsetneq\mathfrak{u}}\Big(\frac{q}{2}\Big)^{|\mathfrak{v}|}\Big(\sum_{k\ge\frac{q}{2}}k^{-\alpha}\Big)^{|\mathfrak{u}|-|\mathfrak{v}|} \\
&\le \frac{2^{|\mathfrak{u}|}}{n}||f||^2_{\mathcal{H}_\alpha}\sum_{\mathfrak{v}\subsetneq\mathfrak{u}}\Big(\frac{q}{2}\Big)^{|\mathfrak{v}|}\Big(\zeta(\alpha)\Big(\frac{q}{2}\Big)^{1-\alpha}\Big)^{|\mathfrak{u}|-|\mathfrak{v}|} \\
&\le \frac{\big(2\zeta(\alpha)\big)^{|\mathfrak{u}|}}{n}||f||^2_{\mathcal{H}_\alpha}\sum_{\mathfrak{v}\subsetneq\mathfrak{u}}\Big(\frac{q}{2}\Big)^{|\mathfrak{u}|-1}\Big(\frac{q}{2}\Big)^{1-\alpha} \\
&\le \frac{\big(2\zeta(\alpha)\big)^{|\mathfrak{u}|}}{\lambda q^t}||f||^2_{\mathcal{H}_\alpha}(2^{|\mathfrak{u}|}-1)\Big(\frac{q}{2}\Big)^{t-\alpha} \\
&= O(q^{-\alpha}),
\end{aligned}$$

iii) for the third term, note that, since A is defect-free, for all $\mathfrak{v} > t$, $B(\mathfrak{v},|\mathfrak{v}|) = n$ and for all $i \ge 1$, $B(\mathfrak{v},t+i) = 0$. Then it comes

$$\begin{aligned}
&\frac{1}{n^2}\sum_{|\mathfrak{v}|>t}\Big(-n+\sum_{r=0}^{|\mathfrak{v}|}B(\mathfrak{v},r)(1-q)^{r-|\mathfrak{v}|}\Big)\sum_{\mathbf{k}\in\mathbb{Z}^*_{\mathfrak{v}}(q)}\big|\widehat{\mathrm{c}}_{\mathbf{k}}(f)\big|^2 \\
&\le \frac{1}{n^2}\sum_{|\mathfrak{v}|>t}\sum_{r=0}^{t}B(\mathfrak{v},r)(q-1)^{r-|\mathfrak{v}|}\sum_{\mathbf{k}\in\mathbb{Z}^*_{\mathfrak{v}}(q)}\Big(\big|\mathrm{c}_{\mathbf{k}}(f)\big|^2+O\big(q^{-\alpha/2}\big)\Big) \\
&\le \frac{1}{n^2}\sum_{|\mathfrak{v}|>t}\sum_{r=0}^{t}B(\mathfrak{v},r)(q-1)^{r-|\mathfrak{v}|}\Big(O(1)+O\big(q^{|\mathfrak{v}|-\alpha/2}\big)\Big) \\
&\le \frac{1}{n^2}\sum_{|\mathfrak{v}|>t}\sum_{r=0}^{t}B(\mathfrak{v},r)(q-1)^{r}\Big(O(q^{-|\mathfrak{v}|})+O\big(q^{-\alpha/2}\big)\Big) \\
&\le O\big(q^{-\min(t+1,\alpha/2)}\big)\frac{1}{n^2}\sum_{|\mathfrak{v}|>t}\sum_{r=0}^{t}B(\mathfrak{v},r)(q-1)^{r} \\
&\le O\big(q^{-\min(t+1,\alpha/2)}\big) \qquad (5.71)
\end{aligned}$$

since for all $r \le t < |\mathfrak{v}|$, $B(\mathfrak{v},r) \le \binom{|\mathfrak{v}|}{r}n^2q^{-r}$. Indeed, consider

$$B'(\mathfrak{v},r) = \sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{1}_{|\{l\in\mathfrak{v},\ A_{il}=A_{jl}\}|\ge r}\ ,$$

we have $B(\mathfrak{v},r) \le B'(\mathfrak{v},r)$ and it easy to prove that

$$B'(\mathfrak{v},t) = B(\mathfrak{v},t) = \binom{|\mathfrak{v}|}{t}n(nq^{-t}-1)$$

and to deduce that for all $r < t$

$$B'(\mathfrak{v},r) \le \binom{|\mathfrak{v}|}{r}n(nq^{-r}-1)\ .$$

The conclusion follows.

5.G Proof of Proposition 5.6

The proof is divided into three parts.

First part. For any complex-valued random variable Z, define

$$\begin{aligned}\mathrm{Var}[Z] &= \mathbb{E}\Big[\big|Z-\mathbb{E}[Z]\big|^2\Big]\\ &= \mathbb{E}\big[|Z|^2\big]-\big|\mathbb{E}[Z]\big|^2.\end{aligned}$$

Hence, note that $\mathbb{E}_\mu\big[\widehat{\mathrm{c}}_{\mathbf{k}}(f,A(\boldsymbol{\pi}))\big]=\widehat{\mathrm{c}}_{\mathbf{k}}(f)$ and deduce

$$\begin{aligned}\mathbb{E}_\mu\Big[\big|\widehat{\mathrm{c}}_{\mathbf{k}}(f,A(\boldsymbol{\pi}))\big|^2\Big] &= \Big|\mathbb{E}_\mu\big[\widehat{\mathrm{c}}_{\mathbf{k}}(f,A(\boldsymbol{\pi}))\big]\Big|^2+\mathrm{Var}_\mu\big[\widehat{\mathrm{c}}_{\mathbf{k}}(f,A(\boldsymbol{\pi}))\big]\\ &= \big|\widehat{\mathrm{c}}_{\mathbf{k}}(f)\big|^2+\mathrm{Var}_\mu\big[\widehat{\mathrm{c}}_{\mathbf{k}}(f,A(\boldsymbol{\pi}))\big]\\ &= \big|\mathrm{c}_{\mathbf{k}}(f)\big|^2+\mathrm{Var}_\mu\big[\widehat{\mathrm{c}}_{\mathbf{k}}(f,A(\boldsymbol{\pi}))\big]+O\big(q^{-\alpha/2}\big)\end{aligned}$$

where, from Theorem 5.2, we have

$$\begin{aligned}\mathrm{Var}_\mu\big[\widehat{\mathrm{c}}_{\mathbf{k}}(f,A(\boldsymbol{\pi}))\big] &= \frac{1}{n}\sum_{|\mathfrak{v}|\geq 1}\sum_{\mathbf{h}\in\mathbb{Z}^*_{\mathfrak{v}}(q)}\big|\widehat{\mathrm{c}}_{\mathbf{k}+\mathbf{h}}(f)\big|^2-\frac{1}{n}\sum_{1\leq|\mathfrak{v}|\leq t}\sum_{\mathbf{h}\in\mathbb{Z}^*_{\mathfrak{v}}(q)}\big|\widehat{\mathrm{c}}_{\mathbf{k}+\mathbf{h}}(f)\big|^2\\ &+ \frac{1}{n^2}\sum_{|\mathfrak{v}|>t}\Big(-n+\sum_{r=0}^{|\mathfrak{v}|}B(\mathfrak{v},r)(1-q)^{r-|\mathfrak{v}|}\Big)\sum_{\mathbf{h}\in\mathbb{Z}^*_{\mathfrak{v}}(q)}\big|\widehat{\mathrm{c}}_{\mathbf{k}+\mathbf{h}}(f)\big|^2. \qquad (5.72)\end{aligned}$$

Denote T_1, T_2 and T_3 the three successive terms on the right-hand side of (5.72). T_3 is given by (5.71) in the proof of Proposition 5.6, and both the other terms are studied in the next parts.

Second part (*details for* T_1). Note that for any $\mathfrak{u}\subseteq\{1,\dots,d\}$ and any $\mathbf{k}\in\mathbb{Z}_{\mathfrak{u}}$,

$$\sum_{\mathbf{h}\in\mathbb{Z}_{\mathfrak{u}}(q)}\big|\widehat{\mathrm{c}}_{\mathbf{k}+\mathbf{h}}(f)\big|^2=\sum_{\mathbf{h}\in\mathbb{Z}_{\mathfrak{u}}(q)}\big|\widehat{\mathrm{c}}_{\mathbf{h}}(f)\big|^2. \qquad (5.73)$$

Indeed, consider

$$\begin{array}{rccc}\Phi_{\mathbf{k}}: & \mathbb{Z}_{\mathfrak{u}}(q) & \longrightarrow & \mathbb{Z}_{\mathfrak{u}}(q)\\ & \mathbf{h} & \longmapsto & \mathbf{h}'\end{array}$$

where for all $i\notin\mathfrak{u}$, $h'_i=0$, and for $i\in\mathfrak{u}$, h'_i is the remainder in $(-\frac{q}{2},\frac{q}{2}]$ of the division of h_i+k_i by q. Then, note that

$$\forall\mathbf{h}\in\mathbb{Z}_{\mathfrak{u}}(q),\ \exists\mathbf{l}_0\in\mathbb{Z}_{\mathfrak{u}},\ \mathbf{k}+\mathbf{h}=\Phi_{\mathbf{k}}(\mathbf{h})+q\mathbf{l}_0\ .$$

Hence, by Poisson summation formula, we have

$$\begin{aligned}\widehat{\mathrm{c}}_{\Phi_{\mathbf{k}}(\mathbf{h})}(f) &= \sum_{\mathbf{l}\in\mathbb{Z}^d}\mathrm{c}_{\Phi_{\mathbf{k}}(\mathbf{h})+q\mathbf{l}}(f)\\ &= \sum_{\mathbf{l}\in\mathbb{Z}^d}\mathrm{c}_{\mathbf{k}+\mathbf{h}+q(\mathbf{l}-\mathbf{l}_0)}(f)\\ &= \widehat{\mathrm{c}}_{\mathbf{k}+\mathbf{h}}(f).\end{aligned}$$

Finally, noting that $\Phi_{\mathbf{k}}$ is bijective, we conclude to (5.73). Then it comes

$$\begin{aligned}
T_1 &= \frac{1}{n}\sum_{|\mathfrak{v}|\geq 1}\sum_{\mathbf{h}\in\mathbb{Z}^*_{\mathfrak{v}}(q)}\left|\widehat{\mathrm{c}}_{\mathbf{k}+\mathbf{h}}(f)\right|^2 \\
&= \frac{1}{n}\Bigg(\sum_{\mathbf{h}\in\mathbb{Z}_{\{1,\dots,d\}}(q)}\left|\widehat{\mathrm{c}}_{\mathbf{k}+\mathbf{h}}(f)\right|^2-\left|\widehat{\mathrm{c}}_{\mathbf{k}}(f)\right|^2\Bigg) \\
&= \frac{1}{n}\Bigg(\sum_{\mathbf{h}\in\mathbb{Z}_{\{1,\dots,d\}}(q)}\left|\widehat{\mathrm{c}}_{\mathbf{h}}(f)\right|^2-\left|\widehat{\mathrm{c}}_{\mathbf{k}}(f)\right|^2\Bigg) \\
&= \frac{1}{n}\Big(\widehat{\mathrm{V}}(f,D(q))+\widehat{\mathrm{c}}_{\mathbf{0}}(f)^2-\left|\widehat{\mathrm{c}}_{\mathbf{k}}(f)\right|^2\Big) \\
&= \frac{1}{n}\Big(\mathrm{V}(f)+\mathrm{c}_{\mathbf{0}}(f)^2-\left|\mathrm{c}_{\mathbf{k}}(f)\right|^2\Big)+O\big(q^{-\alpha/2-t}\big).
\end{aligned}$$

Third part (*details for* T_2). We have

$$\begin{aligned}
T_2 &= -\frac{1}{n}\sum_{1\leq|\mathfrak{v}|\leq t}\sum_{\mathbf{h}\in\mathbb{Z}^*_{\mathfrak{v}}(q)}\left|\widehat{\mathrm{c}}_{\mathbf{k}+\mathbf{h}}(f)\right|^2 \\
&= -\frac{1}{n}\sum_{\substack{1\leq|\mathfrak{v}|\leq t\\ \mathfrak{u}\cap\mathfrak{v}=\emptyset}}\sum_{\mathbf{h}\in\mathbb{Z}^*_{\mathfrak{v}}(q)}\Big(\left|\mathrm{c}_{\mathbf{k}+\mathbf{h}}(f)\right|^2+O\big(q^{-\alpha/2}\big)\Big)-\frac{1}{n}\sum_{\substack{1\leq|\mathfrak{v}|\leq t\\ \mathfrak{u}\cap\mathfrak{v}\neq\emptyset}}\sum_{\mathbf{h}\in\mathbb{Z}^*_{\mathfrak{v}}(q)}\left|\widehat{\mathrm{c}}_{\mathbf{k}+\mathbf{h}}(f)\right|^2 \\
&= -\frac{1}{n}\sum_{\substack{1\leq|\mathfrak{v}|\leq t\\ \mathfrak{u}\cap\mathfrak{v}=\emptyset}}\sum_{\mathbf{h}\in\mathbb{Z}^*_{\mathfrak{v}}(q)}\left|\mathrm{c}_{\mathbf{k}+\mathbf{h}}(f)\right|^2-\frac{1}{n}\sum_{\substack{1\leq|\mathfrak{v}|\leq t\\ \mathfrak{u}\cap\mathfrak{v}\neq\emptyset}}\sum_{\mathbf{h}\in\mathbb{Z}^*_{\mathfrak{v}}(q)}\left|\widehat{\mathrm{c}}_{\mathbf{k}+\mathbf{h}}(f)\right|^2+O(q^{-\alpha/2})\ .
\end{aligned}$$

The first term on the right-hand side is $-R_1(q,t,\lambda,\mathbf{k})/n$ in Proposition 5.7. The second one, that we denote $R_2'(q,t\lambda,\mathbf{k})$, consists of the sum of $-R_2(q,t,\lambda,\mathbf{k})/n$ and an error term of order $O\big(q^{-\alpha/2}\big)$. Indeed, by an application of the Möbius inversion formula (see e.g. [Sta12]), we have

$$R_2'(q,t\lambda,\mathbf{k})=-\frac{1}{n}\sum_{\substack{1\leq|\mathfrak{v}|\leq t\\ \mathfrak{u}\cap\mathfrak{v}\neq\emptyset}}\sum_{\mathfrak{v}'\subseteq\mathfrak{v}}(-1)^{|\mathfrak{v}|-|\mathfrak{v}'|}\sum_{\mathbf{h}\in\mathbb{Z}_{\mathfrak{v}'}(q)}\left|\widehat{\mathrm{c}}_{\mathbf{k}+\mathbf{h}}(f)\right|^2\ .$$

Now note that (5.73) can be generalized as follows

$$\forall\mathbf{k}\in\mathbb{Z}^d,\quad\sum_{\mathbf{h}\in\mathbb{Z}_{\mathfrak{u}}(q)}\left|\widehat{\mathrm{c}}_{\mathbf{k}+\mathbf{h}}(f)\right|^2=\sum_{\mathbf{h}\in\mathbb{Z}_{\mathfrak{u}}(q)}\left|\widehat{\mathrm{c}}_{\mathbf{k}_{\mathfrak{u}}+\mathbf{h}}(f)\right|^2$$

where we recall that $(\mathbf{k}_{\mathfrak{u}})_i=0$ if $i\in\mathfrak{u}$, and $(\mathbf{k}_{\mathfrak{u}})_i=k_i$ otherwise. Then it comes

$$\begin{aligned}
R_2'(q,t\lambda,\mathbf{k}) &= -\frac{1}{n}\sum_{\substack{1\leq|\mathfrak{v}|\leq t\\ \mathfrak{u}\cap\mathfrak{v}\neq\emptyset}}\sum_{\mathfrak{v}'\subseteq\mathfrak{v}}(-1)^{|\mathfrak{v}|-|\mathfrak{v}'|}\sum_{\mathbf{h}\in\mathbb{Z}_{\mathfrak{v}'}(q)}\left|\widehat{\mathrm{c}}_{\mathbf{k}_{\mathfrak{v}'}+\mathbf{h}}(f)\right|^2 \\
&= -\frac{1}{n}\sum_{\substack{1\leq|\mathfrak{v}|\leq t\\ \mathfrak{u}\cap\mathfrak{v}\neq\emptyset}}\sum_{\mathfrak{v}'\subseteq\mathfrak{v}}(-1)^{|\mathfrak{v}|-|\mathfrak{v}'|}\sum_{\mathfrak{v}''\subseteq\mathfrak{v}'}\sum_{\mathbf{h}\in\mathbb{Z}^*_{\mathfrak{v}''}(q)}\left|\widehat{\mathrm{c}}_{\mathbf{k}_{\mathfrak{v}'}+\mathbf{h}}(f)\right|^2 \\
&= O\big(q^{-\alpha/2}\big)-\frac{1}{n}\sum_{\substack{1\leq|\mathfrak{v}|\leq t\\ \mathfrak{u}\cap\mathfrak{v}\neq\emptyset}}\sum_{\mathfrak{v}'\subseteq\mathfrak{v}}(-1)^{|\mathfrak{v}|-|\mathfrak{v}'|}\sum_{\mathfrak{v}''\subseteq\mathfrak{v}'}\sum_{\mathbf{h}\in\mathbb{Z}^*_{\mathfrak{v}''}(q)}\left|\mathrm{c}_{\mathbf{k}_{\mathfrak{v}'}+\mathbf{h}}(f)\right|^2\ .
\end{aligned}$$

Chapitre 6

Calcul des indices de Sobol' en combinant estimateurs Monte Carlo et hypercubes latins

Dans ce chapitre, nous proposons de combiner l'estimateur de la méthode de Sobol' avec des plans d'expériences particuliers afin de rendre le coût de cette méthode indépendant de la dimension d de la fonction étudiée. Les plans d'expériences considérés sont des hypercubes latins répliqués [McK95] et leurs versions généralisées à des tableaux orthogonaux de force quelconque. À l'aide de cette technique, pour tout $k \leq d$, il devient possible d'estimer tous les indices de Sobol' descendants d'ordre k à l'aide de seulement 2 échantillons. Ce chapitre a fait l'objet d'une soumission dans la revue Technometrics [TP12b].

6.1 Introduction and notation

Sobol' indices are quantities defined by normalizing parts of variance in an ANOVA decomposition. They allow to quantify the relative importance of input factors of a function over their entire range of values. They essentially consist of integrals and as a consequence, their computation can become rapidly expensive when the number of factors increases. Many techniques have been developed to estimate these indices including Fast Amplitude Sensitivity Test (FAST) [CLS78, STC99] and Random Balance Design (RBD) [TGM06] — for a recent survey see [TP12c] —, polynomial chaos expansion-based estimators [Sud08, BS10] and the Sobol' Pick-freeze (SPF) scheme [Sob93] (see also [SRA+08] for a review).

Let f be a real square integrable function defined on the unit hypercube $[0,1]^d$ and $\mathbf{X} = (X_1, \ldots, X_d)$ a random vector with independent components uniformly distributed on $[0,1]$. We consider the real random variable $Y = f(\mathbf{X})$. Note that this framework can be generalized to independent arbitrary marginal distributions $(X_i)_{i=1..d}$ by using the inverse transformation method. Then for any $\mathfrak{u} \subseteq \{1, \ldots, d\}$, denote $\mathbf{X}_\mathfrak{u}$ the random vector with components X_i, $i \in \mathfrak{u}$. The ANOVA decomposition [Hoe48, ES81] states that $Y = f(\mathbf{X})$ can be uniquely decomposed into summands of increasing dimensions

$$f(\mathbf{X}) = \sum_{\mathfrak{u} \subseteq \{1,\ldots,d\}} f_\mathfrak{u}(\mathbf{X}_\mathfrak{u}) \tag{6.1}$$

where $f_\emptyset = \mathbb{E}[Y]$ and the other components have mean zero and are mutually uncorrelated. In particular, the sum of functions

$$f_\emptyset + f_1(X_1) + f_2(X_2) + \cdots + f_d(X_d) \tag{6.2}$$

is the so-called additive part of f.

The Sobol' index with respect to the combination of all the variables in $\mathfrak{u} \subseteq \{1,\dots,d\}$ (see [Sob93]) is then defined as

$$S_{\mathfrak{u}} = \frac{\sigma^2_{\mathfrak{u}}}{\sigma^2} = \frac{\mathrm{Var}\big[f_{\mathfrak{u}}(\mathbf{x}_{\mathfrak{u}})\big]}{\mathrm{Var}[Y]}$$

and the Sobol' index with respect to a subset of variables $\mathfrak{u} \subseteq \{1,\dots,d\}$ (see [HS96]) is then defined as

$$\underline{S}_{\mathfrak{u}} = \frac{\underline{\tau}^2_{\mathfrak{u}}}{\sigma^2} = \frac{\mathrm{Var}\big[\sum_{\mathfrak{v}\subseteq\mathfrak{u}} f_{\mathfrak{v}}(\mathbf{x}_{\mathfrak{v}})\big]}{\mathrm{Var}[Y]} \ .$$

In practice, global sensitivity analysis focuses on the first-order — i.e. $|\mathfrak{u}| = 1$ — and the second-order — i.e. $|\mathfrak{u}| = 2$ — terms. Note that, thanks to the properties of the ANOVA decomposition, we have

$$\underline{S}_{\mathfrak{u}} = \sum_{\mathfrak{v}\subseteq\mathfrak{u}} S_{\mathfrak{v}}$$

and the Möbius inversion formula (see e.g. [Sta12]) gives

$$S_{\mathfrak{u}} = \sum_{\mathfrak{v}\subseteq\mathfrak{u}} (-1)^{|\mathfrak{u}|-|\mathfrak{v}|} \underline{S}_{\mathfrak{v}} \ .$$

Concerning notation, when integrals are over a unit hypercube $[0,1]^s$, $s \leq d$, the integration set is generally omitted, and for any $\mathfrak{u} \subseteq \{1,\dots,d\}$, we denote by $\mathfrak{u}^c = \{1,\dots,d\}\backslash\mathfrak{u}$ the relative complement of $\mathfrak{u}$ with respect to $\{1,\dots,d\}$.

Section 6.2 provides a short review of Monte Carlo estimators of Sobol' indices and gives some notation. In Section 6.3 we explain how to combine Monte Carlo estimators and Latin hypercube sampling — see [MCB79] — in a basic way, and we give asymptotic and bias properties of such a technique. In Section 6.4 we study the method introduced in Section 6.3 using replicated Latin hypercubes — see [McK95] —, we give asymptotic and bias properties of this technique and we explain how it allows to compute all the first-order Sobol' indices using only two replicated Latin hypercubes. Potential generalization to orthogonal array-based Latin hypercubes — see [Owe92] — is also discussed in this section. Numerical illustrations are provided in Section 6.5, and Section 6.6 has conclusions. Note that technical lemmas are given in the Appendix.

6.2 Review of Monte Carlo estimators

6.2.1 Notation

Let $\mathfrak{u}$ be a non-empty subset of $\{1,\dots d\}$, and j in $\{1,\dots n\}$, and consider

$$\mathbf{Z}^j_{\mathfrak{u}} = (X^j_1, \dots X^j_{2d-|\mathfrak{u}|})$$

where the X^j_i's are independent random variables uniformly distributed on $[0,1]$. We also denote

$$\begin{aligned}
\mathbf{X}^j_{\mathfrak{u}} &= (X^j_1, \dots, X^j_{|\mathfrak{u}|}) \\
\mathbf{X}^{j,1}_{\mathfrak{u}^c} &= (X^j_{|\mathfrak{u}|+1}, \dots, X^j_d) \\
\mathbf{X}^{j,2}_{\mathfrak{u}^c} &= (X^j_{d+1}, \dots, X^j_{2d-|\mathfrak{u}|})
\end{aligned}$$

so that

$$\mathbf{Z}^j_{\mathfrak{u}} = (\mathbf{X}^j_{\mathfrak{u}}, \mathbf{X}^{j,1}_{\mathfrak{u}^c}, \mathbf{X}^{j,2}_{\mathfrak{u}^c}) \ .$$

Finally, for $k = 1$ and 2, consider

$$Y^{j,k}_{\mathfrak{u}} = f(\mathbf{X}^j_{\mathfrak{u}}, \mathbf{X}^{j,k}_{\mathfrak{u}^c}) \ . \tag{6.3}$$

With this notation, we consider two estimators of the Sobol' indices $\underline{S}_{\mathfrak{u}}$, introduced in [HS96] and [MNM06], which are functions of $(\mathbf{Z}_{\mathfrak{u}}^j)_{j=1..n}$. They are defined by

$$\widetilde{\underline{S}}_{\mathfrak{u},n} = \frac{\widetilde{\underline{\tau}}^2_{\mathfrak{u},n}}{\widetilde{\sigma}^2_n} = \frac{\frac{1}{n}\sum_{j=1}^n Y_{\mathfrak{u}}^{j,1}Y_{\mathfrak{u}}^{j,2} - \left(\frac{1}{n}\sum_{j=1}^n Y_{\mathfrak{u}}^{j,1}\right)\left(\frac{1}{n}\sum_{j=1}^n Y_{\mathfrak{u}}^{j,2}\right)}{\frac{1}{n}\sum_{j=1}^n \left(Y_{\mathfrak{u}}^{j,1}\right)^2 - \left(\frac{1}{n}\sum_{j=1}^n Y_{\mathfrak{u}}^{j,1}\right)^2} \tag{6.4}$$

and

$$\widehat{\underline{S}}_{\mathfrak{u},n} = \frac{\widehat{\underline{\tau}}^2_{\mathfrak{u},n}}{\widehat{\sigma}^2_n} = \frac{\frac{1}{n}\sum_{j=1}^n Y_{\mathfrak{u}}^{j,1}Y_{\mathfrak{u}}^{j,2} - \left(\frac{1}{2n}\sum_{j=1}^n Y_{\mathfrak{u}}^{j,1} + Y_{\mathfrak{u}}^{j,2}\right)^2}{\frac{1}{2n}\sum_{j=1}^n \left(\left(Y_{\mathfrak{u}}^{j,1}\right)^2 + \left(Y_{\mathfrak{u}}^{j,2}\right)^2\right) - \left(\frac{1}{2n}\sum_{j=1}^n Y_{\mathfrak{u}}^{j,1} + Y_{\mathfrak{u}}^{j,2}\right)^2}\,, \tag{6.5}$$

respectively. Note that other Monte Carlo estimators exist (for a recent review, see [Owe12c]).

6.2.2 Statistical properties of the estimators

Asymptotic properties of both the estimators introduced in the previous section are detailed in [JKL+12]. $\widetilde{\underline{S}}_{\mathfrak{u},n}$ and $\widehat{\underline{S}}_{\mathfrak{u},n}$ are strongly consistent and asymptotically normal estimators, and $\widehat{\underline{S}}_{\mathfrak{u},n}$ is, in addition, asymptotically efficient in some sense (see details in Proposition 2.5 in [JKL+12]). Concerning the biases, it is easy to show that

$$\begin{aligned}\mathbb{E}\left[\widetilde{\underline{\tau}}^2_{\mathfrak{u},n}\right] &= \underline{\tau}^2_{\mathfrak{u}} - \frac{1}{n}\underline{\tau}^2_{\mathfrak{u}}\\ \mathbb{E}\left[\widetilde{\sigma}^2_n\right] &= \sigma^2 - \frac{1}{n}\sigma^2\end{aligned}$$

and — see e.g. [Owe12c] —

$$\begin{aligned}\mathbb{E}\left[\widehat{\underline{\tau}}^2_{\mathfrak{u},n}\right] &= \underline{\tau}^2_{\mathfrak{u}} - \frac{1}{2n}(\sigma^2 + \underline{\tau}^2_{\mathfrak{u}})\\ \mathbb{E}\left[\widehat{\sigma}^2_n\right] &= \sigma^2 - \frac{1}{2n}(\sigma^2 + \underline{\tau}^2_{\mathfrak{u}})\end{aligned}$$

but as far as we know, there is no result on the global biases of $\widetilde{\underline{S}}_{\mathfrak{u},n}$ and $\widehat{\underline{S}}_{\mathfrak{u},n}$.

6.3 Monte Carlo estimators and Latin hypercube sampling

6.3.1 Notation and definitions

We begin with the definition of a Latin hypercube :

Definition 6.1. *Let d and n in $\mathbb{N}^*$, and consider Π_n the set of all the permutations of $\{1,\dots,n\}$. We say that $(\mathbf{X}^j)_{j=1..n}$ is a Latin hypercube of size n in $[0,1]^d$ — and we denote $(\mathbf{X}^j)_j \sim \mathcal{LH}(n,d)$ — if for all $j \in \{1,\dots,n\}$,*

$$\mathbf{X}^j = \left(\frac{\pi_1(j) - U_{1,\pi_1(j)}}{n}, \dots, \frac{\pi_d(j) - U_{d,\pi_d(j)}}{n}\right)$$

where the π_i's and the $U_{i,j}$'s are independent random variables uniformly distributed on Π_n and $[0,1]$, respectively.

Now let $\mathfrak{u}$ be a non-empty subset of $\{1,\dots d\}$, and j in $\{1,\dots n\}$, and consider

$$\dot{\mathbf{Z}}_{\mathfrak{u}}^j = (\dot{X}_1^j, \dots, \dot{X}_{2d-|\mathfrak{u}|}^j) \tag{6.6}$$

such that $(\dot{\mathbf{Z}}_{\mathrm{u}}^j)_j \sim \mathcal{LH}(n, 2d-|\mathrm{u}|)$ and denote

$$\begin{aligned}\dot{\mathbf{X}}_{\mathrm{u}}^j &= (\dot{X}_1^j, \ldots, \dot{X}_{|\mathrm{u}|}^j)\\ \dot{\mathbf{X}}_{\mathrm{u}^c}^{j,1} &= (\dot{X}_{|\mathrm{u}|+1}^j, \ldots, \dot{X}_d^j)\\ \dot{\mathbf{X}}_{\mathrm{u}^c}^{j,2} &= (\dot{X}_{d+1}^j, \ldots, \dot{X}_{2d-|\mathrm{u}|}^j)\end{aligned} \tag{6.7}$$

so that $(\dot{\mathbf{X}}_{\mathrm{u}}^j)_j \sim \mathcal{LH}(n, |\mathrm{u}|)$ and $(\dot{\mathbf{X}}_{\mathrm{u}^c}^{j,1})_j$, $(\dot{\mathbf{X}}_{\mathrm{u}^c}^{j,2})_j \sim \mathcal{LH}(n, d-|\mathrm{u}|)$. Finally, for $k=1$ and 2, we denote

$$\dot{Y}_{\mathrm{u}}^{j,k} = f(\dot{\mathbf{X}}_{\mathrm{u}}^j, \dot{\mathbf{X}}_{\mathrm{u}^c}^{j,k}) \ . \tag{6.8}$$

As in the previous section, we consider the estimators defined in (6.4) and (6.5) but we now replace the simple random sample $(\mathbf{Z}_{\mathrm{u}}^j)_{j=1..n}$ by the stratified sample $(\dot{\mathbf{Z}}_{\mathrm{u}}^j)_{j=1..n}$. The resulting estimators are now denoted $\widetilde{\underline{S}}_{\mathrm{u},n}^{LHS} = \widetilde{\underline{\tau}}_{\mathrm{u},n}^{2,LHS}/\widetilde{\sigma}_n^{2,LHS}$ and $\widehat{\underline{S}}_{\mathrm{u},n}^{LHS} = \widehat{\underline{\tau}}_{\mathrm{u},n}^{2,LHS}/\widehat{\sigma}_n^{2,LHS}$, respectively.

6.3.2 Statistical properties of the estimators

The statistical properties of $\widetilde{\underline{S}}_{\mathrm{u},n}^{LHS}$ and $\widehat{\underline{S}}_{\mathrm{u},n}^{LHS}$ are gathered in the following result :

Proposition 6.1.
(i) *If f^4 is integrable then $\widetilde{\underline{S}}_{\mathrm{u},n}^{LHS}$ et $\widehat{\underline{S}}_{\mathrm{u},n}^{LHS}$ are strongly consistent.*

(ii) *If f^6 is integrable then $\sqrt{n}(\widetilde{\underline{S}}_{\mathrm{u},n}^{LHS} - \underline{S}_{\mathrm{u}})$ and $\sqrt{n}(\widehat{\underline{S}}_{\mathrm{u},n}^{LHS} - \underline{S}_{\mathrm{u}})$ converge in law to a zero-mean normal distribution with lower variance than the respective variance given in the central limit theorem (CLT) for the basic estimators $\widetilde{\underline{S}}_{\mathrm{u},n}$ and $\widehat{\underline{S}}_{\mathrm{u},n}$.*

(iii) *We have*

$$\begin{aligned}\mathbb{E}\big[\widetilde{\underline{\tau}}_{\mathrm{u},n}^{2,LHS}\big] &= \underline{\tau}_{\mathrm{u}}^2 + B_{n,1}\\ \mathbb{E}\big[\widetilde{\sigma}_n^{2,LHS}\big] &= \sigma^2 + B_{n,2}\\ \mathbb{E}\big[\widehat{\underline{\tau}}_{\mathrm{u},n}^{2,LHS}\big] &= \underline{\tau}_{\mathrm{u}}^2 + B_{n,3}\\ \mathbb{E}\big[\widehat{\sigma}_n^{2,LHS}\big] &= \sigma^2 + B_{n,3}\end{aligned}$$

where

$$\begin{aligned}-\frac{1}{n-1}\underline{\tau}_{\mathrm{u}}^2 \le \ & B_{n,1} \ \le 0\\ -\frac{1}{n-1}\sigma^2 \le \ & B_{n,2} \ \le 0\\ -\frac{1}{2(n-1)}(\sigma^2 + \underline{\tau}_{\mathrm{u}}^2) \le \ & B_{n,3} \ \le 0 \ .\end{aligned}$$

Démonstration.
(i) This is a consequence of the strong law of large numbers for Latin hypercube sampling given in Theorem 3 in [Loh96].

(ii) The proof consists in translating the original proof, given for simple random sampling — see Proposition 2.2 in [JKL+12] — for Latin hypercube sampling. Concerning $\widetilde{\underline{S}}_{\mathrm{u},n}^{LHS}$, it is easy to show that

$$\widetilde{\underline{S}}_{\mathrm{u},n}^{LHS} = \Phi(\overline{\mathbf{V}}_n)$$

where

$$\overline{\mathbf{V}}_n = \frac{1}{n}\sum_{j=1}^n \mathbf{V}_j$$

$$\mathbf{V}_j = \Big(\big(\dot{Y}_{\mathrm{u}}^{j,1} - \mathbb{E}[Y]\big)\big(\dot{Y}_{\mathrm{u}}^{j,2} - \mathbb{E}[Y]\big) \ , \ \dot{Y}_{\mathrm{u}}^{j,1} - \mathbb{E}[Y] \ , \ \dot{Y}_{\mathrm{u}}^{j,2} - \mathbb{E}[Y] \ , \ \big(\dot{Y}_{\mathrm{u}}^{j,1} - \mathbb{E}[Y]\big)^2\Big)^T$$

and

$$\Phi(x,y,z,t) = \frac{x-yz}{t-y^2} \ .$$

Then we deduce from Theorem 2 in [Loh96] that

$$\sqrt{n}\big(\overline{\mathbf{V}}_n - \mu\big) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}_4(0,\Gamma)$$

where $\mu = (\underline{\tau}_{\mathrm{u}}^2, 0, 0, \sigma^2)^T$ and Γ is the covariance matrix of $\mathbf{R}_1 = \mathbf{V}_1 - \mathbf{A}_1$ — see details in Eq. (3) in [Loh96] — defined by

$$\forall i \in \{1,\ldots,4\},\ A_{1i} \text{ is the additive part — see (6.2) — of } V_{1i} \ .$$

Thus the Delta method — see Theorem 3.1 in [Van98] — gives

$$\sqrt{n}\big(\widetilde{S}_{\mathrm{u},n}^{LHS} - \underline{S}_{\mathrm{u}}\big) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}_1(0, g^T\Gamma g)$$

where $g = \nabla\Phi(\mu)$. Developing the term $g^T\Gamma g$ does not seem to provide any useful information. However, denoting σ_{LHS}^2 this term, and σ_{IID}^2 the analogous quantity in the CLT for simple random sampling, we can show that $\sigma_{LHS}^2 \leq \sigma_{IID}^2$. Indeed we first note that, for simple random sampling, the variance given in [JKL$^+$12] reads

$$\sigma_{IID}^2 = \frac{\mathrm{Var}\big[V_{11} - \underline{S}_{\mathrm{u}} V_{14}\big]}{\sigma^2}$$

and for Latin hypercube sampling, it is easy to show that

$$\sigma_{LHS}^2 = \frac{\mathrm{Var}\big[R_{11} - \underline{S}_{\mathrm{u}} R_{14}\big]}{\sigma^2} \ .$$

Hence,

$$\sigma_{IID}^2 = \sigma_{LHS}^2 + \frac{\mathrm{Var}\big[A_{11} - \underline{S}_{\mathrm{u}} A_{14}\big]}{\sigma^2}$$

and the conclusion of (ii) for $\widetilde{S}_{\mathrm{u},n}^{LHS}$ follows. Concerning $\widehat{S}_{\mathrm{u},n}^{LHS}$, the proof follows the same lines — see Proof of (10) in [JKL$^+$12] for details.

(iii) First we have

$$\begin{aligned}
\mathbb{E}\big[\widetilde{\tau}_{\mathrm{u},n}^{2,LHS}\big] &= \frac{n-1}{n^2}\sum_{j=1}^{n}\mathbb{E}\big[\dot{Y}_{\mathrm{u}}^{j,1}\dot{Y}_{\mathrm{u}}^{j,2}\big] - \frac{1}{n^2}\sum_{j=1}^{n}\sum_{\substack{l=1\\ l\neq j}}^{n}\mathbb{E}\big[\dot{Y}_{\mathrm{u}}^{j,1}\dot{Y}_{\mathrm{u}}^{l,2}\big] \\
&= \frac{n-1}{n}\Big(\mathbb{E}[Y]^2 + \underline{\tau}_{\mathrm{u}}^2\Big) - \frac{n-1}{n}\Big(\mathrm{Cov}(\dot{Y}_{\mathrm{u}}^{1,1}, \dot{Y}_{\mathrm{u}}^{2,2}) + \mathbb{E}[Y]^2\Big)
\end{aligned}$$

and thanks to Lemma 6.4 in Appendix 6.A, it gives

$$\mathbb{E}\big[\widetilde{\tau}_{\mathrm{u},n}^{2,LHS}\big] = \underline{\tau}_{\mathrm{u}}^2 + B_{n,1}$$

with

$$-\frac{1}{n-1}\underline{\tau}_{\mathrm{u}}^2 \leq B_{n,1} \leq 0 \ .$$

Concerning $\widetilde{\sigma}_n^{2,LHS}$, we have

$$\begin{aligned}
\mathbb{E}\big[\widetilde{\sigma}_{\overline{\mathrm{u}},n}^{2,LHS}\big] &= \frac{n-1}{n^2}\sum_{j=1}^{n}\mathbb{E}\big[(\dot{Y}_{\mathrm{u}}^{j,1})^2\big] - \frac{1}{n^2}\sum_{j=1}^{n}\sum_{\substack{l=1\\ l\neq j}}^{n}\mathbb{E}\big[\dot{Y}_{\mathrm{u}}^{j,1}\dot{Y}_{\mathrm{u}}^{l,1}\big] \\
&= \frac{n-1}{n}\mathbb{E}[Y^2] - \frac{n-1}{n}\Big(\mathrm{Cov}(\dot{Y}_{\mathrm{u}}^{1,1}, \dot{Y}_{\mathrm{u}}^{2,1}) + \mathbb{E}[Y]^2\Big)
\end{aligned}$$

and noting that

$$\mathrm{Cov}(\dot{Y}_{\mathrm{u}}^{1,1}, \dot{Y}_{\mathrm{u}}^{2,1}) = \mathrm{Cov}\big(\dot{Y}_{\{1,\ldots,d\}}^{1,1}, \dot{Y}_{\{1,\ldots,d\}}^{2,2}\big)$$

we conclude that

$$\mathbb{E}\big[\widetilde{\sigma}_n^{2,LHS}\big] = \sigma^2 + B_{n,2}$$

with

$$-\frac{1}{n-1}\sigma^2 \leq B_{n,2} \leq 0 \ .$$

As for $\widehat{\underline{\tau}}_{\mathrm{u},n}^{2,LHS}$ and $\widehat{\sigma}_n^{2,LHS}$, we have

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{1}{n}\sum_{j=1}^{n}\frac{\dot{Y}_{\mathrm{u}}^{j,1}+\dot{Y}_{\mathrm{u}}^{j,2}}{2}\right)^2\right] \\
&= \frac{1}{4n}\mathbb{E}\big[(\dot{Y}_{\mathrm{u}}^{1,1}+\dot{Y}_{\mathrm{u}}^{1,2})^2\big] + \frac{1}{4n^2}\sum_{j=1}^{n}\sum_{\substack{l=1\\ l\neq j}}^{n}\mathbb{E}\big[(\dot{Y}_{\mathrm{u}}^{j,1}+\dot{Y}_{\mathrm{u}}^{j,2})(\dot{Y}_{\mathrm{u}}^{l,1}+\dot{Y}_{\mathrm{u}}^{l,2})\big] \\
&= \frac{1}{2n}\Big(\mathbb{E}[Y^2] + \underline{\tau}_{\mathrm{u}}^2 + \mathbb{E}[Y]^2\Big) + \frac{n-1}{n}\mathbb{E}[Y]^2 + \frac{n-1}{2n}\Big(\mathrm{Cov}(\dot{Y}_{\mathrm{u}}^{1,1}, \dot{Y}_{\mathrm{u}}^{2,1}) + \mathrm{Cov}(\dot{Y}_{\mathrm{u}}^{1,1}, \dot{Y}_{\mathrm{u}}^{2,2})\Big) \\
&= \frac{1}{2n}\Big(\sigma^2 + \underline{\tau}_{\mathrm{u}}^2\Big) + \mathbb{E}[Y]^2 + \frac{n-1}{2n}\Big(\mathrm{Cov}(\dot{Y}_{\mathrm{u}}^{1,1}, \dot{Y}_{\mathrm{u}}^{2,1}) + \mathrm{Cov}(\dot{Y}_{\mathrm{u}}^{1,1}, \dot{Y}_{\mathrm{u}}^{2,2})\Big) \ .
\end{aligned}$$

Then it is easy to conclude that

$$\begin{aligned}
\mathbb{E}\big[\widehat{\underline{\tau}}_{\mathrm{u},n}^{2,LHS}\big] &= \underline{\tau}_{\mathrm{u}}^2 + B_{n,3} \\
\mathbb{E}\big[\widehat{\sigma}_n^{2,LHS}\big] &= \sigma^2 + B_{n,3}
\end{aligned}$$

with

$$-\frac{1}{2(n-1)}(\sigma^2 + \underline{\tau}_{\mathrm{u}}^2) \leq B_{n,3} \leq 0 \ .$$

□

Remark 6.1. *Due to their intricate structure, the biases of the estimators $\widetilde{\underline{\tau}}_{\mathrm{u},n}^{2,LHS}$, $\widetilde{\sigma}_n^{2,LHS}$, $\widehat{\underline{\tau}}_{\mathrm{u},n}^{2,LHS}$ and $\widehat{\sigma}_n^{2,LHS}$ can't be easily reduced. Nevertheless we can note that these biases are asymptotically negligible, with a rate of convergence in $O(n^{-1})$ larger than the rate of convergence of the estimators — to their theoretical values — themselves, which is in $O(n^{-1/2})$.*

6.4 Monte Carlo estimators and replicated Latin hypercube sampling

6.4.1 Notation and definitions

We begin with the definition of replicated Latin hypercubes :

Definition 6.2. *Let d and n in $\mathbb{N}^*$, and consider Π_n the set of all the permutations of $\{1,\ldots,n\}$. We say that $(\mathbf{X}^j)_{j=1..n}$ and $(\mathbf{X}'^j)_{j=1..n}$ are two replicated Latin hypercubes of size n in $[0,1]^d$ — and we denote $(\mathbf{X}^j, \mathbf{X}'^j)_j \sim \mathcal{RLH}(n,d)$ — if for all $j \in \{1,\ldots,n\}$,*

$$\mathbf{X}^j = \left(\frac{\pi_1(j) - U_{1,\pi_1(j)}}{n}, \ldots, \frac{\pi_d(j) - U_{d,\pi_d(j)}}{n}\right)$$

and

$$\mathbf{X}'^j = \left(\frac{\pi'_1(j) - U_{1,\pi'_1(j)}}{n}, \ldots, \frac{\pi'_d(j) - U_{d,\pi'_d(j)}}{n}\right)$$

where the π_i's, the π'_i's and the $U_{i,j}$'s are independent random variables uniformly distributed on Π_n, Π_n and $[0,1]$, respectively.

Now let $\mathfrak{u}$ be a non-empty subset of $\{1,\dots d\}$, and j in $\{1,\dots n\}$, and consider

$$\ddot{\mathbf{Z}}_{\mathfrak{u}}^{j} = (\ddot{X}_1^j, \dots, \ddot{X}_{2d-|\mathfrak{u}|}^j)$$

and

$$\begin{aligned} \ddot{\mathbf{X}}_{\mathfrak{u}}^{j} &= (\ddot{X}_1^j, \dots, \ddot{X}_{|\mathfrak{u}|}^j) \\ \ddot{\mathbf{X}}_{\mathfrak{u}^c}^{j,1} &= (\ddot{X}_{|\mathfrak{u}|+1}^j, \dots, \ddot{X}_{d}^j) \\ \ddot{\mathbf{X}}_{\mathfrak{u}^c}^{j,2} &= (\ddot{X}_{d+1}^j, \dots, \ddot{X}_{2d-|\mathfrak{u}|}^j) \end{aligned} \tag{6.9}$$

where $(\ddot{\mathbf{X}}_{\mathfrak{u}}^{j})_j \sim \mathcal{LH}(n,|\mathfrak{u}|)$ and $(\ddot{\mathbf{X}}_{\mathfrak{u}^c}^{j,1}, \ddot{\mathbf{X}}_{\mathfrak{u}^c}^{j,2})_j \sim \mathcal{RLH}(n, d-|\mathfrak{u}|)$, $(\ddot{\mathbf{X}}_{\mathfrak{u}}^{j})_j$ and $(\ddot{\mathbf{X}}_{\mathfrak{u}^c}^{j,1}, \ddot{\mathbf{X}}_{\mathfrak{u}^c}^{j,2})_j$ being independent. Finally, for $k=1$ and 2, we denote

$$\ddot{Y}_{\mathfrak{u}}^{j,k} = f(\ddot{\mathbf{X}}_{\mathfrak{u}}^{j}, \ddot{\mathbf{X}}_{\mathfrak{u}^c}^{j,k}) \; . \tag{6.10}$$

As in Section 6.2, we consider the estimators defined in (6.4) and (6.5) but we now replace the simple random sample $(\mathbf{Z}_{\mathfrak{u}}^{j})_{j=1..n}$ by the stratified sample based on replicated Latin hypercubes $(\ddot{\mathbf{Z}}_{\mathfrak{u}}^{j})_{j=1..n}$. The resulting estimators are now denoted $\widetilde{\underline{S}}_{\mathfrak{u},n}^{RLHS} = \widetilde{\underline{\tau}}_{\mathfrak{u},n}^{2,RLHS}/\widetilde{\sigma}_n^{2,RLHS}$ and $\widehat{\underline{S}}_{\mathfrak{u},n}^{RLHS} = \widehat{\underline{\tau}}_{\mathfrak{u},n}^{2,RLHS}/\widehat{\sigma}_n^{2,RLHS}$, respectively. Note that estimators of sensitivity indices based on r replicated Latin hypercubes have already been introduced by McKay [McK95] (see also the summarized presentation by Saltelli et al. [SCS00]), but these estimators converge to their corresponding analytical Sobol' index only as r tends to $+\infty$.

6.4.2 Statistical properties of the estimators

The statistical properties of $\widetilde{\underline{S}}_{\mathfrak{u},n}^{RLHS}$ and $\widehat{\underline{S}}_{\mathfrak{u},n}^{RLHS}$ are gathered in the following result :

Proposition 6.2.
(i) *If f^4 is integrable then $\widetilde{\underline{S}}_{\mathfrak{u},n}^{RLHS}$ and $\widehat{\underline{S}}_{\mathfrak{u},n}^{RLHS}$ are strongly consistent.*

(ii) *If f^6 is integrable then $\sqrt{n}(\widetilde{\underline{S}}_{\mathfrak{u},n}^{RLHS} - \underline{S}_{\mathfrak{u}})$ and $\sqrt{n}(\widehat{\underline{S}}_{\mathfrak{u},n}^{RLHS} - \underline{S}_{\mathfrak{u}})$ converge in law to a zero-mean normal distribution with the same respective variance given in CLT for the estimators $\widetilde{\underline{S}}_{\mathfrak{u},n}^{LHS}$ and $\widetilde{\underline{S}}_{\mathfrak{u},n}^{LHS}$.*

(iii) *We have*

$$\begin{aligned} \mathbb{E}\big[\widetilde{\underline{\tau}}_{\mathfrak{u},n}^{2,RLHS}\big] &= \underline{\tau}_{\mathfrak{u}}^2 - \frac{1}{n}\underline{\tau}_{\mathfrak{u}}^2 + B_{n,1} + B_{|\mathfrak{u}|,n} \\ \mathbb{E}\big[\widetilde{\sigma}_{n}^{2,RLHS}\big] &= \sigma^2 + B_{n,3} \\ \mathbb{E}\big[\widehat{\underline{\tau}}_{\mathfrak{u},n}^{2,RLHS}\big] &= \underline{\tau}_{\mathfrak{u}}^2 - \frac{1}{2n}\underline{\tau}_{\mathfrak{u}}^2 + B_{n,1} + B_{n,2} + B_{|\mathfrak{u}|,n} \\ \mathbb{E}\big[\widehat{\sigma}_{n}^{2,RLHS}\big] &= \sigma^2 - \frac{1}{2n}\underline{\tau}_{\mathfrak{u}}^2 + B_{n,1} + B_{n,2} \end{aligned}$$

where

$$\begin{aligned} |B_{n,1}| &\leq \left(\frac{d+1}{n}+2\right)\left(\frac{d+1}{n}\right)\mathbb{E}[Y^2] \\ |B_{n,2}| &\leq \frac{\sigma^2}{2n} \\ -\frac{1}{n-1}\sigma^2 \leq \quad B_{n,3} &\leq 0 \\ |B_{|\mathfrak{u}|,n}| &\leq \left(\frac{d-|\mathfrak{u}|+1}{n}+2\right)\left(\frac{d-|\mathfrak{u}|+1}{n-1}\right)\mathbb{E}[Y^2] \; . \end{aligned}$$

Démonstration.
(i) The proof is divided into two parts. In the first one, we only consider continuous functions, and in the second one, we extend the result to the whole class of functions such that f^4 is integrable.

First part : Consistency is obvious as in Proposition 6.1, except for the term

$$\frac{1}{n}\sum_{j=1}^{n} \ddot{Y}_{\mathfrak{u}}^{j,1}\ddot{Y}_{\mathfrak{u}}^{j,2} \ .$$

So denote $\overline{\mathbf{X}}_{\mathfrak{u}^c}^{j,2}$ the Latin hypercube defined by

$$\overline{\mathbf{X}}_{\mathfrak{u}^c}^{j,2} = \frac{\lfloor n\ddot{\mathbf{X}}_{\mathfrak{u}^c}^{j,2}\rfloor + \mathbf{U}_j}{n}$$

where the $\mathbf{U}_j$'s are independent random vectors uniformly distributed in $[0,1[^{d-|\mathfrak{u}|}$ independent from all the permutations and shifts in the definition of $(\ddot{\mathbf{Z}}_{\mathfrak{u}}^j)_j$, and $\lfloor\cdot\rfloor$ is the floor function. We can write

$$\begin{aligned}
\frac{1}{n}\sum_{j=1}^{n} \ddot{Y}_{\mathfrak{u}}^{j,1}\ddot{Y}_{\mathfrak{u}}^{j,2} &= \frac{1}{n}\sum_{j=1}^{n} f(\ddot{\mathbf{X}}_{\mathfrak{u}}^{j}, \ddot{\mathbf{X}}_{\mathfrak{u}^c}^{j,1}) f(\ddot{\mathbf{X}}_{\mathfrak{u}}^{j}, \ddot{\mathbf{X}}_{\mathfrak{u}^c}^{j,2}) \\
&= \frac{1}{n}\sum_{j=1}^{n} f(\ddot{\mathbf{X}}_{\mathfrak{u}}^{j}, \ddot{\mathbf{X}}_{\mathfrak{u}^c}^{j,1}) f(\ddot{\mathbf{X}}_{\mathfrak{u}}^{j}, \overline{\mathbf{X}}_{\mathfrak{u}^c}^{j,2}) \\
&+ \frac{1}{n}\sum_{j=1}^{n} f(\ddot{\mathbf{X}}_{\mathfrak{u}}^{j}, \ddot{\mathbf{X}}_{\mathfrak{u}^c}^{j,1})\Big(f(\ddot{\mathbf{X}}_{\mathfrak{u}}^{j}, \ddot{\mathbf{X}}_{\mathfrak{u}^c}^{j,2}) - f(\ddot{\mathbf{X}}_{\mathfrak{u}}^{j}, \overline{\mathbf{X}}_{\mathfrak{u}^c}^{j,2})\Big) \ . \qquad (6.11)
\end{aligned}$$

The first term on the right-hand side is an estimator as described in Section 6.3 since we note that $(\ddot{\mathbf{X}}_{\mathfrak{u}}^{j}, \ddot{\mathbf{X}}_{\mathfrak{u}^c}^{j,1}, \overline{\mathbf{X}}_{\mathfrak{u}^c}^{j,2})_{j=1..n} \sim \mathcal{LH}(n, 2d-|\mathfrak{u}|)$; so it converges to $\mathbb{E}[Y]^2 + \underline{\tau}_{\mathfrak{u}}^2$ almost surely. The second term on the right-hand side converges to 0 since as f is bounded — by continuity on a compact — it is bounded by

$$\frac{\sup|f|}{n}\sum_{j=1}^{n}\Big|f(\ddot{\mathbf{X}}_{\mathfrak{u}}^{j}, \ddot{\mathbf{X}}_{\mathfrak{u}^c}^{j,2}) - f(\ddot{\mathbf{X}}_{\mathfrak{u}}^{j}, \overline{\mathbf{X}}_{\mathfrak{u}^c}^{j,2})\Big|$$

and by uniform continuity of f — due to Heine-Cantor theorem — this quantity tends to 0 as n tends to $+\infty$. Thus the sum in the right-hand side, i.e. $\frac{1}{n}\sum_{j=1}^{n} \ddot{Y}_{\mathfrak{u}}^{j,1}\ddot{Y}_{\mathfrak{u}}^{j,2}$, converges to $\mathbb{E}[Y]^2 + \underline{\tau}_{\mathfrak{u}}^2$ almost surely.

Second part : Since the space of continuous functions on $[0,1]^d$ — denoted $\mathcal{C}\big([0,1]^d\big)$ — is dense in $L^4\big([0,1]^d\big)$, let $(f_m)_{m\in\mathbb{N}^*}$ be a sequence in $\mathcal{C}\big([0,1]^d\big)$ such that $\mathbb{E}\big[|f_m(\mathbf{X}) - f(\mathbf{X})|^4\big]$ converges to 0 as m tends to $+\infty$, where $\mathbf{X}$ is uniformly distributed on $[0,1]^d$.

Now let $\varepsilon > 0$ and $M = M(\varepsilon) \in \mathbb{N}^*$ such that

$$\mathbb{E}\Big[\big(f_M(\mathbf{X}) - f(\mathbf{X})\big)^2\Big] < \frac{\varepsilon^2}{65\ \mathbb{E}[f^2(\mathbf{X})]}. \qquad (6.12)$$

We can write

$$\begin{aligned}
\frac{1}{n}\sum_{j=1}^{n} \ddot{Y}_{\mathfrak{u}}^{j,1}\ddot{Y}_{\mathfrak{u}}^{j,2} &= \frac{1}{n}\sum_{j=1}^{n} f(\ddot{\mathbf{X}}_{\mathfrak{u}}^{j}, \ddot{\mathbf{X}}_{\mathfrak{u}^c}^{j,1}) f(\ddot{\mathbf{X}}_{\mathfrak{u}}^{j}, \overline{\mathbf{X}}_{\mathfrak{u}^c}^{j,2}) \\
&+ \frac{1}{n}\sum_{j=1}^{n} f(\ddot{\mathbf{X}}_{\mathfrak{u}}^{j}, \ddot{\mathbf{X}}_{\mathfrak{u}^c}^{j,1})\Big(f_M(\ddot{\mathbf{X}}_{\mathfrak{u}}^{j}, \ddot{\mathbf{X}}_{\mathfrak{u}^c}^{j,2}) - f_M(\ddot{\mathbf{X}}_{\mathfrak{u}}^{j}, \overline{\mathbf{X}}_{\mathfrak{u}^c}^{j,2})\Big) \\
&+ \frac{1}{n}\sum_{j=1}^{n} f(\ddot{\mathbf{X}}_{\mathfrak{u}}^{j}, \ddot{\mathbf{X}}_{\mathfrak{u}^c}^{j,1})\Big(f(\ddot{\mathbf{X}}_{\mathfrak{u}}^{j}, \ddot{\mathbf{X}}_{\mathfrak{u}^c}^{j,2}) - f_M(\ddot{\mathbf{X}}_{\mathfrak{u}}^{j}, \ddot{\mathbf{X}}_{\mathfrak{u}^c}^{j,2})\Big) \\
&+ \frac{1}{n}\sum_{j=1}^{n} f(\ddot{\mathbf{X}}_{\mathfrak{u}}^{j}, \ddot{\mathbf{X}}_{\mathfrak{u}^c}^{j,1})\Big(f_M(\ddot{\mathbf{X}}_{\mathfrak{u}}^{j}, \overline{\mathbf{X}}_{\mathfrak{u}^c}^{j,2}) - f(\ddot{\mathbf{X}}_{\mathfrak{u}}^{j}, \overline{\mathbf{X}}_{\mathfrak{u}^c}^{j,2})\Big) \qquad (6.13)
\end{aligned}$$

As noted in the proof of (i) in Proposition 6.1, the first term on the right-hand side of (6.13) converges to $\underline{\tau}_{\mathrm{u}}^2 + \mathbb{E}[Y]^2$ almost surely as n tends to $+\infty$ i.e.

$$\mathbb{P}\Big(\forall \varepsilon > 0,\ \exists N_1 \in \mathbb{N}^*,\ \forall n > N_1,\ \Big|\frac{1}{n}\sum_{j=1}^{n} f(\ddot{\mathbf{X}}_{\mathrm{u}}^{j}, \ddot{\mathbf{X}}_{\mathrm{u}^c}^{j,1}) f(\ddot{\mathbf{X}}_{\mathrm{u}}^{j}, \overline{\ddot{\mathbf{X}}}_{\mathrm{u}^c}^{j,2}) - \underline{\tau}_{\mathrm{u}}^2 - \mathbb{E}[Y]^2\Big| < \frac{\varepsilon}{4}\Big) = 1\ . \quad (6.14)$$

Since f_M is uniformly continuous on $[0,1]^d$, we have that

$$A_n = \sup_{1\le j\le n} \big|f_M(\ddot{\mathbf{X}}_{\mathrm{u}}^{j}, \ddot{\mathbf{X}}_{\mathrm{u}^c}^{j,2}) - f_M(\ddot{\mathbf{X}}_{\mathrm{u}}^{j}, \overline{\ddot{\mathbf{X}}}_{\mathrm{u}^c}^{j,2})\big|$$

converges almost surely to 0 as n tends to $+\infty$. Moreover, since f is integrable, we have that $\frac{1}{n}\sum_{j=1}^{n} |f(\ddot{\mathbf{X}}_{\mathrm{u}}^{j}, \ddot{\mathbf{X}}_{\mathrm{u}^c}^{j,1})|$ converges to $\mathbb{E}\big[|Y|\big]$ as n tends to $+\infty$. Hence

$$\begin{aligned}
&\mathbb{P}\Big(\forall \varepsilon > 0,\ \exists N_1 \in \mathbb{N}^*,\ \forall n > N_1,\ \Big|\frac{1}{n}\sum_{j=1}^{n} f(\ddot{\mathbf{X}}_{\mathrm{u}}^{j}, \ddot{\mathbf{X}}_{\mathrm{u}^c}^{j,1})\Big(f_M(\ddot{\mathbf{X}}_{\mathrm{u}}^{j}, \ddot{\mathbf{X}}_{\mathrm{u}^c}^{j,2}) - f_M(\ddot{\mathbf{X}}_{\mathrm{u}}^{j}, \overline{\ddot{\mathbf{X}}}_{\mathrm{u}^c}^{j,2})\Big)\Big| < \frac{\varepsilon}{4}\Big) \\
&\quad \ge \quad \mathbb{P}\Big(\forall \varepsilon > 0,\ \exists N_2 \in \mathbb{N}^*,\ \forall n > N_2,\ A_n \frac{1}{n}\sum_{j=1}^{n} \big|f(\ddot{\mathbf{X}}_{\mathrm{u}}^{j}, \ddot{\mathbf{X}}_{\mathrm{u}^c}^{j,1})\big| < \frac{\varepsilon}{4}\Big) \\
&\quad = \quad 1\ . \qquad (6.15)
\end{aligned}$$

For the third and the fourth terms on the right-hand side of (6.13), we apply twice the same proof. First the Cauchy-Schwartz inequality gives

$$\mathbb{P}\Big(\forall \varepsilon > 0,\ \exists N_3 \in \mathbb{N}^*,\ \forall n > N_3,\ \Big|\frac{1}{n}\sum_{j=1}^{n} f(\ddot{\mathbf{X}}_{\mathrm{u}}^{j}, \ddot{\mathbf{X}}_{\mathrm{u}^c}^{j,1})\Big(f(\ddot{\mathbf{X}}_{\mathrm{u}}^{j}, \ddot{\mathbf{X}}_{\mathrm{u}^c}^{j,2}) - f_M(\ddot{\mathbf{X}}_{\mathrm{u}}^{j}, \ddot{\mathbf{X}}_{\mathrm{u}^c}^{j,2})\Big)\Big| < \frac{\varepsilon}{4}\Big)$$

$$\ge\ \mathbb{P}\Big(\forall \varepsilon > 0,\ \exists N_3 \in \mathbb{N}^*,\ \forall n > N_3,\ \Big(\frac{1}{n}\sum_{j=1}^{n} f^2(\ddot{\mathbf{X}}_{\mathrm{u}}^{j}, \ddot{\mathbf{X}}_{\mathrm{u}^c}^{j,1})\Big)^{1/2}\Big(\frac{1}{n}\sum_{j=1}^{n}\big(f(\ddot{\mathbf{X}}_{\mathrm{u}}^{j}, \ddot{\mathbf{X}}_{\mathrm{u}^c}^{j,2}) - f_M(\ddot{\mathbf{X}}_{\mathrm{u}}^{j}, \ddot{\mathbf{X}}_{\mathrm{u}^c}^{j,2})\big)^2\Big)^{1/2} < \frac{\varepsilon}{4}\Big).$$

Then note that $\frac{1}{n}\sum_{j=1}^{n} f^2(\ddot{\mathbf{X}}_{\mathrm{u}}^{j}, \ddot{\mathbf{X}}_{\mathrm{u}^c}^{j,1})$ and $\frac{1}{n}\sum_{j=1}^{n}\big(f(\ddot{\mathbf{X}}_{\mathrm{u}}^{j}, \ddot{\mathbf{X}}_{\mathrm{u}^c}^{j,2}) - f_M(\ddot{\mathbf{X}}_{\mathrm{u}}^{j}, \ddot{\mathbf{X}}_{\mathrm{u}^c}^{j,2})\big)^2$ converge almost surely to $\mathbb{E}[Y^2]$ and $\mathbb{E}[(f_M(\mathbf{X}) - f(\mathbf{X}))^2]$ — where $\mathbf{X}$ is uniformly distributed on $[0,1]^d$ — respectively. And deduce that there exists $N_4 \in \mathbb{N}^*$ such that for all $n > N_4$, we have $\frac{1}{n}\sum_{j=1}^{n} f^2(\ddot{\mathbf{X}}_{\mathrm{u}}^{j}, \ddot{\mathbf{X}}_{\mathrm{u}^c}^{j,1}) < 2\ \mathbb{E}[Y^2]$ and $\frac{1}{n}\sum_{j=1}^{n}\big(f(\ddot{\mathbf{X}}_{\mathrm{u}}^{j}, \ddot{\mathbf{X}}_{\mathrm{u}^c}^{j,2}) - f_M(\ddot{\mathbf{X}}_{\mathrm{u}}^{j}, \ddot{\mathbf{X}}_{\mathrm{u}^c}^{j,2})\big)^2 < 2\ \mathbb{E}[(f_M(\mathbf{X}) - f(\mathbf{X}))^2]$ almost surely. As a consequence, deduce from Eq. (6.12) that

$$\begin{aligned}
&\mathbb{P}\Big(\forall \varepsilon > 0,\ \exists N_3 \in \mathbb{N}^*,\ \forall n > N_3,\ \Big|\frac{1}{n}\sum_{j=1}^{n} f(\ddot{\mathbf{X}}_{\mathrm{u}}^{j}, \ddot{\mathbf{X}}_{\mathrm{u}^c}^{j,1})\Big(f(\ddot{\mathbf{X}}_{\mathrm{u}}^{j}, \ddot{\mathbf{X}}_{\mathrm{u}^c}^{j,2}) - f_M(\ddot{\mathbf{X}}_{\mathrm{u}}^{j}, \ddot{\mathbf{X}}_{\mathrm{u}^c}^{j,2})\Big)\Big| < \frac{\varepsilon}{4}\Big) \\
&\quad \ge \quad \mathbb{P}\Big(\forall \varepsilon > 0,\ \exists N_3 > N_4,\ \forall n > N_3,\ \varepsilon\sqrt{\frac{4}{65}} < \frac{\varepsilon}{4}\Big) \\
&\quad = \quad 1 \qquad (6.16)
\end{aligned}$$

Finally, Eqs. (6.14–6.16) gives

$$\mathbb{P}\Big(\forall \varepsilon > 0,\ \exists N \in \mathbb{N}^*,\ \forall n > N,\ \Big|\frac{1}{n}\sum_{j=1}^{n} \ddot{Y}_{\mathrm{u}}^{j,1}\ddot{Y}_{\mathrm{u}}^{j,2}\Big| < \varepsilon\Big) = 1$$

and we have the conclusion.

(ii) As in (i), the only term to treat is

$$\frac{1}{n}\sum_{j=1}^{n} \ddot{Y}_{\mathrm{u}}^{j,1}\ddot{Y}_{\mathrm{u}}^{j,2}\ ,$$

so asymptotic normality is shown in the same way by using the decomposition in (6.11). We always obtain the sum of a term already considered in Section 6.3 which converges in law to a normal distribution and a term which converges to 0 in probability, and the conclusion follows from Slutsky's lemma. We only detail the proof for $\widetilde{S}^{RLHS}_{\mathrm{u},n}$, it is exactly the same for $\widehat{S}^{RLHS}_{\mathrm{u},n}$. So note that following the proof of (ii) in Proposition 1 and the notation above, it is sufficient to show that

$$\sqrt{n}\Big(\frac{1}{n}\sum_{j=1}^{n}\big(f(\ddot{\mathbf{X}}^j_{\mathrm{u}},\ddot{\mathbf{X}}^{j,1}_{\mathrm{u}^c})-\mathbb{E}[Y]\big)\big(f(\ddot{\mathbf{X}}^j_{\mathrm{u}},\ddot{\mathbf{X}}^{j,2}_{\mathrm{u}^c})-f(\ddot{\mathbf{X}}^j_{\mathrm{u}},\overline{\mathbf{X}}^{j,2}_{\mathrm{u}^c})\big)\Big)\xrightarrow[n\to\infty]{\mathbb{P}}0 \tag{6.17}$$

to prove the asymptotic normality of $\widetilde{S}^{RLHS}_{\mathrm{u},n}$.

So consider $\varepsilon,\eta>0$ and prove that there exists $N\in\mathbb{N}^*$ such that for all $n>N$, the quantity

$$P=\mathbb{P}\Big(\Big|\frac{1}{n}\sum_{j=1}^{n}\big(f(\ddot{\mathbf{X}}^j_{\mathrm{u}},\ddot{\mathbf{X}}^{j,1}_{\mathrm{u}^c})-\mathbb{E}[Y]\big)\big(f(\ddot{\mathbf{X}}^j_{\mathrm{u}},\ddot{\mathbf{X}}^{j,2}_{\mathrm{u}^c})-f(\ddot{\mathbf{X}}^j_{\mathrm{u}},\overline{\mathbf{X}}^{j,2}_{\mathrm{u}^c})\big)\Big|>\varepsilon\Big)$$

is less than η. First as f^6 is integrable, there exists a constant $K>0$ such that $\mathbb{P}(|f(\ddot{\mathbf{X}}^j_{\mathrm{u}},\ddot{\mathbf{X}}^{j,1}_{\mathrm{u}^c})|>K)<\eta/4$. Hence

$$\begin{aligned}
P &\le \mathbb{P}\Big(\Big(\Big|\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\big(f(\ddot{\mathbf{X}}^j_{\mathrm{u}},\ddot{\mathbf{X}}^{j,1}_{\mathrm{u}^c})-\mathbb{E}[Y]\big)\big(f(\ddot{\mathbf{X}}^j_{\mathrm{u}},\ddot{\mathbf{X}}^{j,2}_{\mathrm{u}^c})-f(\ddot{\mathbf{X}}^j_{\mathrm{u}},\overline{\mathbf{X}}^{j,2}_{\mathrm{u}^c})\big)\Big|>\varepsilon\Big)\bigcap\big(|f(\ddot{\mathbf{X}}^j_{\mathrm{u}},\ddot{\mathbf{X}}^{j,1}_{\mathrm{u}^c})|\le K\big)\Big)\\
&+ \mathbb{P}\Big(\Big(\Big|\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\big(f(\ddot{\mathbf{X}}^j_{\mathrm{u}},\ddot{\mathbf{X}}^{j,1}_{\mathrm{u}^c})-\mathbb{E}[Y]\big)\big(f(\ddot{\mathbf{X}}^j_{\mathrm{u}},\ddot{\mathbf{X}}^{j,2}_{\mathrm{u}^c})-f(\ddot{\mathbf{X}}^j_{\mathrm{u}},\overline{\mathbf{X}}^{j,2}_{\mathrm{u}^c})\big)\Big|>\varepsilon\Big)\bigcap\big(|f(\ddot{\mathbf{X}}^j_{\mathrm{u}},\ddot{\mathbf{X}}^{j,1}_{\mathrm{u}^c})|>K\big)\Big)\\
&< \mathbb{P}\Big(\frac{K+|\mathbb{E}[Y]|}{\sqrt{n}}\sum_{j=1}^{n}\big|f(\ddot{\mathbf{X}}^j_{\mathrm{u}},\ddot{\mathbf{X}}^{j,2}_{\mathrm{u}^c})-f(\ddot{\mathbf{X}}^j_{\mathrm{u}},\overline{\mathbf{X}}^{j,2}_{\mathrm{u}^c})\big|>\varepsilon\Big)+\frac{\eta}{4}.
\end{aligned}\tag{6.18}$$

Now note that the space of continuous functions on $[0,1]^d$, denoted by $\mathcal{C}([0,1]^d)$, is dense in $L^6([0,1]^d)$ and let $(f_m)_{m\in\mathbb{N}^*}$ be a sequence in $\mathcal{C}([0,1]^d)$ such that $\mathbb{E}[|f_m(\mathbf{X})-f(\mathbf{X})|^6]$ converges to 0 as m tends to $+\infty$ where $\mathbf{X}$ is uniformly distributed on $[0,1]^d$. It is easy to note that there exists $M=M(n)$ such that $\mathbb{P}(|f_M(\mathbf{X})-f(\mathbf{X})|>1/n)<\eta/4$. Thus we get from Eq. (6.18) that

$$\begin{aligned}
P &< \sum_{i=1}^{4}\mathbb{P}\Big(\Big(\frac{K+|\mathbb{E}[Y]|}{\sqrt{n}}\sum_{j=1}^{n}\big(|f_M(\ddot{\mathbf{X}}^j_{\mathrm{u}},\ddot{\mathbf{X}}^{j,2}_{\mathrm{u}^c})-f_M(\ddot{\mathbf{X}}^j_{\mathrm{u}},\overline{\mathbf{X}}^{j,2}_{\mathrm{u}^c})|+|f_M(\ddot{\mathbf{X}}^j_{\mathrm{u}},\ddot{\mathbf{X}}^{j,2}_{\mathrm{u}^c})-f(\ddot{\mathbf{X}}^j_{\mathrm{u}},\ddot{\mathbf{X}}^{j,2}_{\mathrm{u}^c})|\\
&+ |f_M(\ddot{\mathbf{X}}^j_{\mathrm{u}},\overline{\mathbf{X}}^{j,2}_{\mathrm{u}^c})-f(\ddot{\mathbf{X}}^j_{\mathrm{u}},\overline{\mathbf{X}}^{j,2}_{\mathrm{u}^c})|\big)\Big)\bigcap A_i\Big)+\frac{\eta}{4}
\end{aligned}$$

where

$$\begin{aligned}
A_1 &= \big(|f_M(\ddot{\mathbf{X}}^j_{\mathrm{u}},\overline{\mathbf{X}}^{j,2}_{\mathrm{u}^c})-f(\ddot{\mathbf{X}}^j_{\mathrm{u}},\overline{\mathbf{X}}^{j,2}_{\mathrm{u}^c})|>\frac{1}{n}\big)\cap\big(|f_M(\ddot{\mathbf{X}}^j_{\mathrm{u}},\ddot{\mathbf{X}}^{j,2}_{\mathrm{u}^c})-f(\ddot{\mathbf{X}}^j_{\mathrm{u}},\ddot{\mathbf{X}}^{j,2}_{\mathrm{u}^c})|>\frac{1}{n}\big)\\
A_2 &= \big(|f_M(\ddot{\mathbf{X}}^j_{\mathrm{u}},\overline{\mathbf{X}}^{j,2}_{\mathrm{u}^c})-f(\ddot{\mathbf{X}}^j_{\mathrm{u}},\overline{\mathbf{X}}^{j,2}_{\mathrm{u}^c})|>\frac{1}{n}\big)\cap\big(|f_M(\ddot{\mathbf{X}}^j_{\mathrm{u}},\ddot{\mathbf{X}}^{j,2}_{\mathrm{u}^c})-f(\ddot{\mathbf{X}}^j_{\mathrm{u}},\ddot{\mathbf{X}}^{j,2}_{\mathrm{u}^c})|<\frac{1}{n}\big)
\end{aligned}$$

and A_3 and A_4 are the complementary events of A_1 and A_2, respectively. So we deduce

$$\begin{aligned}
P &< \mathbb{P}\Big(\Big(\frac{K+|\mathbb{E}[Y]|}{\sqrt{n}}\sum_{j=1}^{n}\big(|f_M(\ddot{\mathbf{X}}^j_{\mathrm{u}},\ddot{\mathbf{X}}^{j,2}_{\mathrm{u}^c})-f_M(\ddot{\mathbf{X}}^j_{\mathrm{u}},\overline{\mathbf{X}}^{j,2}_{\mathrm{u}^c})|+|f_M(\ddot{\mathbf{X}}^j_{\mathrm{u}},\ddot{\mathbf{X}}^{j,2}_{\mathrm{u}^c})-f(\ddot{\mathbf{X}}^j_{\mathrm{u}},\ddot{\mathbf{X}}^{j,2}_{\mathrm{u}^c})|\\
&+ |f_M(\overline{\mathbf{X}}^{j}_{\mathrm{u}},\ddot{\mathbf{X}}^{j,2}_{\mathrm{u}^c})-f(\ddot{\mathbf{X}}^j_{\mathrm{u}},\overline{\mathbf{X}}^{j,2}_{\mathrm{u}^c})|\big)\Big)\bigcap A_3\Big)+\mathbb{P}(A_1)+\mathbb{P}(A_2)+\mathbb{P}(A_4)+\frac{\eta}{4}\\
&< \mathbb{P}\Big(\frac{K+|\mathbb{E}[Y]|}{\sqrt{n}}\Big(2+\sum_{j=1}^{n}|f_M(\ddot{\mathbf{X}}^j_{\mathrm{u}},\ddot{\mathbf{X}}^{j,2}_{\mathrm{u}^c})-f_M(\ddot{\mathbf{X}}^j_{\mathrm{u}},\overline{\mathbf{X}}^{j,2}_{\mathrm{u}^c})|\Big)>\varepsilon\Big)+\eta.
\end{aligned}$$

Now by an other density argument, note that there exists a sequence of Lipschitz continuous functions with constant 1, denoted $(f_{M,q})_{q\in\mathbb{N}^*}$, such that $\sup_{[0,1]^d}|f_{M,q}(\mathbf{x})-f_M(\mathbf{x})|$ converges to 0 as q tends

to $+\infty$. Then there exists $Q = Q(n) \in \mathbb{N}^*$ such that $\sup_{[0,1]^d} |f_{M,Q}(\mathbf{x}) - f_M(\mathbf{x})| < 1/n$ and deduce that

$$\begin{aligned} P &< \mathbb{P}\Bigg(\frac{K+|\mathbb{E}[Y]|}{\sqrt{n}}\Big(2+\sum_{j=1}^{n}\big(|f_{M,Q}(\ddot{\mathbf{X}}^j_{\mathrm{u}},\ddot{\mathbf{X}}^{j,2}_{\mathrm{u}^c}) - f_{M,Q}(\ddot{\mathbf{X}}^j_{\mathrm{u}},\overline{\mathbf{X}}^{j,2}_{\mathrm{u}^c})| + |f_{M,Q}(\ddot{\mathbf{X}}^j_{\mathrm{u}},\ddot{\mathbf{X}}^{j,2}_{\mathrm{u}^c}) - f_M(\ddot{\mathbf{X}}^j_{\mathrm{u}},\ddot{\mathbf{X}}^{j,2}_{\mathrm{u}^c})| \\ &+ |f_{M,Q}(\ddot{\mathbf{X}}^j_{\mathrm{u}},\overline{\mathbf{X}}^{j,2}_{\mathrm{u}^c}) - f_M(\ddot{\mathbf{X}}^j_{\mathrm{u}},\overline{\mathbf{X}}^{j,2}_{\mathrm{u}^c})|\big)\Big) > \varepsilon\Bigg) + \eta \\ &< \mathbb{P}\Bigg(\frac{5(K+|\mathbb{E}[Y]|)}{\sqrt{n}} > \varepsilon\Bigg) + \eta. \end{aligned}$$

and the conclusion follows.

(iii) The proof is given in Appendix B. □

6.4.3 Estimating all the first-order Sobol' indices using only two replicated Latin hypercube

First note that for any independent random permutations π_1 and π_2 uniformly distributed in Π_n, we have that π_1 and $\pi_1 \circ \pi_2$ are independent.

Then, let $(\mathbf{D}^{j,1}, \mathbf{D}^{j,2})_j \sim \mathcal{RLH}(n,d)$ be a design of experiments of $2n$ points in dimension d defined as in Definition 6.2. Note that, by keeping the notation used in Definition 6.2, for any $i \in \{1,\dots,d\}$, we have :

(i) $(D_i^{j,1})_j \sim \mathcal{LH}(n,1)$

(ii) for all $j \in \{1,\dots,n\}$, $D_i^{j,1} = D_i^{\pi_i'^{-1}\circ\pi_i(j),2}$

(iii) $\big(\mathbf{D}^{i,1}_{\{i\}^c}, \mathbf{D}^{\pi_i'^{-1}\circ\pi_i(j),2}_{\{i\}^c}\big)_j \sim \mathcal{RLH}(n,d-1)$

(iv) $(D_i^{i,1})_j$ and $\big(\mathbf{D}^{i,1}_{\{i\}^c}, \mathbf{D}^{\pi_i'^{-1}\circ\pi_i(j),2}_{\{i\}^c}\big)_j$ are independent.

As a consequence, we can estimate all the S_i's with the two replicated Latin hypercubes $(\mathbf{D}^{j,1}, \mathbf{D}^{j,2})_j$ by considering successively as in Eqs. (6.9–6.10), for any $i \in \{1,\dots,d\}$ and $j \in \{1,\dots,n\}$,

$$\begin{aligned} \ddot{X}^j_{\{i\}} &= D_i^{i,1} = D_i^{\pi_i'^{-1}\circ\pi_i(j),2} \\ \ddot{\mathbf{X}}^{j,1}_{\{i\}^c} &= \mathbf{D}^{i,1}_{\{i\}^c} \\ \ddot{\mathbf{X}}^{j,2}_{\{i\}^c} &= \mathbf{D}^{\pi_i'^{-1}\circ\pi_i(j),2}_{\{i\}^c} \end{aligned}$$

and for $k = 1$ and 2,

$$\ddot{Y}^{j,k}_{\{i\}} = f\big(\ddot{X}^j_{\{i\}}, \ddot{\mathbf{X}}^{j,k}_{\{i\}^c}\big).$$

6.4.4 Construction with Latin hypercubes based on general orthogonal arrays

We first begin with the definition of an orthogonal array (OA) :

Definition 6.3. *An orthogonal array in dimension d, with q levels, strength $t \le d$ and index λ is a matrix with $n = \lambda q^t$ rows and d columns such that in every n-by-t submatrix each of the q^t possible rows — i.e. the distinct t-tuples $(l_1,\dots,l_t)$ where the l_i's take their values in the set of the q levels — occurs exactly the same number λ of times.*

We now recall the definition of OA-based Latin hypercubes — see [Owe92] — and introduce the general notion of replicated OA-based Latin hypercubes.

Definition 6.4. *Let $(A_i^j)_{i=1..d, j=1..n}$ be an orthogonal array in dimension d, with n points and q levels in $\{1,\dots,q\}$, and consider Π_q the set of all the permutations of $\{1,\dots,q\}$. We say that $(\mathbf{X}^j)_{j=1..n}$*

is a Latin hypercube based on the orthogonal array $(\mathbf{A}^j)_{j=1..n}$ *— and we denote* $(\mathbf{X}^j)_j \sim \mathcal{LH}\big((\mathbf{A}^j)_j\big)$ *— if for all* $j \in \{1,\dots,n\}$,

$$\mathbf{X}^j = \left(\frac{\pi_1(A_1^j) - U_{1,\pi_1(A_1^j)}}{q}, \dots, \frac{\pi_d(A_d^j) - U_{d,\pi_d(A_d^j)}}{q}\right)$$

where the π_i*'s and the* $U_{i,j}$*'s are independent random variables uniformly distributed on* Π_q *and* $[0,1]$*, respectively.*

Definition 6.5. *Let* $(A_i^j)_{i=1..d,j=1..n}$ *an orthogonal array in dimension* d*, with* n *points and* q *levels in* $\{1,\dots,q\}$*, and consider* Π_q *the set of all the permutations of* $\{1,\dots,q\}$*. We say that* $(\mathbf{X}^j)_{j=1..n}$ *and* $(\mathbf{X}'^j)_{j=1..n}$ *are two replicated Latin hypercubes based on the orthogonal array* $(\mathbf{A}^j)_{j=1..n}$ *— and we denote* $(\mathbf{X}^j, \mathbf{X}'^j)_j \sim \mathcal{RLH}\big((\mathbf{A}^j)_j\big)$ *— if for all* $j \in \{1,\dots,n\}$,

$$\mathbf{X}^j = \left(\frac{\pi_1(A_1^j) - U_{1,\pi_1(A_1^j)}}{q}, \dots, \frac{\pi_d(A_d^j) - U_{d,\pi_d(A_d^j)}}{q}\right)$$

and

$$\mathbf{X}'^j = \left(\frac{\pi'_1(A_1^j) - U_{1,\pi'_1(A_1^j)}}{q}, \dots, \frac{\pi'_d(A_d^j) - U_{d,\pi'_d(A_d^j)}}{q}\right)$$

where the π_i*'s, the* π'_i*'s and the* $U_{i,j}$*'s are independent random variables uniformly distributed on* Π_q*,* Π_q *and* $[0,1]$*, respectively.*

Note that in the particular case of the orthogonal array $(\mathbf{A}^j)_{j=1}$ with strength 1 and index unity defined by

$$\forall i \in \{1,\dots d\},\ \forall j \in \{1,\dots n\},\quad A_i^j = j,$$

these definitions are exactly Definitions 6.1 and 6.2.

Now the designs of experiments introduced in Definition 6.5 allow to estimate all the $\underline{S}_{\mathfrak{u}}$'s with $|\mathfrak{u}| = t$ where t is the strength of the underlying orthogonal array of the replicated Latin hypercube. More precisely, let $(\mathbf{A}^j)_j$ be an orthogonal array with q levels in $\{1,\dots,q\}$, strength t and index unity, and consider $(\mathbf{D}^{j,1}, \mathbf{D}^{j,2})_j \sim \mathcal{RLH}\big((\mathbf{A}^j)_j\big)$. For any $\mathfrak{u} \subseteq \{1,\dots,d\}$, with $\mathfrak{u} = \{i_1,\dots,i_t\}$, and any $k = 1$ or 2, define $(\mathbf{D}^{j(\mathfrak{u}),k})_j$ as the set of points $(\mathbf{D}^{j,k})_j$ ranked in increasing lexicographic order with respect to the i_1-, …, i_t-th coordinates. Then we can define estimators of the $\underline{S}_{\mathfrak{u}}$'s by considering those in Eqs. (6.4) and (6.5) and by replacing (6.3) by

$$Y_{\mathfrak{u}}^{j,k} = f\big(\mathbf{D}^{j(\mathfrak{u}),k}\big)\ .$$

Remark 6.2. *Theoretical properties of the estimators for this generalisation remain open issues and will consist of a further work. The first step for strong consistency will be to state a strong law of large numbers for OA-based Latin hypercubes with strength* $t > 1$ *since, as far as we know, such a result does not exist. Asymptotic normality has already been proved for OA-based Latin hypercube with strength* $t = 2$ *under smoothness conditions — see* [Loh08] *— but it is not sufficient to conclude in the case of replicated OA-based Latin hypercubes since formulas as in* (6.25) *and* (6.26) *are necessary. As for the biases of the estimators, it will be necessary to study covariances in OA-based Latin hypercubes with strength* $t > 1$ *in order to state formulas as in* (6.25) *and* (6.26) *as well.*

6.5 Numerical illustrations

6.5.1 Application to an analytical test-case

In this section, we apply the new method proposed in Section 6.4 to the Ishigami function, see [IH90b] :

$$f(X_1, X_2, X_3) = \sin(X_1) + 7\sin^2(X_2) + 0.1X_3^4\sin(X_1)$$

where the X_i's are independent random variables uniformly distributed on $[-\pi, \pi]$. Analytical values of Sobol' indices of this model are

$$\underline{S}_1 = 0.3139, \quad \underline{S}_2 = 0.4424, \quad \underline{S}_3 = 0, \quad \underline{S}_{12} = 0.7563, \quad \underline{S}_{23} = 0.4424, \quad \underline{S}_{13} = 0.5575 \text{ and } \underline{S}_{123} = 1.$$

We are interested in comparing the new method, with the classic one based on crude Monte Carlo method and which need $d+1$ samples to estimate all the first-order Sobol' indices, and $2d+2$ samples to estimate all the second-order Sobol' indices, see [Sal02]). Here, both methods are compared at the same sample size n in order to investigate the estimators themselves, but keep in mind that the new method is definitely more efficient since only two samples are needed to estimate all the first-order Sobol' indices or all the second-order Sobol' indices. In the experiment, we focus on the empirical coverage — i.e. the empirical proportion of confidence interval containing the analytical value of the Sobol' index — of both estimators at different sample sizes between 10^2 and 10^5, and for $r = 100000$ replicates. We first investigate estimators $\widehat{\underline{S}}_{\{i\},n}$ and $\widehat{\underline{S}}^{RLHS}_{\{i\},n}$, $i \in \{1, \ldots, d\}$ and in both cases, we provide asymptotic confidence intervals from the estimation of the asymptotic variance given in [JKL+12] (see end of the proof of Prop. 2.2). Indeed, as we know that this asymptotic variance is :

$$\sigma^2_{IID,\mathfrak{u}} = \frac{\mathrm{Var}\big[(Y^1_{\mathfrak{u}} - \mathbb{E}[Y^1_{\mathfrak{u}}])(Y^2_{\mathfrak{u}} - \mathbb{E}[Y^1_{\mathfrak{u}}]) - \underline{S}_{\mathfrak{u}}/2\big((Y^1_{\mathfrak{u}} - \mathbb{E}[Y^1_{\mathfrak{u}}])^2\big)\big(Y^2_{\mathfrak{u}} - \mathbb{E}[Y^1_{\mathfrak{u}}]\big)\big]}{\mathrm{Var}[Y]^2} \geq \sigma^2_{RLHS,\mathfrak{u}}, \quad (6.19)$$

we can provide an estimator of the asymptotic confidence interval for the classic method

$$I_{IID,\mathfrak{u},\alpha} = \Big[\underline{S}_{\mathfrak{u}} - \frac{\sigma^2_{IID,\mathfrak{u}}\, u_{\alpha/2}}{\sqrt{n}}, \underline{S}_{\mathfrak{u}} + \frac{\sigma^2_{IID,\mathfrak{u}}\, u_{\alpha/2}}{\sqrt{n}}\Big]$$

and an other one for the new method

$$I_{RLHS,\mathfrak{u},\alpha} = \Big[\underline{S}_{\mathfrak{u}} - \frac{\sigma^2_{RLHS,\mathfrak{u}}\, u_{\alpha/2}}{\sqrt{n}}, \underline{S}_{\mathfrak{u}} + \frac{\sigma^2_{RLHS,\mathfrak{u}}\, u_{\alpha/2}}{\sqrt{n}}\Big]$$

where $u_{\alpha/2}$ is the normal quantile at the significance level α. By using the estimator of the asymptotic variance given in (6.19) in both cases, the confidence interval lengths of the classic and the new estimators are the same. More specifically, the estimated length of the new estimator is greater or equal than its optimal value. Thus the asymptotic value of the empirical coverage of the new method is greater or equal than the expected one. However at the moment, we do not know how to estimate correctly $\sigma^2_{RLHS,\mathfrak{u}}$ because of its singular expression (see Proof of (ii) in Proposition 6.1 in Section 6.3.2). We just say few words about it in the next subsection and more fundamentally, it should consist of a further work.

We also investigate estimators $\widehat{\underline{S}}_{\{i,j\},n}$ and $\widehat{\underline{S}}^{OA2-RLHS}_{\{i,j\},n}$, $i \neq j \in \{1, \ldots, d\}$, where the notation $OA2 - RLHS$ refers to the generalization to replicated latin hypercube based on orthogonal array of strength 2 presented in Section 6.4.4. In this case, we conjecture that the Central Limit theorem established in (ii) in Proposition 6.2 is also true under some smoothness assumption — note that, here, Ishigami function is $\mathcal{C}^\infty$. Results are gathered in Figures 6.1 to 6.4. For the second-order Sobol' indices, we can observe that the bivariate stratification has a bad effect on the new estimator at low sample size, but we can notice its good properties as the number of simulations increases.

Remark on the confidence interval length of the new estimator

Concerning the estimation of the right confidence interval length of the new estimators, note that if the asymptotic empirical coverage — estimated using Formula (6.19) — is $1 - \alpha'$ instead of the expected value $1-\alpha$, then it means that the true asymptotic confidence interval should be $u_{\alpha/2}/u_{\alpha'/2}$ time as long, where $u.$ denote the normal quantiles. More specifically in our first application, we obtain in this way the true asymptotic normalized ($\times\sqrt{n}$) confidence interval length of $\underline{S}_1$, $\underline{S}_2$, $\underline{S}_{12}$, $\underline{S}_{13}$ and $\underline{S}_{23}$; they are gathered in Table 6.1. Moreover considering these right normalized confidence interval lengths, we can observe on Figures 6.5 and 6.6 that the empirical coverage of the new estimator converges to the expected level 0.99 as n increases, and so we confirm the reliability of the empirical confidence intervals constructed with the true asymptotic length. Unfortunately, evaluating the true asymptotic confidence interval length is infeasible in practice since it requires a lot of replications to estimate the empirical coverage. So the issue related to the construction of optimal confidence intervals remains open.

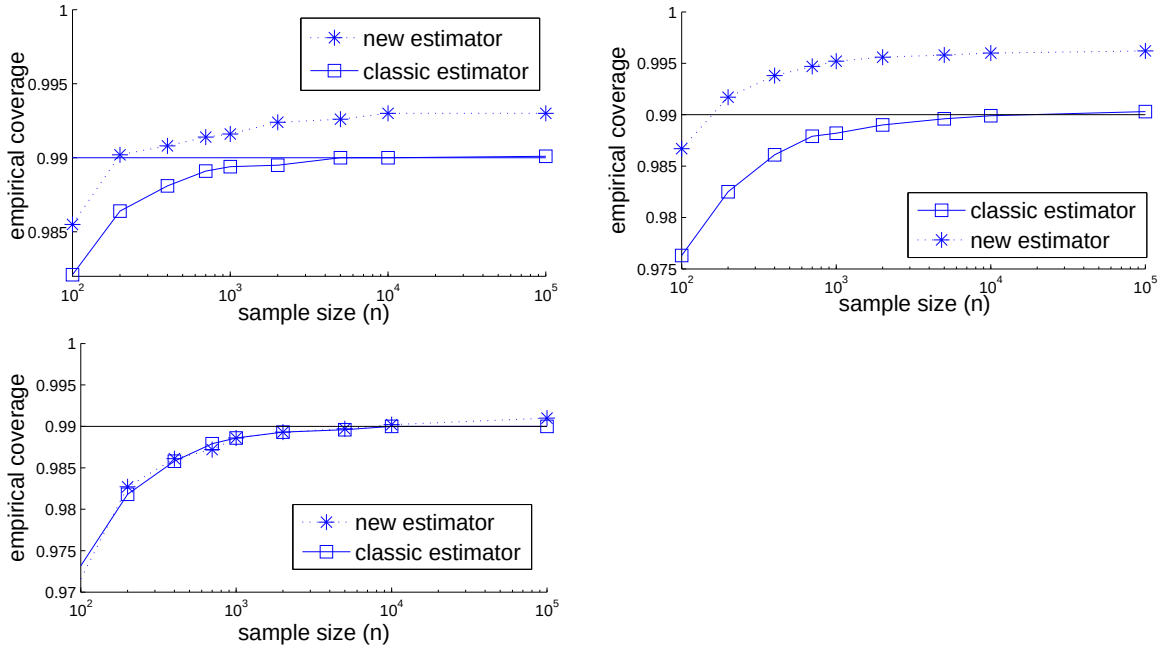


FIGURE 6.1 – Empirical coverage of confidence intervals for $\underline{S}_1$ (top left), $\underline{S}_2$ (top right) and $\underline{S}_3$ (bottom).

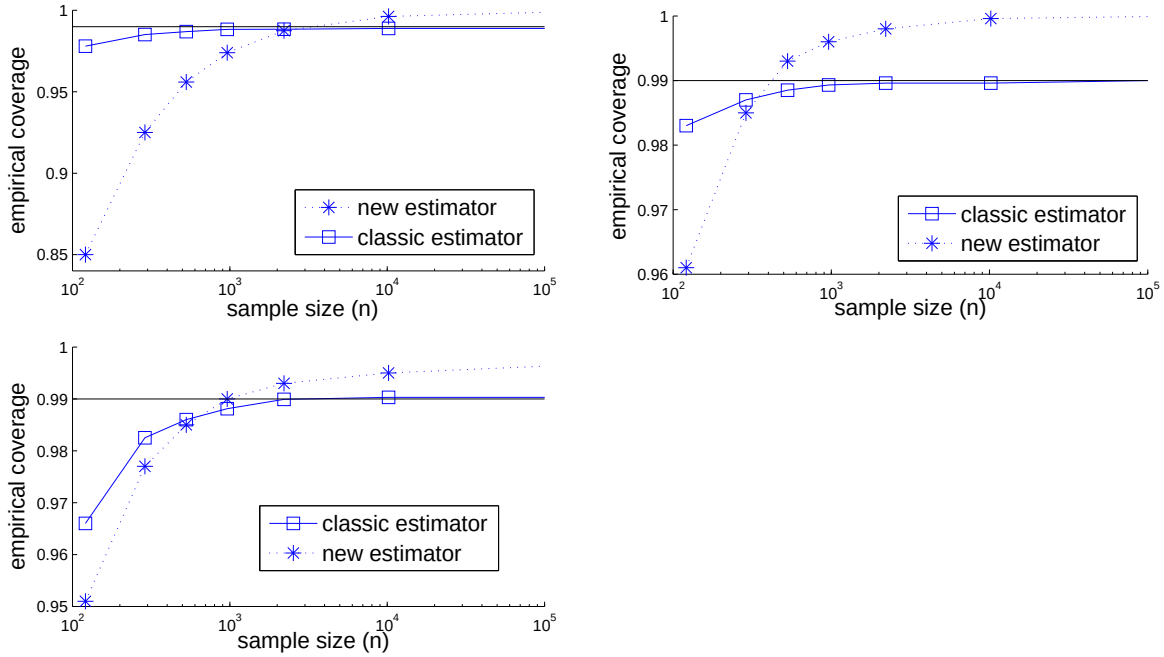


FIGURE 6.2 – Empirical coverage of confidence intervals for $\underline{S}_{12}$ (top left), $\underline{S}_{13}$ (top right) and $\underline{S}_{23}$ (bottom).

6.5.2 Application to a marine ecosystem simulator

We now illustrate the new method to a one-dimensional coupled hydrodynamical– biological model developed and applied to the Ligurian Sea (northwestern Mediterranean). This ecosystem simulator,

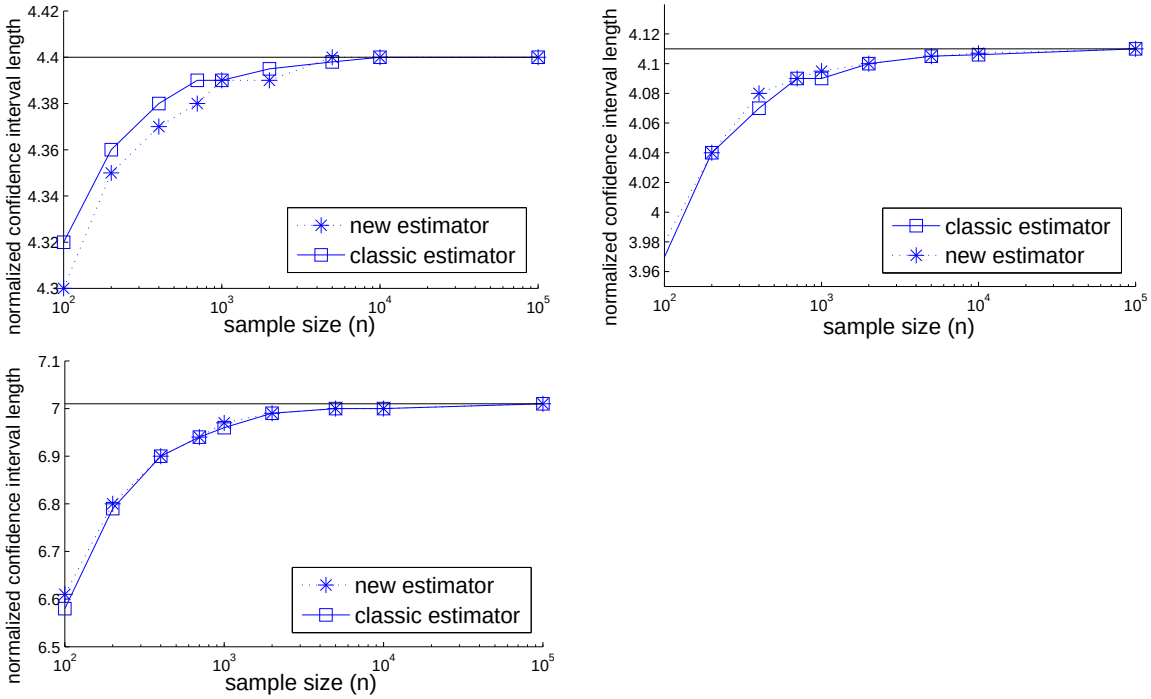


FIGURE 6.3 – Normalized length of the empirical interval for $\underline{S}_1$ (top left), $\underline{S}_2$ (top right) and $\underline{S}_3$ (bottom).

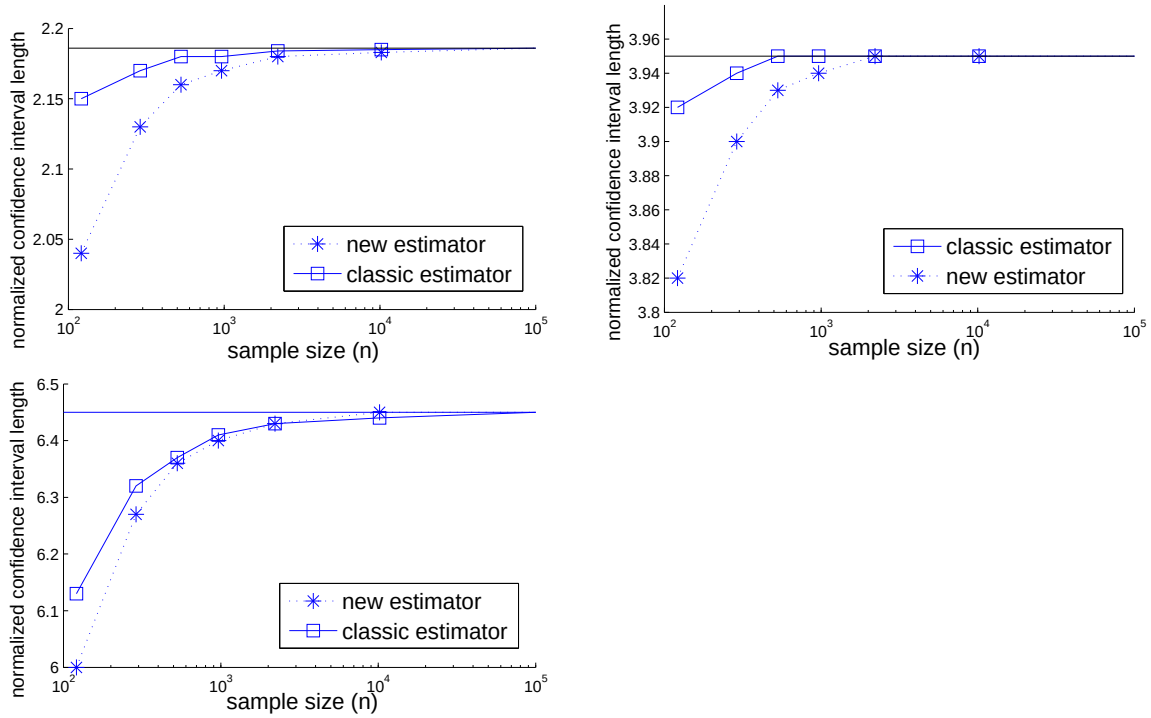


FIGURE 6.4 – Normalized length of the empirical interval for $\underline{S}_{12}$ (top left), $\underline{S}_{13}$ (top right) and $\underline{S}_{23}$ (bottom).

MODèle d'ÉCOsystème du GHER et du LOBEPM[1] (MODECOGeL), combines a 1D (vertical)

1. GHER : GeoHydrodynamics and Environment Research, Université de Liège, Belgium. LOBEPM : Laboratoire d'Océanologie Biologique et d'Écologie du Plancton Marin, Université Pierre et Marie Curie, France

	$\underline{S}_1$	$\underline{S}_2$	$\underline{S}_{12}$	$\underline{S}_{13}$	$\underline{S}_{23}$
estimated lengths using (6.19)	4.40	4.15	2.19	3.95	6.45
right lengths	3.96	3.28	1.53	2.37	5.16

TABLE 6.1 – Comparison between confidence interval lengths estimated using (6.19) and the right lengths for $\underline{S}_1$, $\underline{S}_2$, $\underline{S}_{12}$, $\underline{S}_{13}$ and $\underline{S}_{23}$

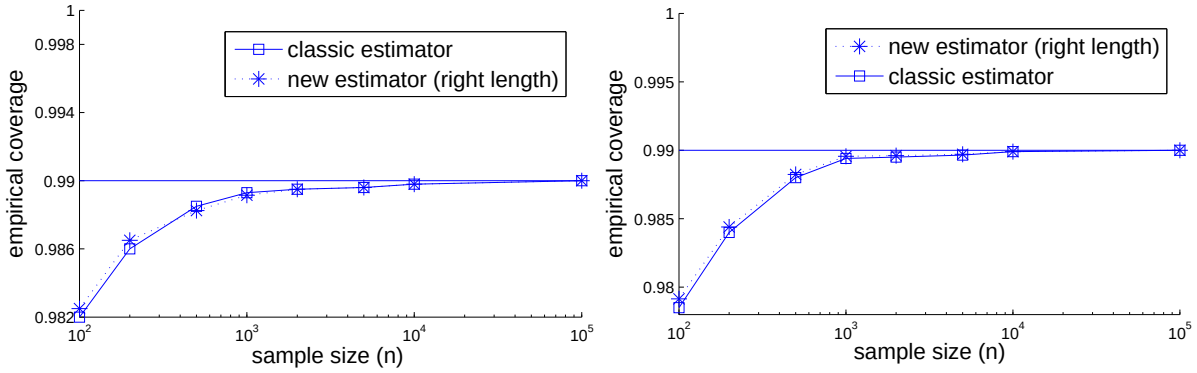


FIGURE 6.5 – Empirical coverage of confidence intervals for $\underline{S}_1$ (left) and $\underline{S}_2$ (right).

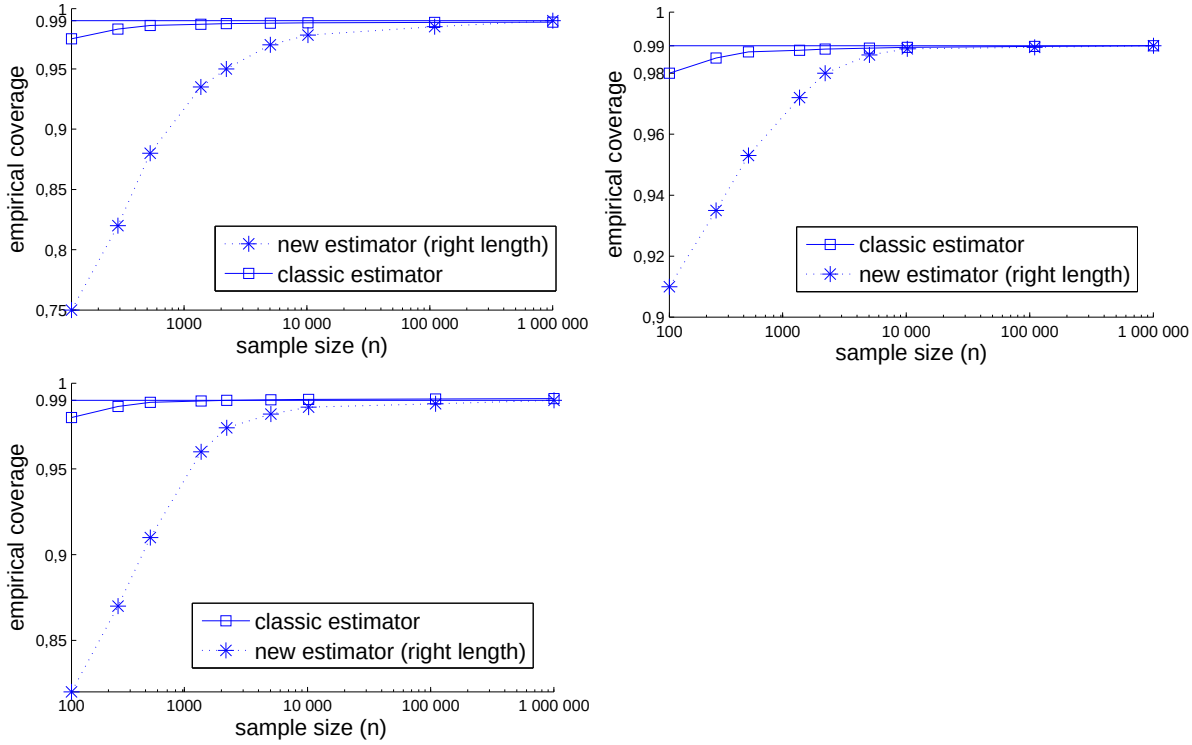


FIGURE 6.6 – Empirical coverage of confidence intervals for $\underline{S}_{12}$ (top left), $\underline{S}_{13}$ (top right) and $\underline{S}_{23}$ (bottom).

version of the 3D GHER model which takes into account momentum and heat surface fluxes computed from a real meteorological data set, and a biogeochemical model defined by a nitrogen cycle of 12 biological state variables (see Figure 6.7) controlled by 87 input parameters (see [LN98]). Here we focus on the chlorophyll-a concentration which is defined as a function of time and depth

$$\texttt{chla}(t,z) = 1.59 * \big(\texttt{pp}(t,z) + \texttt{np}(t,z) + \texttt{mp}(t,z)\big)$$

where pp, np and mp are the phyto-, nano- and microphytoplankton biomasses, respectively. The

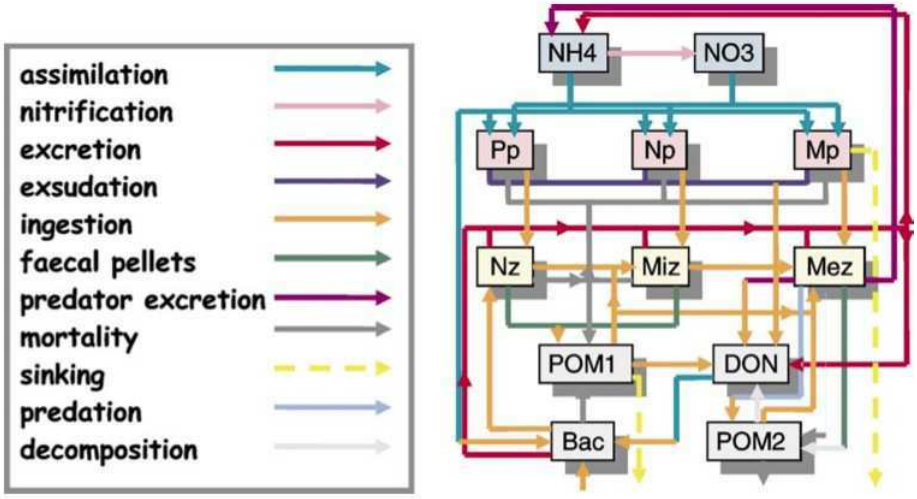


FIGURE 6.7 – Biogeochemical model (NH4 : Ammonium ; NH3 : nitrate ; Pp, Np, Mp : pico-, nano-, microphytoplankton ; Nz, Miz, Mez : nano-, micro-, mesozooplankton ; PON1, PON2 : type 1 and 2 particulate organic nitrogen ; Bac : bacteria ; DON : dissolved organic nitrogen).

behavior of these three state variables are modeled by the following reaction-diffusion and reaction-advection-diffusion equation

$$\begin{aligned}
\frac{\partial\,\mathtt{pp}}{\partial t} &= \frac{\partial}{\partial z}\left(\lambda\frac{\partial\,\mathtt{pp}}{\partial z}\right) + \big((1-exud_{\mathtt{pp}})\mu_{\mathtt{pp}} - mort_{\mathtt{pp}}\big)\mathtt{pp} - ing_{\mathtt{pp},\mathtt{nz}}\mathtt{nz}\\
\frac{\partial\,\mathtt{np}}{\partial t} &= \frac{\partial}{\partial z}\left(\lambda\frac{\partial\,\mathtt{np}}{\partial z}\right) + \big((1-exud_{\mathtt{np}})\mu_{\mathtt{np}} - mort_{\mathtt{np}}\big)\mathtt{np} - ing_{\mathtt{np},\mathtt{miz}}\mathtt{miz}\\
\frac{\partial\,\mathtt{mp}}{\partial t} &= \frac{\partial}{\partial z}\left(\lambda\frac{\partial\,\mathtt{mp}}{\partial z}\right) + \big((1-exud_{\mathtt{mp}})\mu_{\mathtt{mp}} - mort_{\mathtt{mp}}\big)\mathtt{mp} - ing_{\mathtt{mp},\mathtt{mez}}\mathtt{mez} - sin_{\mathtt{mp}}\frac{\partial\mathtt{mp}}{\partial z}
\end{aligned}$$

where nz, miz and mez are the nano-, micro- and mesozooplankton biomasses, respectively, and the other notations are

λ	vertical turbulent diffusivity ($\mathrm{m}^2.\mathrm{s}^{-1}$)
$exud_A$	exudation of A (percentage)
μ_A	growth rate of A (day^{-1})
$mort_A$	mortality rate of A (day^{-1})
$ing_{A,B}$	ingestion rate of A by predator B (mgChl)
$sin_{\mathtt{mp}}$	sinking velocity of microphytoplankton ($\mathrm{m.day}^{-1}$)

In our experiment, we focus on two different outputs : the annual maximum of chlorophyll-a concentration in surface water Y_{surf} and the annual maximum of the mean of cholorphyll-a concentration between 20 and 50 meters in depth Y_{depth}. These are practical indicators of biological activity. We are interested in the influence of eight parameters among the 87 input factors. On the one hand, we consider 6 a priori influent parameters $\mu_{max\mathtt{pp}}$, $\mu_{max\mathtt{np}}$, $\mu_{max\mathtt{mp}}$, $I_{opt\mathtt{pp}}$, $I_{opt\mathtt{np}}$ and $I_{opt\mathtt{mp}}$ where μ_{maxA} and I_{optA} denote the maximum growth rate of A and the optimum insolation for A, respectively. These input factors are directly related to the growth rate of A, μ_A (see details in Appendix C). On the other hand, we consider the maximum growth rate of bacteria $\mu_{max\mathtt{bac}}$ and the sinking velocity of particulate organic nitrogen (type 1) $sin_{\mathtt{pon1}}$ which have a priori a negligible effect on chlorophyll-a concentration since they do not act directly on pp, np and mp but on the state variables bac and pon1. We take these eight parameters to be independent gamma distributed random variables with parameters given in Table 6.2. We estimate all first- and second-order Sobol' indices of both ouputs Y_{surf} and Y_{depth} by using the estimators defined in Sections 6.4.3 and 6.4.4 with sample sizes $n = 65536$ and $n = 66049$, respectively.

	label	k	θ	mean	standard deviation
$\mu_{max\texttt{pp}}$ (day^{-1})	1	9	0.33	3	1
$\mu_{max\texttt{np}}$ (day^{-1})	2	9	0.28	2.5	0.83
$\mu_{max\texttt{mp}}$ (day^{-1})	3	9	0.22	2	0.67
$I_{opt\texttt{pp}}$ (W.m^{-2})	4	9	1.11	10	3.33
$I_{opt\texttt{np}}$ (W.m^{-2})	5	9	1.67	15	5
$I_{opt\texttt{mp}}$ (W.m^{-2})	6	9	2.22	20	6.67
$\mu_{max\texttt{bac}}$ (day^{-1})	7	9	0.22	2	0.67
$sin_{\texttt{pon1}}$ (m.day^{-1})	8	9	0.17	1.5	0.5

TABLE 6.2 – Distributions of variables using gamma density $f(x;k,\theta) = x^{k-1}\exp(-x/\theta)/(\Gamma(\theta)\theta^k)$, where $\Gamma(\cdot)$ is the gamma function.

The first-order Sobol' indices are estimated by using nested replicated latin hypercubes following Qian's construction [Qia09]. They allow to visualize empirical convergence of the estimated indices as shown in Figure 6.8. The estimated indices at the biggest sample size ($n = 65536$) are reported in Tables 6.3 and 6.4; we can notice that both outputs do not define an additive model since in both cases, the sum of the first-order Sobol' indices are less than sixty percents. We also notice that $\mu_{max\texttt{pp}}$ is important in both outputs, while three other a priori important parameters — $\mu_{max\texttt{np}}$, $I_{opt\texttt{np}}$ and $I_{opt\texttt{mp}}$— have actually no effect. At last, it is surprising to observe that the parameter $\mu_{max\texttt{bac}}$, which does not act directly on both outputs, has non-zero values.

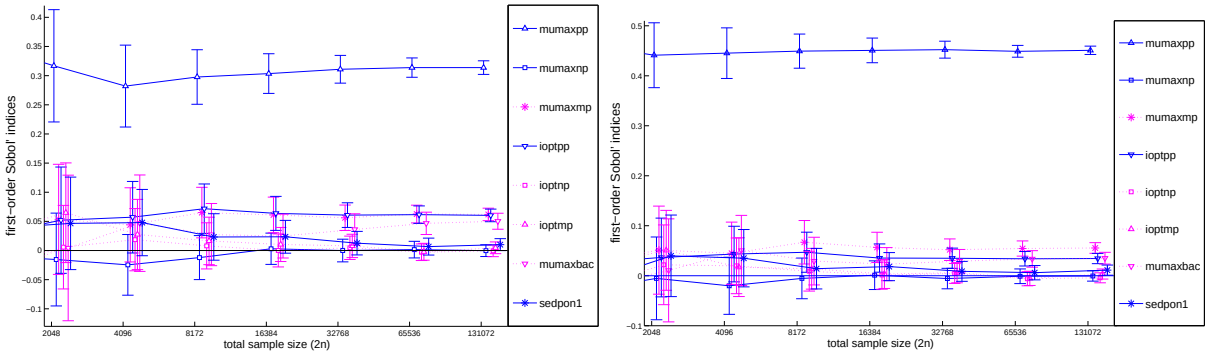


FIGURE 6.8 – Plots of first-order Sobol' indices with error bars — 99% confidence interval — for both outputs Y_{surf} (left) and Y_{depth} (right).

The second-order Sobol' indices are estimated by using a replicated latin hypercube based on an orthogonal array with 257 levels, index 1 and strength 2 — i.e. $n = 66049$ — following Bose's construction [Bos38]. The results are reported in Tables 6.5 and 6.6; they confirm that $\mu_{max\texttt{pp}}$ has the main role in both outputs since the non-negligible second-order Sobol' indices are all related to the latter. As a conclusion, we can notice that both outputs are extremely complex and contain, without any doubt, interactions of order more than or equal to 3. Such an analysis with the Monte Carlo estimator of Sobol' indices would be less efficient without the new approach we proposed in this paper. More precisely, both order 1 and order 2 analysis using the classic Monte Carlo estimator — i.e. estimating all the Sobol' indices of order 1 or 2 — only could use a sample size of 30000 instead of 132000 since this classic approach needs 9 independent samples while the new one only needs 2 for the order 1 analysis and 18 independent samples while the new one only needs 4 for the order 2 analysis (see [Sal02]).

6.6 Conclusion

We have introduced a new method to estimate all the k-th order Sobol' indices by using only 2 samples, for any k. This outperforms existing methods including the combinatorial results established

	$\underline{S}_{\{1\}}$	$\underline{S}_{\{2\}}$	$\underline{S}_{\{3\}}$	$\underline{S}_{\{4\}}$	$\underline{S}_{\{5\}}$	$\underline{S}_{\{6\}}$	$\underline{S}_{\{7\}}$	$\underline{S}_{\{8\}}$
estimated index	0.314	0	0.061	0.060	0	0.003	0.051	0.010
estimated error	0.010	0.011	0.012	0.011	0.010	0.010	0.013	0.012

TABLE 6.3 – Estimation of first-order Sobol' indices for the output Y_{surf}. The estimated error is the radius of the 99% confidence interval.

	$\underline{S}_{\{1\}}$	$\underline{S}_{\{2\}}$	$\underline{S}_{\{3\}}$	$\underline{S}_{\{4\}}$	$\underline{S}_{\{5\}}$	$\underline{S}_{\{6\}}$	$\underline{S}_{\{7\}}$	$\underline{S}_{\{8\}}$
estimated index	0.451	0	0.055	0.034	0	0	0.035	0.011
estimated error	0.009	0.010	0.010	0.010	0.010	0.010	0.012	0.010

TABLE 6.4 – Estimation of first-order Sobol' indices for the output Y_{depth}. The estimated error is the radius of the 99% confidence interval.

	$\underline{S}_{\{1,2\}}$	$\underline{S}_{\{1,3\}}$	$\underline{S}_{\{1,4\}}$	$\underline{S}_{\{1,5\}}$	$\underline{S}_{\{1,6\}}$	$\underline{S}_{\{1,7\}}$	$\underline{S}_{\{1,8\}}$	$\underline{S}_{\{2,3\}}$	$\underline{S}_{\{2,4\}}$	$\underline{S}_{\{2,5\}}$
estimated index	0.374	0.479	0.424	0.339	0.324	0.400	0.318	0.069	0.066	0.016
estimated error	0.012	0.011	0.013	0.011	0.011	0.010	0.011	0.011	0.011	0.011

	$\underline{S}_{\{2,6\}}$	$\underline{S}_{\{2,7\}}$	$\underline{S}_{\{2,8\}}$	$\underline{S}_{\{3,4\}}$	$\underline{S}_{\{3,5\}}$	$\underline{S}_{\{3,6\}}$	$\underline{S}_{\{3,7\}}$	$\underline{S}_{\{3,8\}}$	$\underline{S}_{\{4,5\}}$	$\underline{S}_{\{4,6\}}$
estimated index	0.015	0.069	0.015	0.125	0.074	0.075	0.128	0.072	0.077	0.070
estimated error	0.010	0.015	0.010	0.011	0.011	0.011	0.013	0.011	0.011	0.011

	$\underline{S}_{\{4,7\}}$	$\underline{S}_{\{4,8\}}$	$\underline{S}_{\{5,6\}}$	$\underline{S}_{\{5,7\}}$	$\underline{S}_{\{5,8\}}$	$\underline{S}_{\{6,7\}}$	$\underline{S}_{\{6,8\}}$	$\underline{S}_{\{7,8\}}$
estimated index	0.121	0.066	0.017	0.055	0.014	0.056	0.009	0.050
estimated error	0.013	0.011	0.010	0.015	0.010	0.014	0.010	0.015

TABLE 6.5 – Estimation of second-order Sobol' indices for the output Y_{surf}. The estimated error is the radius of the 99% confidence interval.

	$\underline{S}_{\{1,2\}}$	$\underline{S}_{\{1,3\}}$	$\underline{S}_{\{1,4\}}$	$\underline{S}_{\{1,5\}}$	$\underline{S}_{\{1,6\}}$	$\underline{S}_{\{1,7\}}$	$\underline{S}_{\{1,8\}}$	$\underline{S}_{\{2,3\}}$	$\underline{S}_{\{2,4\}}$	$\underline{S}_{\{2,5\}}$
estimated index	0.506	0.593	0.510	0.455	0.450	0.515	0.447	0.056	0.034	0.005
estimated error	0.010	0.009	0.010	0.009	0.009	0.008	0.009	0.011	0.011	0.011

	$\underline{S}_{\{2,6\}}$	$\underline{S}_{\{2,7\}}$	$\underline{S}_{\{2,8\}}$	$\underline{S}_{\{3,4\}}$	$\underline{S}_{\{3,5\}}$	$\underline{S}_{\{3,6\}}$	$\underline{S}_{\{3,7\}}$	$\underline{S}_{\{3,8\}}$	$\underline{S}_{\{4,5\}}$	$\underline{S}_{\{4,6\}}$
estimated index	0.008	0.055	0.009	0.087	0.057	0.064	0.109	0.063	0.041	0.043
estimated error	0.010	0.014	0.010	0.011	0.011	0.011	0.013	0.011	0.010	0.010

	$\underline{S}_{\{4,7\}}$	$\underline{S}_{\{4,8\}}$	$\underline{S}_{\{5,6\}}$	$\underline{S}_{\{5,7\}}$	$\underline{S}_{\{5,8\}}$	$\underline{S}_{\{6,7\}}$	$\underline{S}_{\{6,8\}}$	$\underline{S}_{\{7,8\}}$
estimated index	0.082	0.041	0.009	0.040	0.007	0.046	0.006	0.041
estimated error	0.013	0.010	0.010	0.014	0.010	0.014	0.010	0.014

TABLE 6.6 – Estimation of second-order Sobol' indices for the output Y_{depth}. The estimated error is the radius of the 99% confidence interval.

by Saltelli in 2002 [Sal02]. We derive theoretical results in the particular case of first-order Sobol' indices from the work by Janon et al. [JKL+12] on asymptotical properties of Sobol' indices and from the work by Loh [Loh96] on asymptotical properties of LHS. A further work will consists in deriving these theoretical results to higher-order Sobol' indices.

Acknowledgments

The authors thank Pierre Brasseur, Jean-Michel Brankart and Eric Blayo for valuable discussions on the simulator MODECOGeL and more generally on marine ecosystem models. They also thank Art Owen for his helpful comments. This work has been partially supported by French National Research Agency (ANR) through COSINUS program (project COSTA-BRAVA n° ANR-09-COSI-015).

6.A Lemmas for Proposition 6.1

Let $\mathbf{X}^1$ and $\mathbf{X}^2$ two distinct points of a Latin hypercube of size n in $[0,1]^d$. For any function f defined on $[0,1]^d$, consider $Y^1 = f(\mathbf{X}^1)$ et $Y^2 = f(\mathbf{X}^2)$. In Theorem 1 in [Ste87], Stein gives the following result

Theorem 6.1. *If f is a square integrable function then as n tends to $+\infty$, we have*

$$\mathrm{Cov}(Y^1, Y^2) = -\frac{1}{n}\sum_{i=1}^{d}\sigma_i^2 + o(n^{-1}) \ .$$

In this section, we prove an analogous result with more general settings and without the asymptotic assumption on n (see Lemma 6.4).

6.A.1 Notation and definitions

For s and n in $\mathbb{N}^*$, define the partition of $[0,1)^s$ in elementary hypercubes of side $1/n$,

$$\mathcal{Q}_s(n) = \left\{ Q \subseteq [0,1)^s \mid Q = \prod_{i=1}^{s}[\alpha_i, \beta_i),\ \alpha_i \in \left\{0, \frac{1}{n}, \ldots, \frac{n-1}{n}\right\},\ \beta_i = \alpha_i + \frac{1}{n} \right\}.$$

For any square integrable function g defined on $[0,1)^s$, $s \leq d$, define the sequence with general term

$$u_n(g) = n^s \sum_{Q \in \mathcal{Q}_s(n)} \left(\int_Q g(\mathbf{x}) d\mathbf{x} \right)^2 ,\ n \in \mathbb{N}.$$

6.A.2 Preliminary results

The first lemma is the analogous result for Lebesgue integrability of a result given in Equation (A.4) in [Ste87] for Riemann integrability. The second one gives an important inequality which allows to work without asymptotic assumption on n. The last one consists in simplifying integrals under Latin hypercube sampling using the ANOVA decomposition.

Lemma 6.1. *If g is a square integrable function, the sequence $\big(u_n(g)\big)$ converges to $\int g^2(\mathbf{x})d\mathbf{x}$ as n tends to $+\infty$.*

Démonstration. Noting that

$$u_n(g) = \int g_n(\mathbf{x}) d\mathbf{x}$$

where

$$\forall \mathbf{x} \in [0,1)^s,\ g_n(\mathbf{x}) = \sum_{Q \in \mathcal{Q}_s(n)} \left(n^s \int_Q g(\mathbf{y}) d\mathbf{y} \right)^2 \mathbf{1}_Q(\mathbf{x})$$

Lemma 6.1 is a straightforward consequence of the dominated convergence theorem. So let us prove that there exists an integrable function h such that for all $n \in \mathbb{N}^*$, $|g_n| \leq h$ almost surely, and g_n converges pointwise to g^2, and the conclusion will follow.

First since g is a square integrable function, we have $|g(\mathbf{x})| \leq M$ a.s., and by their definition, the g_n's are as well. Hence their exists an integrable function $(h : \mathbf{x} \mapsto M)$ such that $|g_n| \leq h$ almost surely. Concerning the pointwise convergence, let us prove that for any $\mathbf{x} \in [0,1)^s$,

$$\forall \varepsilon > 0,\ \exists\ N > 0,\ \forall n \geq N,\ \left| \left(n^s \int_{Q_{\mathbf{x}}} g(\mathbf{y}) d\mathbf{y} \right)^2 - g^2(\mathbf{x}) \right| < \varepsilon$$

where $Q_{\mathbf{x}}$ is the set Q in $\mathcal{Q}_s(n)$ containing $\mathbf{x}$. This is obvious if g^2 is a simple function and we easily generalize the result to any g^2 since any measurable function is a pointwise limit of simple functions. □

Lemma 6.2. *The sequence* $\big(u_n(g)\big)$ *is dominated by* $\int g^2(\mathbf{x})d\mathbf{x}$.

Démonstration. Let $n \in \mathbb{N}^*$, the result is proved by showing that the sequence of general term $v_k(g) = u_{2^k n}(g)$ is increasing. In this case, by Lemma 6.1, we have $\lim v_k(g) = \int g^2(\mathbf{x})d\mathbf{x}$, and since v_k is increasing, all the terms of this sequence are dominated by $\int g^2(\mathbf{x})d\mathbf{x}$, hence $v_0(g) = u_n(g) \leq \int g^2(\mathbf{x})d\mathbf{x}$. To prove that the sequence $\big(v_k(g)\big)$ is increasing, note that

$$\begin{aligned} v_{k+1}(g) &= (2^{k+1}n)^s \sum_{Q \in \mathcal{Q}_s(2^{k+1}n)} \left(\int_Q g(\mathbf{x})d\mathbf{x} \right)^2 \\ &= (2^k n)^s \sum_{Q \in \mathcal{Q}_s(2^k n)} \left(2^s \sum_{P \in \mathcal{P}(Q, 2^{k+1}n)} \left(\int_P g(\mathbf{x})d\mathbf{x} \right)^2 \right) \end{aligned}$$

where $\mathcal{P}(Q, 2^{k+1}n) = \mathcal{Q}(2^{k+1}n) \cap Q$. Then by Jensen inequality, we have

$$2^s \sum_{P \in \mathcal{P}(Q, 2^{k+1}n)} \left(\int_P g(\mathbf{x})d\mathbf{x} \right)^2 \geq \left(\int_Q g(\mathbf{x})d\mathbf{x} \right)^2$$

and we conclude that $v_{k+1}(g) \geq v_k(g)$. □

For $0 \leq x_1,\ x_2 \leq 1$ define

$$\begin{aligned} r_n(x_1, x_2) &= 1 \ \text{ if } \lfloor nx_1 \rfloor = \lfloor nx_2 \rfloor \\ &= 0 \ \text{ otherwise}, \end{aligned}$$

where $\lfloor \cdot \rfloor$ is the floor function. We now end with the following result

Lemma 6.3. *Let* $\mathfrak{v}$ *be a subset of* $\{1, \ldots, d\}$, *we have*

$$\int f(\mathbf{x}_1) f(\mathbf{x}_2) \prod_{i \in \mathfrak{v}} r_n(x_{1i}, x_{2i}) d\mathbf{x}_1 d\mathbf{x}_2 = \int \sum_{\mathfrak{w} \subseteq \mathfrak{v}} f_{\mathfrak{w}}(\mathbf{x}_{1\mathfrak{w}}) f_{\mathfrak{w}}(\mathbf{x}_{2\mathfrak{w}}) \prod_{i \in \mathfrak{v}} r_n(x_{1i}, x_{2i}) d\mathbf{x}_1 d\mathbf{x}_2 \ .$$

Démonstration. By the ANOVA decomposition — see (6.1) — we have

$$\int f(\mathbf{x}_1) f(\mathbf{x}_2) \prod_{i \in \mathfrak{v}} r_n(x_{1i}, x_{2i}) d\mathbf{x}_1 d\mathbf{x}_2 = \int \sum_{\mathfrak{w}_1 \subseteq \{1,..,d\}} \sum_{\mathfrak{w}_2 \subseteq \{1,..,d\}} f_{\mathfrak{w}_1}(\mathbf{x}_{1\mathfrak{w}_1}) f_{\mathfrak{w}_2}(\mathbf{x}_{2\mathfrak{w}_2}) \prod_{i \in \mathfrak{v}} r_n(x_{1i}, x_{2i}) d\mathbf{x}_1 d\mathbf{x}_2 \ .$$

Then note that a certain number of terms in the member on the right-hand side vanishe. If $(\mathfrak{w}_1 \cap \mathfrak{v}^c) \cup (\mathfrak{w}_2 \cap \mathfrak{v}^c) \neq \emptyset$ then suppose without loss of generality that there exists $k \in \mathfrak{w}_1 \setminus \mathfrak{v}$; we have

$$\int f_{\mathfrak{w}_1}(\mathbf{x}_{1\mathfrak{w}_1}) f_{\mathfrak{w}_2}(\mathbf{x}_{2\mathfrak{w}_2}) \prod_{i \in \mathfrak{v}} r_n(x_{1i}, x_{2i}) d\mathbf{x}_1 d\mathbf{x}_2 = \int \underbrace{\left(\int f_{\mathfrak{w}_1}(\mathbf{x}_{1\mathfrak{w}_1}) dx_{1k} \right)}_{I_1} f_{\mathfrak{w}_2}(\mathbf{x}_{2\mathfrak{w}_2}) \prod_{i \in \mathfrak{v}} r_n(x_{1i}, x_{2i}) d\mathbf{x}_{1\{k\}^c} d\mathbf{x}_2$$

and note that, by a basic property of the ANOVA decomposition, $I_1 = 0$. If $(\mathfrak{w}_1 \cap \mathfrak{v}^c) \cup (\mathfrak{w}_2 \cap \mathfrak{v}^c) = \emptyset$ and $\mathfrak{w}_1 \neq \mathfrak{w}_2$, then suppose without loss of generality that there exists $k \in \mathfrak{w}_1 \setminus \mathfrak{w}_2$. In this case, we have

$$\int f_{\mathfrak{w}_1}(\mathbf{x}_{1\mathfrak{w}_1}) f_{\mathfrak{w}_2}(\mathbf{x}_{2\mathfrak{w}_2}) \prod_{i\in\mathfrak{v}} r_n(x_{1i}, x_{2i}) d\mathbf{x}_1 d\mathbf{x}_2$$

$$= \int \underbrace{\Big(\int f_{\mathfrak{w}_1}(\mathbf{x}_{1\mathfrak{w}_1}) r_n(x_{1k}, x_{2k}) dx_{1k} dx_{2k} \Big)}_{I_2} f_{\mathfrak{w}_2}(\mathbf{x}_{2\mathfrak{w}_2}) \Big(\prod_{i\in\mathfrak{v}\setminus\{k\}} r_n(x_{1i}, x_{2i}) \Big) d\mathbf{x}_{1\{k\}^c} d\mathbf{x}_{2\{k\}^c}$$

and note that by the definition of r_n, we have

$$\int f_{\mathfrak{w}_1}(\mathbf{x}_{1\mathfrak{w}_1}) r_n(x_{1k}, x_{2k}) dx_{1k} dx_{2k} = \int f_{\mathfrak{w}_1}(\mathbf{x}_{1\mathfrak{w}_1}) dx_{1k}$$

and thus $I_2 = 0$. The conclusion of the lemma follows. □

6.A.3 Main result

Let $\mathfrak{u}$ be a non-empty subset of $\{1, \dots d\}$ and consider $(\dot{\mathbf{Z}}_{\mathfrak{u}}^j)_j \sim \mathcal{LH}(n, 2d - |\mathfrak{u}|)$. For any function f defined on $[0,1]^d$, consider

$$\dot{Y}_{\mathfrak{u}}^{1,1} = f(\dot{\mathbf{X}}_{\mathfrak{u}}^1, \dot{\mathbf{X}}_{\mathfrak{u}^c}^{1,1}) \text{ and } \dot{Y}_{\mathfrak{u}}^{2,2} = f(\dot{\mathbf{X}}_{\mathfrak{u}}^2, \dot{\mathbf{X}}_{\mathfrak{u}^c}^{2,2}) \, .$$

We have the following result

Lemma 6.4. *If f is a square integrable function then we have*

$$- \sum_{\substack{\emptyset \neq \mathfrak{w} \subseteq \mathfrak{u} \\ |\mathfrak{w}| \ odd}} \frac{\sigma_{\mathfrak{w}}^2}{(n-1)^{|\mathfrak{w}|}} \leq \mathrm{Cov}(\dot{Y}_{\mathfrak{u}}^{1,1}, \dot{Y}_{\mathfrak{u}}^{2,2}) \leq \sum_{\substack{\emptyset \neq \mathfrak{w} \subseteq \mathfrak{u} \\ |\mathfrak{w}| \ even}} \frac{\sigma_{\mathfrak{w}}^2}{(n-1)^{|\mathfrak{w}|}} \, .$$

Démonstration. Recall that for $0 \leq x_1, \ x_2 \leq 1$,

$$\begin{aligned} r_n(x_1, x_2) &= 1 \ \text{ if } \lfloor n x_1 \rfloor = \lfloor n x_2 \rfloor \\ &= 0 \ \text{ otherwise} \, , \end{aligned}$$

where $\lfloor \cdot \rfloor$ is the floor function. For $\mathbf{x}_1 = (x_{11}, \dots, x_{1d})$ in $[0,1)^d$, define $\mathbf{x}_{1\mathfrak{v}} = (x_{1i_1}, \dots, x_{1i_{|\mathfrak{v}|}})$ where $\mathfrak{v} = \{i_1, \dots, i_{|\mathfrak{v}|}\}$. Due to the joint density of $(\dot{\mathbf{X}}_{\mathfrak{u}}^1, \dot{\mathbf{X}}_{\mathfrak{u}}^2)$ under Latin hypercube sampling — see [MCB79] or [Ste87] — and by Lemma 6.3, we have

$$\mathrm{Cov}(\dot{Y}_{\mathfrak{u}}^{1,1}, \dot{Y}_{\mathfrak{u}}^{2,2}) + \Big(\int f(\mathbf{x}) d\mathbf{x} \Big)^2 = \int f(\mathbf{x}_1) f(\mathbf{x}_2) \Big(\frac{n}{n-1} \Big)^{|\mathfrak{u}|} \prod_{i\in\mathfrak{u}} \big(1 - r_n(x_{1i}, x_{2i}) \big) d\mathbf{x}_1 d\mathbf{x}_2$$

$$\begin{aligned} &= \Big(\frac{n}{n-1} \Big)^{|\mathfrak{u}|} \sum_{\mathfrak{v}\subseteq\mathfrak{u}} (-1)^{|\mathfrak{v}|} \int f(\mathbf{x}_1) f(\mathbf{x}_2) \prod_{i\in\mathfrak{v}} r_n(x_{1i}, x_{2i}) d\mathbf{x}_1 d\mathbf{x}_2 \\ &= \Big(\frac{n}{n-1} \Big)^{|\mathfrak{u}|} \sum_{\mathfrak{v}\subseteq\mathfrak{u}} (-1)^{|\mathfrak{v}|} \int \sum_{\mathfrak{w}\subseteq\mathfrak{v}} f_{\mathfrak{w}}(\mathbf{x}_{1\mathfrak{w}}) f_{\mathfrak{w}}(\mathbf{x}_{2\mathfrak{w}}) \prod_{i\in\mathfrak{v}} r_n(x_{1i}, x_{2i}) d\mathbf{x}_1 d\mathbf{x}_2 \\ &= \Big(\frac{n}{n-1} \Big)^{|\mathfrak{u}|} \sum_{\mathfrak{v}\subseteq\mathfrak{u}} (-1)^{|\mathfrak{v}|} \sum_{\mathfrak{w}\subseteq\mathfrak{v}} \Big(\frac{1}{n} \Big)^{|\mathfrak{v}|-|\mathfrak{w}|} \int f_{\mathfrak{w}}(\mathbf{x}_{1\mathfrak{w}}) f_{\mathfrak{w}}(\mathbf{x}_{2\mathfrak{w}}) \prod_{i\in\mathfrak{w}} r_n(x_{1i}, x_{2i}) d\mathbf{x}_{1\mathfrak{w}} d\mathbf{x}_{2\mathfrak{w}} \quad (6.20) \end{aligned}$$

Then note that for any function of $\mathfrak{w}$ denoted by $A(\mathfrak{w})$, we have

$$\begin{aligned}
&\sum_{\mathfrak{v}\subseteq\mathfrak{u}}(-1)^{|\mathfrak{v}|}\sum_{\mathfrak{w}\subseteq\mathfrak{v}}\Big(\frac{1}{n}\Big)^{|\mathfrak{v}|-|\mathfrak{w}|}A(\mathfrak{w}) \\
&\qquad = \sum_{\mathfrak{v}\subseteq\mathfrak{u}}\Big(-\frac{1}{n}\Big)^{|\mathfrak{v}|}\sum_{\mathfrak{w}\subseteq\mathfrak{v}}\Big(\frac{1}{n}\Big)^{-|\mathfrak{w}|}A(\mathfrak{w}) \\
&\qquad = \sum_{\mathfrak{w}\subseteq\mathfrak{u}}\left(\sum_{k=0}^{|\mathfrak{u}|-|\mathfrak{w}|}\binom{|\mathfrak{u}|-|\mathfrak{w}|}{k}\Big(-\frac{1}{n}\Big)^{k+|\mathfrak{w}|}\right)\Big(\frac{1}{n}\Big)^{-|\mathfrak{w}|}A(\mathfrak{w}) \\
&\qquad = \sum_{\mathfrak{w}\subseteq\mathfrak{u}}\Big(\frac{n-1}{n}\Big)^{|\mathfrak{u}|-|\mathfrak{w}|}(-1)^{|\mathfrak{w}|}A(\mathfrak{w}) \qquad (6.21)
\end{aligned}$$

Hence, we deduce that

$$\mathrm{Cov}(\dot{Y}_{\mathfrak{u}}^{1,1},\dot{Y}_{\mathfrak{u}}^{2,2}) = \sum_{\substack{\mathfrak{w}\subseteq\mathfrak{u}\\ \mathfrak{w}\neq\emptyset}}\Big(\frac{n}{n-1}\Big)^{|\mathfrak{w}|}(-1)^{|\mathfrak{w}|}\int f_{\mathfrak{w}}(\mathbf{x}_{1\mathfrak{w}})f_{\mathfrak{w}}(\mathbf{x}_{2\mathfrak{w}})\prod_{i\in\mathfrak{w}}r_n(x_{1i},x_{2i})d\mathbf{x}_{1\mathfrak{w}}d\mathbf{x}_{2\mathfrak{w}}\ . \qquad (6.22)$$

Finally by the definition of r_n, we have

$$0\le\int f_{\mathfrak{w}}(\mathbf{x}_{1\mathfrak{w}})f_{\mathfrak{w}}(\mathbf{x}_{2\mathfrak{w}})\prod_{i\in\mathfrak{w}}r_n(x_{1i},x_{2i})d\mathbf{x}_{1\mathfrak{w}}d\mathbf{x}_{2\mathfrak{w}}\le\sum_{Q\in\mathcal{Q}_{|\mathfrak{w}|}(n)}\left(\int_Q f_{\mathfrak{w}}(\mathbf{x}_{1\mathfrak{w}})d\mathbf{x}_{1\mathfrak{w}}\right)^2$$

and by Lemma 6.2, this gives

$$0\le\int f_{\mathfrak{w}}(\mathbf{x}_{1\mathfrak{w}})f_{\mathfrak{w}}(\mathbf{x}_{2\mathfrak{w}})\prod_{i\in\mathfrak{w}}r_n(x_{1i},x_{2i})d\mathbf{x}_{1\mathfrak{w}}d\mathbf{x}_{2\mathfrak{w}}\le\frac{\sigma_{\mathfrak{w}}^2}{n^{|\mathfrak{w}|}}\ . \qquad (6.23)$$

The latter inequalities and (6.22) lead to Lemma 6.4. □

6.B Proof of (iii) in Proposition 6.2

We first give three lemmas. The proof of (iii) in Proposition 6.2 is given in Section 6.B.2.

6.B.1 Preliminary results

Lemma 6.5. *Let $d\in\mathbb{N}^*$, if $n\ge\frac{d^2}{2}$ then*

$$\left(1+\frac{1}{n}\right)^d-1\le\frac{d+1}{n}\ .$$

Démonstration. If $d=1$, the result is obvious. Otherwise, for any $x>0$, consider the function g_d defined by

$$g_d(x)=\left(1+\frac{1}{x}\right)^d-1-\frac{d+1}{x}\ .$$

We show that
(1) if there exists $x_0>0$ such that $g_d(x_0)\le 0$ then for all $x\ge x_0$, $g_d(x)\le 0$
(2) $g_d(d^2/2)\le 0$
and the conclusion follows. Concerning (1) note that

$$g_d(x)=1+\frac{d}{x}+O(x^{-2})-1-\frac{d}{x}-\frac{1}{x}=-\frac{1}{x}+O(x^{-2})$$

and then that g_d is negative as x tends to $+\infty$. Moreover for any $d > 1$, g_d is first decreasing and then increasing. Indeed, we have

$$g_d'(x) = -\frac{d}{x^2}\left(1+\frac{1}{x}\right)^{d-1} + \frac{d+1}{x^2}$$

and we deduce that $g_d'(x_0) = 0$ with

$$x_0 = \frac{1}{\left(\frac{d+1}{d}\right)^{1/(d-1)} - 1} > 0$$

and is negative on the left side and positive on the right side. The conclusion of (1) follows. Concerning (2), it is easy to check that it is true for $d = 1$ and 2, and for $d \geq 3$ we have

$$\begin{aligned}
g_d\left(\frac{d^2}{2}\right) &= \sum_{k=0}^{d}\binom{d}{k}\left(\frac{2}{d^2}\right)^k - 1 - \frac{2}{d} - \frac{2}{d^2} \\
&= -\frac{2}{d^3} + \sum_{k=3}^{d}\binom{d}{k}\left(\frac{2}{d^2}\right)^k \\
&\leq -\frac{2}{d^3} + \sum_{k=3}^{d}\frac{1}{k!}\left(\frac{2}{d}\right)^k \\
&\leq -\frac{2}{d^3} + \frac{1}{d^3} + \frac{1}{3d^3} + \frac{2}{3}\sum_{k=4}^{d}\frac{1}{d^k} \\
&\leq -\frac{2}{d^3} + \sum_{k=3}^{d}\frac{1}{d^k} + \frac{1}{3d^3}\left(1 - \sum_{k=1}^{d-3}\frac{1}{d^k}\right) \\
&\leq -\frac{2}{d^3} + \sum_{k=3}^{d}\frac{1}{d^k} \\
&\leq -\frac{2}{d^3} + \frac{2}{d^3}
\end{aligned}$$

and the conclusion follows. □

With the same notation as at the beginning of Section 6.A.3, we have the following result

Lemma 6.6. *If f is a square integrable function, we have*

$$\mathbb{E}\big[f(\dot{\mathbf{X}}^1_{\mathfrak{u}}, \dot{\mathbf{X}}^{1,1}_{\mathfrak{u}^c}) f(\dot{\mathbf{X}}^1_{\mathfrak{u}}, \dot{\mathbf{X}}^{2,1}_{\mathfrak{u}^c})\big] = \mathbb{E}[Y]^2 + \underline{\tau}^2_{\mathfrak{u}} + B_{\mathfrak{u},n}$$

where

$$-\mathbb{E}[Y^2] \sum_{\substack{\emptyset \neq \mathfrak{v} \subseteq \mathfrak{u}^c \\ |v| \ odd}} \frac{1}{(n-1)^{|v|}} \leq B_{\mathfrak{u},n} \leq \mathbb{E}[Y^2] \sum_{\substack{\emptyset \neq \mathfrak{v} \subseteq \mathfrak{u}^c \\ |v| \ even}} \frac{1}{(n-1)^{|v|}} \ . \tag{6.24}$$

Démonstration. First, due to the joint density of $(\dot{\mathbf{X}}^{1,1}_{\mathfrak{u}^c}, \dot{\mathbf{X}}^{2,1}_{\mathfrak{u}^c})$ under Latin hypercube sampling — see [MCB79] or [Ste87] — we have

$$\begin{aligned}
\mathbb{E}\big[f(\dot{\mathbf{X}}^1_{\mathfrak{u}}, \dot{\mathbf{X}}^{1,1}_{\mathfrak{u}^c}) f(\dot{\mathbf{X}}^1_{\mathfrak{u}}, \dot{\mathbf{X}}^{2,1}_{\mathfrak{u}^c})\big] &= \int f(\mathbf{x}, \mathbf{x}_1) f(\mathbf{x}, \mathbf{x}_2) \left(\frac{n}{n-1}\right)^{d-|\mathfrak{u}|} \prod_{i \in \mathfrak{u}^c} \big(1 - r_n(x_{1i}, x_{2i})\big) d\mathbf{x} d\mathbf{x}_1 d\mathbf{x}_2 \\
&= \left(\frac{n}{n-1}\right)^{d-|\mathfrak{u}|} \int \Bigg(\underbrace{\sum_{\mathfrak{v} \subseteq \mathfrak{u}^c} (-1)^{|\mathfrak{v}|} \int f(\mathbf{x}, \mathbf{x}_1) f(\mathbf{x}, \mathbf{x}_2) \prod_{i \in \mathfrak{v}} r_n(x_{1i}, x_{2i}) d\mathbf{x}_1 d\mathbf{x}_2}_{I(\mathbf{x})} \Bigg) d\mathbf{x} \ .
\end{aligned}$$

We now denote $f_{\mathbf{x}} : \mathbf{y} \mapsto f(\mathbf{x}, \mathbf{y})$ and then by (6.20) and (6.21) we have

$$\begin{aligned} I(\mathbf{x}) &= \sum_{\mathfrak{v} \subseteq \mathfrak{u}^c} (-1)^{|\mathfrak{v}|} \int f_{\mathbf{x}}(\mathbf{x}_1) f_{\mathbf{x}}(\mathbf{x}_2) \prod_{i \in \mathfrak{v}} r_n(x_{1i}, x_{2i}) d\mathbf{x}_1 d\mathbf{x}_2 \\ &= \sum_{\mathfrak{v} \subseteq \mathfrak{u}^c} (-1)^{|\mathfrak{v}|} \sum_{\mathfrak{w} \subseteq \mathfrak{v}} \Big(\frac{1}{n}\Big)^{|\mathfrak{v}|-|\mathfrak{w}|} \int f_{\mathbf{x},\mathfrak{w}}(\mathbf{x}_{1\mathfrak{w}}) f_{\mathbf{x},\mathfrak{w}}(\mathbf{x}_{2\mathfrak{w}}) \prod_{i \in \mathfrak{w}} r_n(x_{1i}, x_{2i}) d\mathbf{x}_{1\mathfrak{w}} d\mathbf{x}_{2\mathfrak{w}} \\ &= \sum_{\mathfrak{w} \subseteq \mathfrak{u}^c} (-1)^{|\mathfrak{w}|} \Big(\frac{n-1}{n}\Big)^{d-|\mathfrak{u}|-|\mathfrak{w}|} \int f_{\mathbf{x},\mathfrak{w}}(\mathbf{x}_{1\mathfrak{w}}) f_{\mathbf{x},\mathfrak{w}}(\mathbf{x}_{2\mathfrak{w}}) \prod_{i \in \mathfrak{w}} r_n(x_{1i}, x_{2i}) d\mathbf{x}_{1\mathfrak{w}} d\mathbf{x}_{2\mathfrak{w}} \ . \end{aligned}$$

Hence by (6.23) we have for all $\mathfrak{w} \neq \emptyset$,

$$0 \leq \int f_{\mathbf{x},\mathfrak{w}}(\mathbf{x}_{1\mathfrak{w}}) f_{\mathbf{x},\mathfrak{w}}(\mathbf{x}_{2\mathfrak{w}}) \prod_{i \in \mathfrak{w}} r_n(x_{1i}, x_{2i}) d\mathbf{x}_{1\mathfrak{w}} d\mathbf{x}_{2\mathfrak{w}} \leq \frac{\int f^2_{\mathbf{x},\mathfrak{w}}(\mathbf{x}_{1\mathfrak{w}}) d\mathbf{x}_{1\mathfrak{w}}}{n^{|\mathfrak{w}|}} \leq \frac{\int f^2_{\mathbf{x}}(\mathbf{x}_1) d\mathbf{x}_1}{n^{|\mathfrak{w}|}}$$

and note that

$$\int f_{\mathbf{x},\emptyset}(\mathbf{x}_{1\emptyset}) f_{\mathbf{x},\emptyset}(\mathbf{x}_{2\emptyset}) \prod_{i \in \emptyset} r_n(x_{1i}, x_{2i}) d\mathbf{x}_{1\emptyset} d\mathbf{x}_{2\emptyset} = f^2_{\mathbf{x},\emptyset} \ .$$

Finally, note that

$$\int f^2_{\mathbf{x},\emptyset} d\mathbf{x} = \underline{\tau}^2_{\mathfrak{u}} + \mathbb{E}[Y]^2$$

and

$$\int \int f^2_{\mathbf{x}}(\mathbf{x}_1) d\mathbf{x}_1 d\mathbf{x} = \mathbb{E}\big[Y^2\big]$$

and conclude that

$$\mathbb{E}\big[f(\dot{\mathbf{X}}^1_{\mathfrak{u}}, \dot{\mathbf{X}}^{1,1}_{\mathfrak{u}^c}) f(\dot{\mathbf{X}}^1_{\mathfrak{u}}, \dot{\mathbf{X}}^{2,1}_{\mathfrak{u}^c})\big] = \Big(\frac{n}{n-1}\Big)^{d-|\mathfrak{u}|} \int I(\mathbf{x}) d\mathbf{x} = \underline{\tau}^2_{\mathfrak{u}} + \mathbb{E}[Y]^2 + B_{\mathfrak{u},n}$$

with

$$-\mathbb{E}[Y^2] \sum_{\substack{\emptyset \neq \mathfrak{v} \subseteq \mathfrak{u}^c \\ |v| \text{ odd}}} \frac{1}{(n-1)^{|v|}} \leq B_{\mathfrak{u},n} \leq \mathbb{E}[Y^2] \sum_{\substack{\emptyset \neq \mathfrak{v} \subseteq \mathfrak{u}^c \\ |v| \text{ even}}} \frac{1}{(n-1)^{|v|}} \ .$$

□

Lemma 6.7. *The inequalities in Equation* (6.24) *imply that*

$$\Big| \sum_{\mathfrak{w} \subseteq u^c} \Big(\frac{1}{n}\Big)^{|\mathfrak{w}|} B_{\mathfrak{u}\cup\mathfrak{w},n} \Big| \leq \Big(\frac{d-|\mathfrak{u}|+1}{n} + 1\Big)\Big(\frac{d-|\mathfrak{u}|+1}{n-1}\Big)\mathbb{E}\big[Y^2\big] \ .$$

Démonstration. By (6.24), we have

$$\begin{aligned} \sum_{\mathfrak{w} \subseteq \mathfrak{u}^c} \Big(\frac{1}{n}\Big)^{|\mathfrak{w}|} B_{\mathfrak{u}\cup\mathfrak{w},n} &\leq \mathbb{E}\big[Y^2\big] \sum_{\mathfrak{w} \subseteq \mathfrak{u}^c} \Big(\frac{1}{n}\Big)^{|\mathfrak{w}|} \sum_{\emptyset \neq \mathfrak{v} \subseteq (\mathfrak{u}\cup\mathfrak{w})^c} \frac{1}{(n-1)^{|\mathfrak{w}|}} \\ &\leq \mathbb{E}\big[Y^2\big] \sum_{\mathfrak{w} \subseteq \mathfrak{u}^c} \Big(\frac{1}{n}\Big)^{|\mathfrak{w}|} \Big(\Big(1 + \frac{1}{n-1}\Big)^{d-|\mathfrak{u}|-|\mathfrak{w}|} - 1\Big) \\ &\leq \mathbb{E}\big[Y^2\big] \Big[\Big(1 + \frac{1}{n-1}\Big)^{d-|\mathfrak{u}|} \sum_{\mathfrak{w} \subseteq \mathfrak{u}^c} \Big(\frac{1}{n}\Big)^{|\mathfrak{w}|} - \sum_{\mathfrak{w} \subseteq \mathfrak{u}^c} \Big(\frac{1}{n}\Big)^{|\mathfrak{w}|}\Big] \\ &\leq \mathbb{E}\big[Y^2\big] \Big[\Big(1 + \frac{1}{n-1}\Big)^{d-|\mathfrak{u}|} \Big(1 + \frac{1}{n}\Big)^{d-|\mathfrak{u}|} - \Big(1 + \frac{1}{n}\Big)^{d-|\mathfrak{u}|}\Big] \end{aligned}$$

and the conclusion follows by applying twice Lemma 6.5. □

6.B.2 Proof of (iii) in Proposition 6.2

Démonstration. First note that by the definition of $(\ddot{\mathbf{Z}}^j_{\mathfrak{u}})_{j=1..n}$ we have

$$\ddot{Y}^{j,1}_{\mathfrak{u}} = f(\ddot{\mathbf{X}}^j_{\mathfrak{u}}, \ddot{\mathbf{X}}^{j,1}_{\mathfrak{u}^c}) \quad \text{and} \quad \ddot{Y}^{j,2}_{\mathfrak{u}} = f(\ddot{\mathbf{X}}^j_{\mathfrak{u}}, \ddot{\mathbf{X}}^{j,2}_{\mathfrak{u}^c})$$

with

$$\ddot{\mathbf{X}}^{j,1}_{\mathfrak{u}^c} = \left(\frac{\pi_1(j) - U_{1,\pi_1(j)}}{n}, \ldots, \frac{\pi_d(j) - U_{d,\pi_d(j)}}{n} \right)$$

and

$$\ddot{\mathbf{X}}^{j,2}_{\mathfrak{u}^c} = \left(\frac{\pi'_1(j) - U_{1,\pi'_1(j)}}{n}, \ldots, \frac{\pi'_d(j) - U_{d,\pi'_d(j)}}{n} \right)$$

where the π_i's, the π'_i's and the $U_{i,j}$'s are independent random variables uniformly distributed on Π_n — see Definition 6.1 —, Π_n and $[0,1]$, respectively. Moreover note that if for an index $i \in \mathfrak{u}^c$, we have $\pi_i(j) = \pi'_i(j)$ then $\ddot{X}^{j,1}_i = \ddot{X}^{j,2}_i$; and if $\pi_i(j) \neq \pi'_i(j)$ then $U_{i,\pi_i(j)}$ and $U_{i,\pi'_i(j)}$ are independent and therefore $\ddot{X}^{j,1}_i$ and $\ddot{X}^{j,2}_i$ are two distinct points of a Latin hypercube of size n in $[0,1]$. For $j \in \{1,\ldots,n\}$, denote by $\mathfrak{e}(j)$ the set of integers $i \in \mathfrak{u}^c$ such that $\pi_i(j) = \pi'_i(j)$. Thus we have

$$\begin{aligned}
\mathbb{E}\big[\ddot{Y}^{j,1}_{\mathfrak{u}} \ddot{Y}^{j,2}_{\mathfrak{u}}\big] \quad &= \quad \frac{1}{n^{2d-|\mathfrak{u}|}} \sum_{\mathfrak{w} \subseteq \mathfrak{u}^c} \sum_{\substack{\boldsymbol{\pi}(j) \in \\ \{1,\ldots,n\}^d}} \sum_{\substack{\boldsymbol{\pi}'_{\mathfrak{u}^c}(j) \in \\ \{1,\ldots,n\}^{d-|\mathfrak{u}|}}} \mathbf{1}_{\{\mathfrak{e}(j)=\mathfrak{w}\}} \cdots \\
&\cdots \quad \int f\left(\frac{\pi_1(j) - u_{11}}{n}, \ldots, \frac{\pi_d(j) - u_{1d}}{n} \right) f\left(\frac{\pi'_1(j) - u_{21}}{n}, \ldots, \frac{\pi'_d(j) - u_{2d}}{n} \right) d\mathbf{u}_1 d\mathbf{u}_{2(\mathfrak{u}\cup\mathfrak{w})^c}
\end{aligned}$$

where for all $i \in \mathfrak{u} \cup \mathfrak{e}(j)$, $\pi_i(j) = \pi'_i(j)$ and $u_{1i}(j) = u_{2i}(j)$. And noting that

$$\frac{1}{(n-1)^{d-|\mathfrak{u}|-|\mathfrak{w}|} n^d} \sum_{\substack{\boldsymbol{\pi}(j) \in \\ \{1,\ldots,n\}^d}} \sum_{\substack{\boldsymbol{\pi}'_{\mathfrak{u}^c}(j) \in \\ \{1,\ldots,n\}^{d-|\mathfrak{u}|}}} \mathbf{1}_{\{\mathfrak{e}(j)=\mathfrak{w}\}} \cdots$$

$$\cdots \int f\left(\frac{\pi_1(j) - u_{11}}{n}, \ldots, \frac{\pi_d(j) - u_{1d}}{n} \right) f\left(\frac{\pi'_1(j) - u_{21}}{n}, \ldots, \frac{\pi'_d(j) - u_{2d}}{n} \right) d\mathbf{u}_1 d\mathbf{u}_{2(\mathfrak{u}\cup\mathfrak{w})^c}$$

is equal to $\mathbb{E}\big[f(\dot{\mathbf{X}}^1_{\mathfrak{u}\cup\mathfrak{w}}, \dot{\mathbf{X}}^{1,1}_{(\mathfrak{u}\cup\mathfrak{w})^c}) f(\dot{\mathbf{X}}^1_{\mathfrak{u}\cup\mathfrak{w}}, \dot{\mathbf{X}}^{2,1}_{(\mathfrak{u}\cup\mathfrak{w})^c})\big]$ where $(\dot{\mathbf{X}}^j_{\mathfrak{u}\cup\mathfrak{w}}, \dot{\mathbf{X}}^{j,1}_{(\mathfrak{u}\cup\mathfrak{w})^c})_j \sim \mathcal{LH}(n,d)$, Lemma 6.6 gives

$$\mathbb{E}\big[\ddot{Y}^{j,1}_{\mathfrak{u}} \ddot{Y}^{j,2}_{\mathfrak{u}}\big] = \sum_{\mathfrak{w} \subseteq \mathfrak{u}^c} \left(\frac{1}{n} \right)^{|\mathfrak{w}|} \Big(\mathbb{E}[Y]^2 + \underline{\tau}^2_{\mathfrak{u}\cup\mathfrak{w}} + B_{\mathfrak{u}\cup\mathfrak{w},n} \Big) \, .$$

By Lemmas 6.5 and 6.7, and noting that $\mathbb{E}[Y]^2 + \underline{\tau}^2_{\mathfrak{u}\cup\mathfrak{w}} \leq \mathbb{E}[Y^2]$, we obtain

$$\mathbb{E}\big[\ddot{Y}^{j,1}_{\mathfrak{u}} \ddot{Y}^{j,2}_{\mathfrak{u}}\big] = \mathbb{E}[Y]^2 + \underline{\tau}^2_{\mathfrak{u}} + B_{|\mathfrak{u}|,n} \tag{6.25}$$

where

$$\big|B_{|\mathfrak{u}|,n}\big| \leq \left(\frac{d - |\mathfrak{u}| + 1}{n} + 2 \right) \left(\frac{d - |\mathfrak{u}| + 1}{n - 1} \right) \mathbb{E}[Y^2] \, . \tag{6.26}$$

Following the same proof, it is easy to show that for $j \neq l$, we have

$$\mathbb{E}\big[\ddot{Y}^{j,1}_{\mathfrak{u}} \ddot{Y}^{l,2}_{\mathfrak{u}}\big] = \mathbb{E}[Y]^2 + B_{n,1}$$

where

$$\big|B_{n,1}\big| \leq \left(\frac{d+1}{n} + 2 \right) \left(\frac{d+1}{n-1} \right) \mathbb{E}[Y^2] \, . \tag{6.27}$$

Thus noting that

$$\mathbb{E}\big[\underline{\tau}^{2,RLHS}_{\mathfrak{u},n}\big] = \frac{n-1}{n} \mathbb{E}[\ddot{Y}^{1,1}_{\mathfrak{u}} \ddot{Y}^{1,2}_{\mathfrak{u}}] - \frac{n-1}{n} \mathbb{E}[\ddot{Y}^{1,1}_{\mathfrak{u}} \ddot{Y}^{2,2}_{\mathfrak{u}}]$$

we conclude that

$$\mathbb{E}\big[\widetilde{\mathcal{I}}^{2,RLHS}_{\mathfrak{u},n}\big] = \underline{\mathcal{I}}^2_{\mathfrak{u}} - \frac{1}{n}\underline{\mathcal{I}}^2_{\mathfrak{u}} + \frac{n-1}{n}\big(B_{n,1} + B_{|\mathfrak{u}|,n}\big)$$

where the biases are $O(n^{-1})$ as specified above. Concerning $\widetilde{\sigma}^{2,RLHS}_n$, note that $\widetilde{\sigma}^{2,RLHS}_n = \widetilde{\sigma}^{2,LHS}_n$ and the conclusion follows from (iii) in Proposition 6.1. Concerning $\widehat{\mathcal{I}}^{2,RLHS}_{\mathfrak{u},n}$ and $\widehat{\sigma}^{2,RLHS}_n$, we have

$$\begin{aligned}\mathbb{E}\left[\left(\frac{1}{n}\sum_{j=1}^{n}\frac{\ddot{Y}^{j,1}_{\mathfrak{u}} + \ddot{Y}^{j,2}_{\mathfrak{u}}}{2}\right)^2\right] &= \frac{1}{4n}\mathbb{E}\big[(\ddot{Y}^{1,1}_{\mathfrak{u}} + \ddot{Y}^{1,2}_{\mathfrak{u}})^2\big] + \frac{1}{4n^2}\sum_{j=1}^{n}\sum_{\substack{l=1\\ l\neq j}}^{n}\mathbb{E}\big[(\ddot{Y}^{j,1}_{\mathfrak{u}} + \ddot{Y}^{j,2}_{\mathfrak{u}})(\ddot{Y}^{l,1}_{\mathfrak{u}} + \ddot{Y}^{l,2}_{\mathfrak{u}})\big] \\ &= \frac{1}{2n}\Big(\mathbb{E}\big[(\ddot{Y}^{1,1}_{\mathfrak{u}})^2\big] + \mathbb{E}[\ddot{Y}^{1,1}_{\mathfrak{u}}\ddot{Y}^{1,2}_{\mathfrak{u}}]\Big) + \frac{n-1}{2n}\Big(\mathbb{E}[\ddot{Y}^{1,1}_{\mathfrak{u}}\ddot{Y}^{2,1}_{\mathfrak{u}}] + \mathbb{E}[\ddot{Y}^{1,1}_{\mathfrak{u}}\ddot{Y}^{2,2}_{\mathfrak{u}}]\Big)\ .\end{aligned}$$

Then using notation in (6.6–6.8), note that

$$\mathbb{E}\big[\ddot{Y}^{1,1}_{\mathfrak{u}}\ddot{Y}^{2,1}_{\mathfrak{u}}\big] = \mathbb{E}\big[\dot{Y}^{1,1}_{\mathfrak{u}}\dot{Y}^{2,1}_{\mathfrak{u}}\big] = \mathrm{Cov}\big(\dot{Y}^{1,1}_{\mathfrak{u}}, \dot{Y}^{2,1}_{\mathfrak{u}}\big) + \mathbb{E}[Y]^2 = \mathrm{Cov}\big(\dot{Y}^{1,1}_{\{1,\ldots,d\}}, \dot{Y}^{2,2}_{\{1,\ldots,d\}}\big) + \mathbb{E}[Y]^2$$

and by (6.26), (6.27) and Lemma 6.4, we deduce

$$\mathbb{E}\left[\left(\frac{1}{n}\sum_{j=1}^{n}\frac{\ddot{Y}^{j,1}_{\mathfrak{u}} + \ddot{Y}^{j,2}_{\mathfrak{u}}}{2}\right)^2\right] = \frac{1}{2n}\underline{\mathcal{I}}^2_{\mathfrak{u}} + \mathbb{E}[Y]^2 + B_{n,1} + B_{n,2}\ .$$

where

$$|B_{n,2}| \leq \frac{\sigma^2}{2n}$$

and $B_{n,1}$ is specified in (6.27). Then it is easy to conclude that

$$\begin{aligned}\mathbb{E}\big[\widehat{\mathcal{I}}^{2,RLHS}_{\mathfrak{u},n}\big] &= \underline{\mathcal{I}}^2_{\mathfrak{u}} - \frac{1}{2n}\underline{\mathcal{I}}^2_{\mathfrak{u}} + B_{n,1} + B_{n,2} + \frac{n-1}{n}B_{|\mathfrak{u}|,n} \\ \mathbb{E}\big[\widehat{\sigma}^{2,RLHS}_n\big] &= \sigma^2 - \frac{1}{2n}\underline{\mathcal{I}}^2_{\mathfrak{u}} + B_{n,1} + B_{n,2}\end{aligned}$$

where the biases are $O(n^{-1})$ as specified above. □

6.C Phytoplankton growth model

The phytoplankton growth is given by the five following equations, where A stands for `pp`, `np` or `mp`.

$$\begin{aligned}\mu_A &= \mu_{max\mathrm{A}}\lim_{\mathrm{IA}}\lim_{\mathrm{TA}}(\lim_{\mathrm{NO_3A}} + \lim_{\mathrm{NH_4A}}) \\ \lim_{\mathrm{NO_3A}} &= \left(\frac{\mathrm{NO_3}}{\mathrm{NO_3} + \mathrm{K_{NO_3A}}}\right)\exp(-\Psi\mathrm{NH_4}) \\ \lim_{\mathrm{NH_4A}} &= \left(\frac{\mathrm{NH_4}}{\mathrm{NH_4} + \mathrm{K_{NH_4A}}}\right) \\ \lim_{\mathrm{IA}} &= \frac{2(1+\beta_{\mathrm{IA}})\frac{\mathrm{PAR}}{I_{optA}}}{\big(\frac{\mathrm{PAR}}{I_{optA}}\big)^2 + 2\beta_{\mathrm{IA}}\frac{\mathrm{PAR}}{I_{optA}} + 1} \\ \lim_{\mathrm{TA}} &= \max\left(\frac{2(1+\beta_{\mathrm{TA}})\frac{T-T_{letA}}{T_{optA}-T_{letA}}}{\big(\frac{T-T_{letA}}{T_{optA}-T_{letA}}\big)^2 + 2\beta_{\mathrm{IA}}\frac{T-T_{letA}}{T_{optA}-T_{letA}} + 1}, 0\right)\end{aligned}$$

where the parameters are defined in the following table

parameter	definition
μ_A	growth rate of A
$\mu_{max\mathrm{A}}$	maximum growth rate of A
$\mathrm{lim}_{\mathrm{NO_3A}}$	limitation by NO_3 for A
$\mathrm{lim}_{\mathrm{NH_4A}}$	limitation by NH_4 for A
$\mathrm{K}_{\mathrm{NO_3A}}$	half-saturation coefficient of NO_3 for A
$\mathrm{K}_{\mathrm{NH_4A}}$	half-saturation coefficient of NH_4 for A
Ψ	inhibition coefficient by NH_4
NO_3	NO_3 concentration
NH_4	NH_4 concentration
$\mathrm{lim}_{\mathrm{IA}}$	limitation by light for A
β_{IA}	shape factor for photoinhibition curve
$I_{opt\mathrm{A}}$	optimum insolation for A
PAR	photosynthetic active radiation
$\mathrm{lim}_{\mathrm{TA}}$	limitation by temperature for A
β_{TA}	shape factor for thermoinhibition curve
$T_{opt\mathrm{A}}$	optimum temperature for A
$T_{let\mathrm{A}}$	lower lethal temperature for A
T	temperature

TABLE 6.7 – Parameters of the phytoplankton growth model

Troisième partie

Applications

Chapitre 7

Applications à des modèles analytiques

Ce chapitre propose une étude synthétique des différents estimateurs des indices de Sobol' que nous avons évoqués précédemment. Les applications sont effectuées sur 3 modèles analytiques : une fonction continue et multiplicative, une fonction continue mais non-multiplicative et enfin une fonction discontinue. Ces cas d'études sont très élémentaires et n'ont pour but que de mettre en évidence les faiblesses éventuelles des différentes méthodes testées.

Nous nous intéressons à l'estimation des indices de Sobol' d'ordre 1 et 2. Les méthode comparées sont : la méthode de Sobol', RBD, FAST et l'approche par quasi-régression suivant des polynômes de chaos (PC). La méthode de Sobol' est évaluée dans sa version classique (notée MC), ainsi qu'avec la nouvelle approche introduite au Chapitre 6 (notée RLHS pour l'estimation des indices d'ordre 1 et RLHS-OA2 pour les indices d'ordre 2). La méthode RBD est testée dans sa version classique (notée OA1), dans sa version classique avec correction de biais (notée OA1/corrigée), dans la version construite à l'aide de tableaux orthogonaux de force 2 introduite au Chapitre 5 (notée OA2 et OA2/corrigée). La méthode FAST est appliquée dans sa nouvelle version introduite au Chapitre 5 à l'aide de jeux de fréquences — disponibles sur la page internet personnelle de Frances Kuo : http ://web.maths.unsw.edu.au/∼fkuo/ — construits par les algorithmes "composante par composante" récents (voir [CKN06, Nuy07, CKN10]). La randomisation du plan d'expérience de FAST se fait par la rotation de Cranley-Patterson qui consiste à translater, modulo 1 et composante par composante, tous les points du plan original par un même vecteur aléatoire uniformément distribué sur $[0,1]^d$. Enfin, l'approche par quasi-régression suivant des polynômes de chaos est effectuée dans un premier temps relativement à des hypercubes latins, puis généralisée à des hypercubes latins basés sur des tableaux orthogonaux de force 2 (notée OA2).

Pour l'estimation des indices d'ordre 1 par les méthodes spectrales, l'ensemble de troncature choisi est du type $\{1,\dots,M\}$ pour les polynômes de chaos et $\{-M,\dots,-1,1,\dots M\}$ pour les polynômes trigonométriques où M est un entier naturel non nul. Pour l'estimation des indices d'ordre 2, l'ensemble de troncature est une croix de Zaremba du type $\{i*j \mid i,j \in \mathbb{N}^* \text{ et } i*j \leq M\}$ pour les polynômes de chaos et $\{i*j \mid i,j \in \mathbb{Z}^* \text{ et } |i*j| \leq M\}$ pour les polynômes trigonométriques où M est également un entier naturel.

Le critère de comparaison utilisé dans ces tests est la racine carrée de l'erreur quadratique moyenne (RMSE) calculée sur 1000 réplications. Il est important de noter que lors de l'estimation des indices de Sobol' d'ordre 2, les quantités évaluées par la méthode de Sobol' sont des indices descendants et pour les méthodes spectrales, des indices élémentaires.

7.1 Test sur la g-fonction

7.1.1 Descriptif du test

La première fonction testée est la g-fonction (voir e.g. [SS95], et Figure 7.1) :

$$f(\mathbf{X}) = \prod_{i=1}^{d} f_i(X_i)$$

avec pour tout $i \in [1:d]$

$$f_i(X_i) = \frac{|4X_i - 2| + a_i}{1 + a_i} \quad , \quad a_i > 0$$

où les X_i désignent des variables aléatoires indépendantes uniformément distribuées sur $[0,1]$.

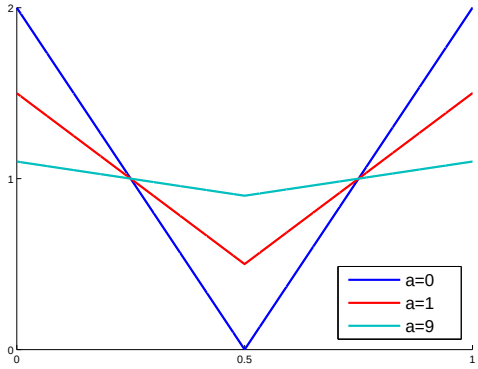


FIGURE 7.1 – Courbe des composantes d'une g-fonction pour quelques valeurs de paramètres a.

Il est important de remarquer que, même si la définition de cette fonction laisse entrevoir une singularité au centre de son ensemble de définition, la g-fonction possède une représentation en série de Fourier relativement simple puisque son spectre décroît à l'infini de manière très rapide. En effet, on note que pour tout $\mathbf{k} \in \mathbb{Z}^d$

$$\mathrm{c}_{\mathbf{k}}(f) = \begin{cases} 0 & \text{si } \exists i \mid k_i \neq 0 \text{ et pair} \\ \dfrac{\displaystyle\prod_{i \mid k_i \neq 0} 4\pi^{-2}(1+a_i)^{-1}}{\displaystyle\prod_{i \mid k_i \neq 0} k_i^2} & \text{sinon.} \end{cases}$$

Le test est effectué sur une g-fonction de 12 variables dont les paramètres a_i sont donnés par $\mathbf{a} = (0,0,0,1,1,1,1,9,9,9,9,9)$. Les valeurs théoriques des indices de Sobol' de cette fonction peuvent être obtenues en notant que

$$\begin{aligned} \sigma^2_{\{i\}} &= \frac{1}{3(1+a_i)^2}, \\ \sigma^2_{\mathrm{u}} &= \prod_{i \in \mathrm{u}} \sigma^2_{\{i\}} \end{aligned}$$

et

$$\sigma^2 = \prod_{i=1}^{n} (\sigma^2_{\{i\}} + 1) - 1.$$

Les estimations sont réalisées pour plusieurs tailles d'échantillons : 1000, 4000, 10000, 40000 et 160000 ; et pour chacune, 1000 réplications sont effectuées. Le nombre d'harmoniques prises en compte dans le calcul des indices de Sobol' d'ordre 1 pour les méthodes spectrales est fixé à 8, 13, 17, 24

et 30 pour chacune des tailles d'échantillons. Ces valeurs représentent également le paramètre des croix de Zaremba (voir Formule (5.41)) considérées lors de l'estimation des indices d'ordre 2 pour les méthodes spectrales.

7.1.2 Résultats du test et discussions

Les estimations des différents indices de Sobol' d'ordre 1 et d'ordre 2 présentant toutes le même comportement, nous nous contentons de tracer uniquement les courbes pour deux indices d'ordre 1 — l'un dont la valeur est la plus grande (0,145) et l'autre dont la valeur est la plus petite (0,001) — et pour un indice d'ordre 2 — dont la valeur est la plus grande (0.048). Les résultats sont présentés dans les Figures 7.2 à 7.4.

Concernant l'estimation des indices d'ordre 1, on remarque principalement que deux groupes se dessinent. D'une part, on constate que la méthode FAST atteint une vitesse de convergence en $O(n^{-1})$, et d'autre part que les autres méthodes convergent à une vitesse de l'ordre de $O(n^{-1/2})$. La vitesse de convergence observée de la méthode FAST est une illustration de la Proposition 5.4 qui établit que les estimateurs de $\sigma^2_{\{i\}}$ et σ^2 produits par cette méthode peuvent converger à des vitesses supérieures à celle de la méthode de Monte Carlo. Parmi les autres méthodes, on remarque que les méthodes de Sobol' sont les moins précises, et que la méthode de quasi-régression suivant les polynômes de Legendre est comparable à la méthode RBD classique. Enfin, on note que la correction de biais et la stratification plus forte — par un tableau orthogonal de force 2 — tend à réduire l'erreur quadratique.

Concernant les indices de Sobol' d'ordre 2, les remarques précédentes s'appliquent dans leur ensemble. On remarque en outre que la méthode de Sobol' permet d'atteindre un niveau d'erreur comparable à celui de la méthode RBD,.

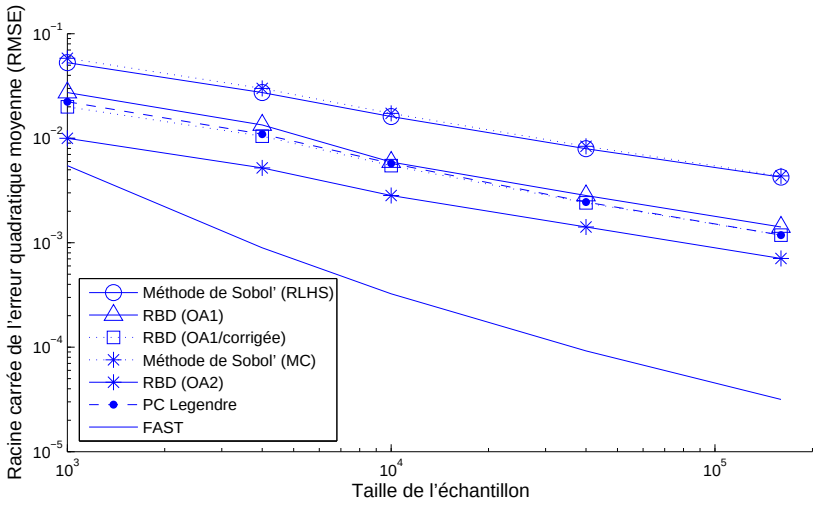


FIGURE 7.2 – Estimation des indices de Sobol' d'ordre 1 de la g-fonction ayant pour valeur 0.145.

Remarque 7.1. *Dans ce test, nous avons "normalisé" la méthode de Sobol' basée sur le hypercubes latins (notée RLHS), i.e. que pour un test sur un échantillon de taille n, nous avons considéré une méthode de Sobol' sur deux hypercubes latins répliqués de taille $n/2$. Le comparatif est donc parfaitement juste vis-à-vis des méthodes spectrales. Par contre, la méthode de Sobol' classique n'a pas été "normalisée" de manière à comparer la qualité de son estimateur relativement à celui de la méthode de Sobol' RLHS. On remarquera toutefois que sa RMSE doit, en toute rigueur, être multipliée par un facteur non-négligeable — ici $\sqrt{6.5} > 2,5$ puisqu'elle nécessite au minimum 13 échantillons pour estimer tous les indices d'ordre 1 et 26 pour l'estimation des indices d'ordre 2, au lieu de 2 seulement pour la méthode RLHS — qui se dégrade avec l'augmentation de la dimension. Cette remarque s'appliquera également pour les tests suivants.*

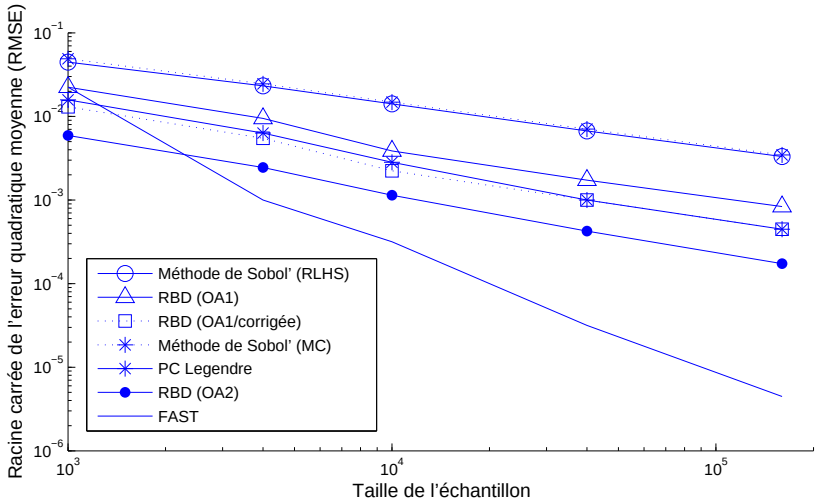


FIGURE 7.3 – Estimation des indices de Sobol' d'ordre 1 de la g-fonction ayant pour valeur 0.001.

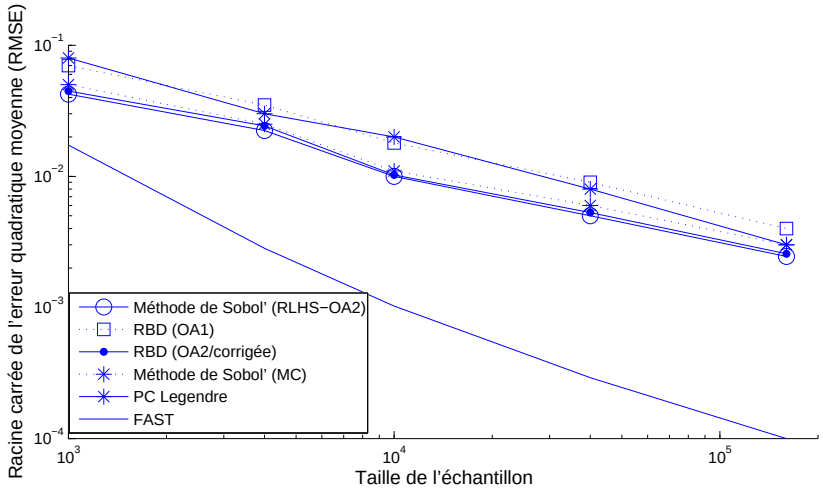


FIGURE 7.4 – Estimation des indices de Sobol' d'ordre 2 de la g-fonction

7.2 Test sur une fonction continue non-multiplicative

7.2.1 Descriptif du test

La seconde fonction testée est une fonction continue mais non multiplicative en dimension 10 définie par

$$f(\mathbf{X}) = \min(X_1, \ldots, X_5)$$

avec $\mathbf{X} = (X_1, \ldots X_{10})$, où les X_i désignent des variables aléatoires indépendantes uniformément distribuées sur $[0, 1]$. On peut la trouver par exemple dans [LO06].

La fonction présente donc uniquement $s = 5$ variables actives jouant des rôles symétriques. Les valeurs théoriques de ses indices de Sobol' sont accessibles par les formules

$$\underline{\tau}^2_{\mathrm{u}} = \frac{|\mathrm{u}|}{(s+1)^2(2s - |\mathrm{u}| + 2)} \quad \mathrm{u} \subseteq \{1, \ldots, 5\}$$

et

$$\sigma^2 = \frac{s}{(s+1)^2(s+2)}.$$

Dans notre cas, les indices de Sobol' d'ordre 1 des variables actives prennent pour valeur 7/55, et 7/25 pour les indices de Sobol' d'ordre 2 descendants.

Les estimations sont réalisées pour plusieurs tailles d'échantillons : 2209, 10201, 41209, 109561, 491401 et 1002001 ; et pour chacune, 1000 réplications sont effectuées. Le nombre d'harmoniques prises en compte dans le calcul des indices de Sobol' d'ordre 1, ainsi que le paramètre des croix de Zaremba (voir Formule (5.41)) considérées lors de l'estimation des indices d'ordre 2 pour les méthodes spectrales, est maintenant choisi de manière à minimiser l'erreur quadratique[1] Dans ce cadre, les indices de Sobol' dont les valeurs sont nulles ne sont pas estimés par les méthodes spectrales, car le paramètre optimal pour le nombre d'harmoniques à prendre en compte est 0. Ils sont néanmoins estimés par les méthodes de Sobol', et présentent dans tous les cas une RMSE 6 fois inférieure à celle des indices non nuls présenté dans les Figures 7.5 et 7.6.

7.2.2 Résultats du test et discussions

Pour l'estimation des indices d'ordre 1, même si contrairement au test précédent, toutes les méthodes présentent la même vitesse de convergence, les résultats des Figures 7.5 et 7.6 montrent quand même une grande disparité entre les différentes méthodes. On constate d'abord un rapport égal à 20 entre la RMSE de la méthode de Sobol' et celle de la quasi-régression suivant les polynômes de Legendre. Ceci est principalement expliqué par une performance exceptionnelle de la quasi-régression qui tient au fait que seules les 5 premiers coefficients du développement en chaos polynomial de Legendre sont prises en compte dans le calcul des indices de Sobol'. Seule la méthode FAST est capable d'obtenir une RMSE aussi faible que celle de la quasi-régression sur les polynômes de Legendre. Toutefois, le nombre d'harmoniques nécessaires pour atteindre ce niveau de précision varie entre 15, pour l'échantillon de taille 2000, et 60 pour l'échantillon de taille 1000000. Ce nombre important est désavantageux pour FAST car il fait apparaître des interférences lors de l'estimation des coefficients de Fourier ; ils se traduisent d'ailleurs par la trajectoire non lisse de la RMSE de FAST dans la Figure 7.5. Quant aux méthodes RBD, elles sont nettement moins concurrentielles que la quasi-régression sur les polynômes de Legendre, du fait principalement d'un nombre imposant d'harmoniques nécessaires à l'estimation des indices de Sobol' : 9 pour l'échantillon de taille 2000, et 70 pour l'échantillon de taille 1000000. En effet, plus le nombre d'harmoniques prises en compte dans l'estimation de l'indice est important, et plus la variance et le biais résiduel après correction augmentent.

Pour l'estimation des indices d'ordre 2, les constats et les explications sont les mêmes. La quasi-régression suivant les polynôme de Legendre reste la méthode la plus compétitive ; et les méthodes spectrales RBD et FAST doivent composer avec des paramètres de croix de Zaremba très élevés — jusqu'à 60 pour l'échantillon de taille 1000000 — qui contraignent FAST à une vitesse de décroissance en $O(n^{-1/2})$ et qui empêche RBD d'atteindre une RMSE inférieure à 10^{-3}. Ce dernier point illustre la problématique principale de l'estimation des indices de Sobol' par méthode spectrale : si l'indice en question est la série de coefficients de Fourier — ou Fourier-Hermite, Fourier-Legendre etc. — qui décroissent lentement à l'infini, alors l'estimateur de l'indice de Sobol' doit prendre en compte un nombre important de ces coefficients pour atteindre un niveau de précision donné, et malheureusement, plus ce nombre augmente et plus la variance de l'estimateur aussi. Ceci peut conduire dans des cas pathologiques à un estimateur qui ne converge pas, c'est ce qu'on observe pour la méthode RBD sur la Figure 7.6, car vraisemblablement la base trigonométrique ne permet pas d'approcher l'effet d'ordre 2 de la fonction étudiée par des harmoniques basse-fréquences. Enfin, on termine en notant que le remplacement de l'hypercube latin par un hypercube latin basé sur un tableau orthogonal de force 2 dans la quasi-régression suivant les polynômes de Legendre permet de réduire l'erreur quadratique de manière importante.

1. Cela signifie qu'on réalise le test pour chacune des valeurs jusqu'à observé une dégradation de l'estimation du fait d'un trop grand nombre de coefficients pris en compte, et on renvoie la meilleure estimation.

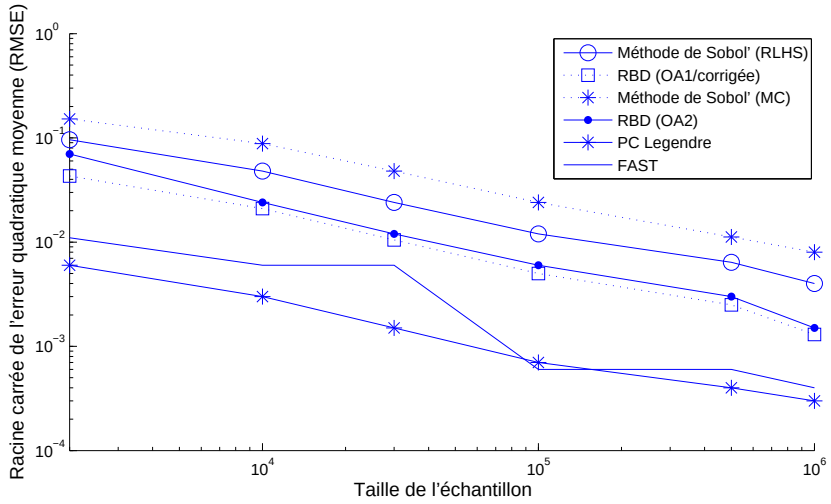


FIGURE 7.5 – Estimation des indices de Sobol' d'ordre 1 de la fonction-test numéro 2 ayant pour valeur 7/55.

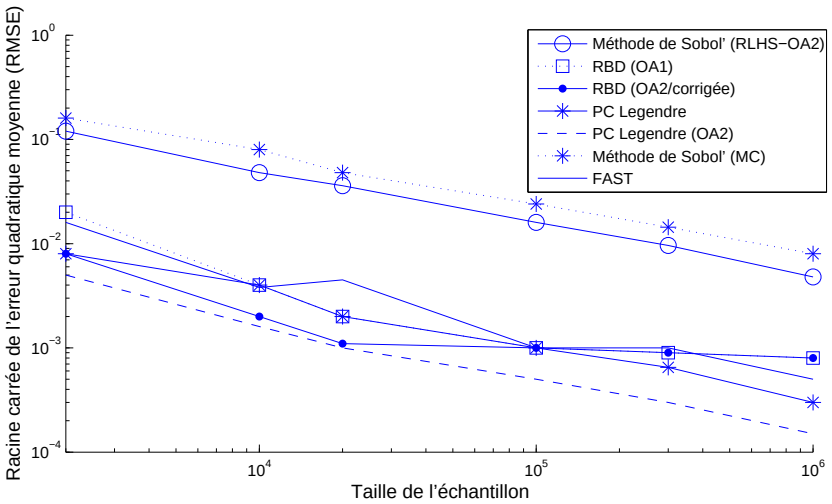


FIGURE 7.6 – Estimation des indices de Sobol' d'ordre 2 de la fonction-test numéro 2 ayant pour valeur 7/25 (indice descendant) ou 7/275 (indice élémentaire).

7.3 Test sur une fonction discontinue

7.3.1 Descriptif du test

La dernière fonction testée est une fonction discontinue en dimension 10 définie par

$$f(\mathbf{X}) = \mathbf{1}_{\{X_1<X_2\}}(\mathbf{X})$$

avec $\mathbf{X} = (X_1, \dots X_{10})$, où les X_i désignent des variables aléatoires indépendantes uniformément distribuées sur $[0, 1]$.

La fonction présente donc uniquement 2 variables actives jouant des rôles symétriques. Les indices de Sobol' des variables actives, $S_{\{1\}}$, $S_{\{2\}}$et $S_{\{1,2\}}$, prennent pour valeur 1/3.

Les estimations sont réalisées pour plusieurs tailles d'échantillons : 2000, 10000, 20000, 100000, 200000 et 1000000 ; et pour chacune, 1000 réplications sont effectuées. Le nombre d'harmoniques prises en compte dans le calcul des indices de Sobol' d'ordre 1, ainsi que le paramètre des croix de Zaremba (voir Formule (5.41)) considérées lors de l'estimation des indices d'ordre 2 pour les méthodes

spectrales, est maintenant choisi de manière à minimiser l'erreur quadratique. Dans ce cadre, les indices de Sobol' dont les valeurs sont nulles ne sont pas estimés par les méthodes spectrales, car le paramètre optimal pour le nombre d'harmoniques à prendre en compte est 0. Ils sont néanmoins estimés par les méthodes de Sobol', et présentent dans tous les cas une RMSE 3 fois inférieure à celle des indices non nuls présenté dans les Figures 7.7 et 7.8.

7.3.2 Résultats du test et discussions

Concernant l'estimation d'ordre 1, une fois encore la méthode FAST est incapable d'atteindre une vitesse de convergence supérieure à $O(n^{-1/2})$. De plus, pour atteindre cette vitesse, l'estimateur FAST a dû prendre en compte un nombre gigantesque d'harmoniques : 25 pour l'échantillon de taille 2000, et jusqu'à 300 pour l'échantillon de taille 1000000. Cette précision obtenue à l'aide d'un tel nombre d'harmoniques n'est possible que parce que le spectre global de la fonction-test est très creux — car seule 2 variables sont actives —; dans un modèle présentant la même discontinuité et plus de variables actives, une telle précision serait tout simplement impossible du fait des interférences. En terme de performance, arrivent ensuite la méthode de Sobol' nouvellement introduite au Chapitre 6 et la quasi-régression suivant les polynômes de Legendre. Pour cette dernière, il suffit d'une seule harmonique pour converger vers l'indice de Sobol' d'ordre 1, ce qui explique son bon comportement. Rajoutons qu'en utilisant un hypercube latin basé sur un tableau orthogonal d'ordre 2 pour cette quasi-régression, on réduit la racine carrée de l'erreur quadratique d'un facteur 3 à 4. Pour conclure sur ces premières remarques, on note simplement que la présence de discontinuités dans un modèle — même la plus simple — fait rapidement apparaître la méthode de Sobol' comme une méthode performante. On constate en outre que l'écart entre la méthode de Sobol' originale et la méthode exploitant les hypercubes latins est ici plus important que dans les tests précédents. Enfin, on note que comme dans le test précédent, la méthode RBD n'arrive pas à converger du fait d'un trop grand nombre d'harmoniques à prendre en compte.

Concernant l'estimation des indices d'ordre 2, seules les méthodes de Sobol' et la méthode FAST sont représentées. Comme précédemment pour l'estimation des indices d'ordre 1, la méthode FAST ne doit son semblant de convergence qu'à la prise en compte d'un nombre gigantesque d'harmoniques — par exemple, une croix de Zaremba contenant plus de 4000 coefficients pour l'échantillon de taille 1000000. Pour la méthode RBD, la convergence difficile observée pour l'estimation des indices d'ordre 1 se reproduit dans une plus grande ampleur, et la RMSE reste largement supérieure à 10^{-2}. Le même phénomène s'observe pour la quasi-régression suivant les polynômes de Legendre. Et finalement, seule la méthode de Sobol' est robuste à la discontinuité de la fonction-test.

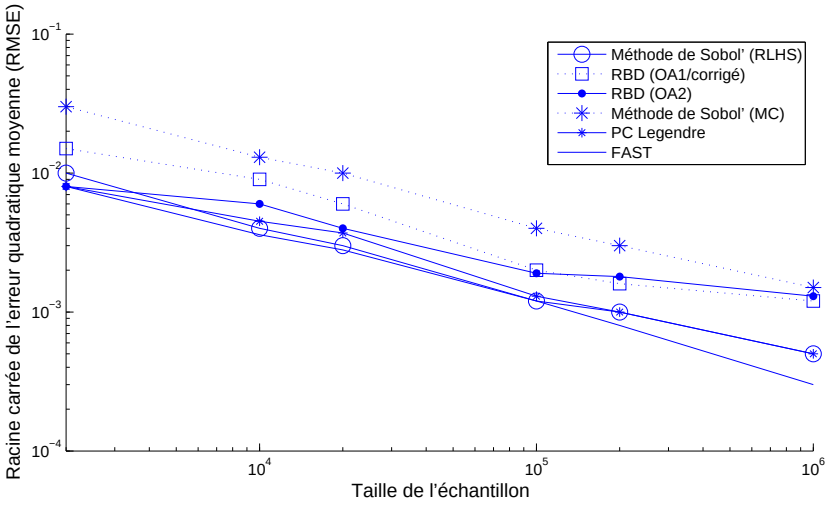


FIGURE 7.7 – Estimation des indices de Sobol' d'ordre 1 de la fonction-test numéro 3 ayant pour valeur 1/3.

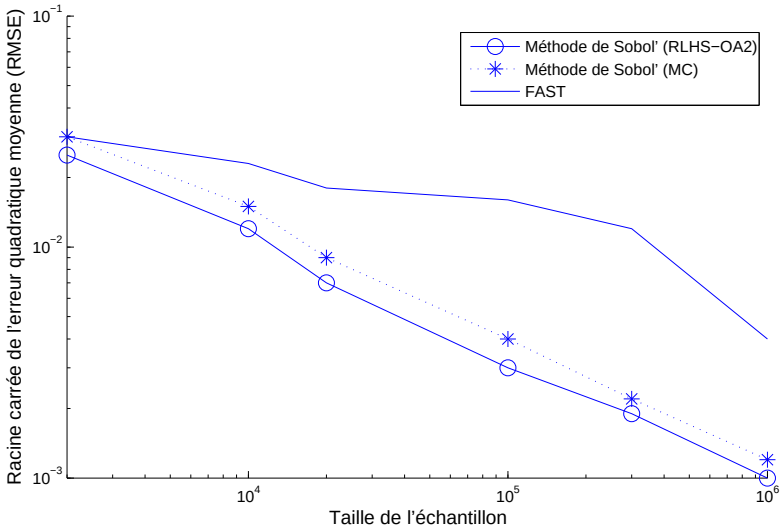


FIGURE 7.8 – Estimation des indices de Sobol' d'ordre 2 de la fonction-test numéro 3 ayant pour valeur 1/3 (indice élémentaire) et 1 (indice descendant).

7.4 Conclusion

Après tout ces tests, on constate sans surprise que la méthode de Sobol' est la seule à être robuste vis-à-vis d'une discontinuité. Plus encore, elle montre des erreurs quadratiques très faibles pour l'estimation des indices d'ordre 2. En effet pour ce type d'indices, même si certains tracés de RMSE peuvent faire apparaître cette méthode en retrait des autres, on constate en réalité que sa RMSE renormalisée par la valeur de l'indice estimé plaide plutôt en sa faveur. Rappelons par exemple, pour le test numéro 2, les méthodes spectrales estiment l'indice élémentaire de valeur 7/275 et la méthode de Sobol' l'indice descendant de valeur 7/25.

En second lieu, on remarque que les erreurs quadratiques observées pour les méthodes de quasi-régression suivant les polynômes de Legendre ou suivant les polynômes trigonométriques (RBD) peuvent varier énormément suivant le nombre de coefficients pris en compte dans le calcul de l'indice de Sobol'. En particulier, un développement spectral à décroissance lente conduit à des estimateurs qui n'arrivent pas à converger. Le problème central dans ce type de méthode est donc de trouver la base orthogonale sur laquelle le modèle possède un spectre à décroissance rapide. Nous avons pu constater que les polynômes de Legendre et les polynômes trigonométriques échouent tous les deux dans le cas d'une fonction présentant un front discontinu. Il existe bien évidemment d'autres bases orthogonales, citons par exemple, les fonctions de Walsh [Bea75] — qui ont fait l'objet d'une généralisation de la méthode FAST [PC81] — ou les ondelettes [Mey93].

Enfin, concernant la méthode FAST, nous avons pu observer qu'elle peut atteindre une vitesse de convergence de l'ordre de $O(n^{-1})$ pour une fonction dont le spectre de Fourier est à décroissance rapide, mais qu'elle se heurte au problème des interférences dès que cette propriété n'est pas vérifiée. En dimension réduite, nous avons pu constater qu'elle peut rester compétitive si la fonction ne présente pas de discontinuité, mais au regard de l'étude théorique du Chapitre 5, il n'apparaît pas raisonnable de l'appliquer en grande dimension ($d > 20$).

Chapitre 8

Application à un modèle d'écosystème marin

Ce chapitre constitue une application des méthodes d'analyse de sensibilité à un simulateur d'écosystème marin dépendant de 85 paramètres. L'étude que nous proposons constitue un travail introductif qui nous permet de tester les méthodes discutées dans cette thèse sur un modèle non-analytique qui s'exprime de manière implicite. D'autre part, cette première étude avec un nombre réduit de paramètres incertains, nous permet de prendre la mesure de la complexité du modèle et ainsi d'évaluer quels types de méthodes sont pertinents et réalisables en vue d'expériences à plus grande échelle. À moyen terme, ce travail a idéalement pour but l'aide au développement de ce modèle (réduction/complexification, calibration, etc.), en complétant les analyses de sensibilité déjà existantes, comme par exemple l'analyse de colinéarité [BRK01] (voir, e.g., [RSG06]) ou des approches élémentaires de calcul de taux d'accroissement (voir, e.g., [Lac98]).

8.1 Description du modèle MODECOGeL

MODECOGeL est un modèle 1D couplé (hydrodynamique/biogéochimie) simulant l'écosystème marin régional en mer Ligure (Méditerranée nord-ouest, voir Figure 8.1) suivant la dimension verticale — jusqu'à $-400m$ — en vue d'une étude de variabilité interannuelle entre les années 1984 et 1988. Il a été développé par Lacroix [Lac98, LN98] et ensuite affiné par Lacroix et Grégoire [LG02]. Il consiste en un code FORTRAN de trois milles lignes résultant de la discrétisation par la méthode des volumes finis d'un certain nombre d'équations différentielles couplées contrôlées par 85 paramètres. Les pas de temps et d'espace sont respectivement 10 minutes ou 1 heure, et 1 mètre ou 5 mètres. La résolution la plus grossière permet de réaliser une simulation annuelle en 4 secondes avec un ordinateur de bureau (processeur quatre coeurs cadencé à 3GHz, 4Go de mémoire vive). Dans ce qui suit, en nous référant exclusivement à la thèse de Lacroix, nous résumons les enjeux et la desription de ce modèle. Pour plus de détails, nous renvoyons le lecteur au manuscrit original de Lacroix [Lac98].

8.1.1 Enjeux de la modélisation des processus biologiques de l'océan

L'étude des océans, et plus généralement des eaux naturelles, intéresse les chercheurs autant dans l'optique de considérations climatiques, qu'écologiques. Dans l'introduction générale de sa thèse consacrée à la construction du modèle MODECOGeL, Lacroix [Lac98] pointe trois phénomènes dont l'intérêt ne cesse de grandir auprès de la communauté scientifique. Il s'agit d'une part de l'évolution de la concentration en dioxyde de carbone (CO_2) dans l'océan, et plus généralement du cycle global du CO_2 (océan/atmosphère) d'autre part, de l'évolution des ressources halieutiques, et enfin du problème lié à l'eutrophisation — i.e. surplus de phosphore et d'azote dans les eaux naturelles qui ont principalement pour conséquence une dégradation du milieu marin.

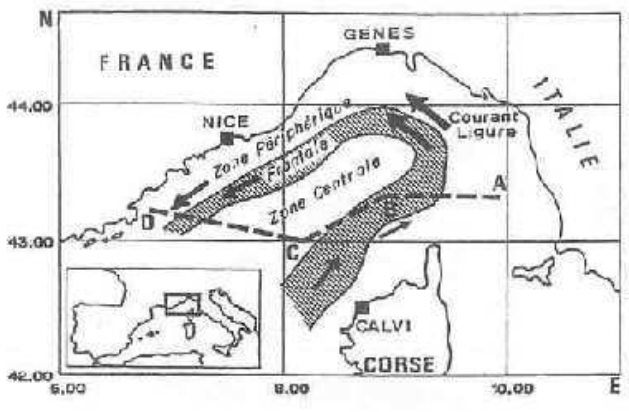


FIGURE 8.1 – Schéma de la Mer Ligure.

Évolution de la concentration en CO_2 Le processus qui régule la température au sol de la Terre — 15°C en moyenne — se résume principalement à un phénomène naturel communément appelé *effet de serre*. En effet, certains gaz présents dans l'atmosphère terrestre comme la vapeur d'eau, le CO_2, le méthane ou encore l'ozone piègent une partie du rayonnement infrarouge émis par la Terre et le réémettent vers cette dernière, ce qui a pour conséquence d'augmenter la température au sol. Sans ce processus naturel, la température moyenne terrestre au sol serait seulement de −18°C [Bar07]. Selon les données du rapport de l'année 2011 du Groupe d'Experts Intergouvernemental sur l'Évolution du Climat (GIEC), la contribution des différents gaz à effet de serre se répartit comme suit

- vapeur d'eau, 60%
- dioxyde de carbone, 26%
- ozone, 8%
- méthane et oxyde nitreux, 6%.

Par conséquent, le cycle du CO_2, et en particulier l'accroissement de sa concentration atmosphérique dû à l'activité humaine, apparaît comme un objet d'étude important. Notons en outre que près d'un tiers des rejets anthropiques en CO_2 est absorbé par l'océan via la photosynthèse effectuée par les organismes autotrophes [SS93] et peut être piégé au fond de l'océan plusieurs centaines d'années [LM98]. La quantification des flux de carbone océanique revêt donc une importance particulière pour réaliser des prévisions de la concentration en CO_2 atmosphérique à l'échelle climatique.

Évolution des ressources halieutiques L'estimation des stocks de poissons est une problématique centrale dans le développement durable des ressources halieutiques. Elle permet en effet de mettre en place une régulation de la pêche et ainsi d'éviter l'épuisement des stocks de poissons. Cette estimation est néanmoins extrêmement difficile en pratique, et une technique alternative à l'évaluation directe des stocks consiste à évaluer la quantité de CO_2 exportée dans les réseaux trophiques supérieurs. Cette connaissance permet en effet d'estimer approximativement les stocks de poissons [PTH84].

Eutrophisation L'accroissement des concentrations en nitrate et phosphate dans les terres — notamment par l'utilisation intensive d'engrais dans l'agriculture moderne — produit un surplus d'azote et de phosphore dans les eaux naturelles. Ceci a principalement pour conséquence une croissance excessive et une prolifération des végétaux aquatiques qui peut mener à des extinctions d'espèces et à des problèmes de qualité de l'eau [Sch96]. Par conséquent, l'étude du cycle de l'azote et du phosphore est tout aussi important que celui du dioxyde de carbone.

Même si les observations directes in situ — i.e. par analyse de prélèvements au niveau d'une balise permanente ou par campagne en mer (température, salinité, etc.) — et les observations par satellite

— i.e. par analyse de la couleur de l'eau pour évaluer de manière indirecte la concentration en phytoplancton — se font de plus en plus nombreuses sur un maillage de plus en plus fin, elles restent néanmoins insuffisantes pour comprendre certains phénomènes et les prévoir. Par conséquent, la modélisation des processus biologiques peut apporter une aide considérable en océanographie. Néanmoins un simple modèle biologique peut se révèler insuffisant pour reproduire les phénomènes naturels originaux en adéquation avec les données observées. En effet, les phénomènes hydrodynamiques ont un impact majeur sur les flux de nutriments, et les conditions de température et d'insolation contrôlent en partie les processus de croissance des autotrophes. Notons en outre que comme le remarque Lacroix [Lac98], les différents processus hydrodynamiques qui se déroulent à des échelles de temps proches de ceux des processus biologiques vont avoir tendance à imposer leur structure spatiale [Nih81] — on parle d'*ajustement écohydrodynamique* [ND90]. Par conséquent, le couplage du modèle biologique à un modèle physique apparaît comme incontournable. C'est ce qui est réalisé dans le modèle MODECOGeL que nous détaillons maintenant.

8.1.2 Description du modèle physique

Modélisation de la quantité de lumière disponible pour la photosynthèse

Le processus de photosynthèse réalisé par les organismes autotrophes dépend essentiellement de la quantité de lumière reçue par ceux-ci. Or la lumière arrivant à la surface de l'océan est partiellement réfléchie, suivant l'état de la surface de l'eau, et la quantité de lumière décroît en outre avec la profondeur du fait de la capacité d'absorption de l'eau et de l'auto-ombrage du phytoplancton [PTH84]. On considère généralement que 43% à 50% de la quantité de lumière est active pour la photosynthèse une fois la surface de l'eau traversée, et que cette quantité décroît exponentiellement avec la profondeur [AN88]. On aboutit alors au modèle suivant (voir e.g. [Lac98]).

$$PAR(z,t) = PAR_0 * IS(t) * (1-\alpha) * \exp(-K_{ext} * z)$$

où le coefficient d'extinction de la lumière avec la profondeur (m^{-1}), K_{ext}, est donné par

$$K_{ext} = K_w + K_{chl1} \times \texttt{chla} + K_{chl2} \times \texttt{chla}^{2/3}$$

avec

$PAR(z,t)$	: quantité de lumière disponible pour la photosynthèse à la profondeur z et au temps t (Wm^2)
PAR_0	: proportion de radiations disponibles pour la photosynthèse (sans unité, dans $[0,1]$)
$IS(t)$	: insolation totale arrivant en surface après réflexion (Wm^{-2})
α	: albédo de surface (sans unité, dans $[0,1]$)
K_w	: coefficient d'extinction de la lumière avec la profondeur pour l'eau pure (m^{-1})
K_{chl1} et K_{chl2}	: coefficients d'extinction de la lumière due à la biomasse phytoplanctonique ($(mgChl)^{-1}m^2$) et ($(mgChl)^{-2/3}m$)
$\texttt{chla}(z,t)$	: concentration en chlorophylle-a ($mgChlm^3$).

Modélisation hydrodynamique

Le modèle hydrodynamique est construit à partir des *équations primitives* décrivant la dynamique de l'océan, et ramené à la dimension verticale en supposant l'homogénéité horizontale [Lac90, LD92]. Il possède cinq variables d'état : la température T, la salinité S, les vitesses horizontales u et v, et l'énergie cinétique turbulente k. Nous résumons les équations, les conditions initiales et les conditions aux limites qui le régissent dans les deux paragraphes suivants (voir [Lac98], Annexe 2, pour plus de détails). Notons qu'une version améliorée de ce modèle, à six variables d'état, a été proposée par Lacroix et Grégoire en 2002 [LG02].

Équations hydrodynamiques Les équations sont les suivantes :

$$\begin{aligned}
\frac{\partial u}{\partial t} &= \frac{\partial}{\partial z}\left(\tilde{\lambda}\frac{\partial u}{\partial z}\right) + fv \\
\frac{\partial v}{\partial t} &= \frac{\partial}{\partial z}\left(\tilde{\lambda}\frac{\partial v}{\partial z}\right) + fu \\
\frac{\partial T}{\partial t} &= \frac{\partial}{\partial z}\left(\tilde{\lambda}^T\frac{\partial T}{\partial z}\right) \\
\frac{\partial S}{\partial t} &= \frac{\partial}{\partial z}\left(\tilde{\lambda}^S\frac{\partial S}{\partial z}\right) \\
\frac{\partial k}{\partial t} &= \tilde{\lambda}\left|\left|\frac{\partial(u,v)}{\partial z}\right|\right|^2(1-R_f) - \varepsilon + \frac{\partial}{\partial z}\left(\tilde{\lambda}\frac{\partial k}{\partial z}\right)
\end{aligned}$$

avec

f : facteur de Coriolis (s^{-1})
$\tilde{\lambda}$: coefficient de diffusion turbulente associée à la vitesse (m^2s^{-1})
$\tilde{\lambda}^T$: coefficient de diffusion turbulente associée à la température (m^2s^{-1})
$\tilde{\lambda}^S$: coefficient de diffusion turbulente associée à la salinité (m^2s^{-1})
R_f : facteur de Richardson de flux (sans unité)
ε : taux de dissipation de l'énergie cinétique turbulente (m^2s^{-3}).

Les coefficients $\tilde{\lambda}$, $\tilde{\lambda}^S$, $\tilde{\lambda}^T$ et ε sont donnés par

$$\tilde{\lambda} = 0.5 * l * \sqrt{k}$$

où

$$l = \kappa * z * \left(1 - 0.75\frac{z}{H}\right) * (1 - R_f)$$

avec

κ : constante de Von Karman ($= 0.4$)
z : hauteur d'eau à partir du fond (m)
H : hauteur d'eau totale (m) ;

et pour $y = T$ ou S,

$$\tilde{\lambda}^y = \tilde{\lambda} * 1.1 * \sqrt{1 - R_f};$$

et enfin,

$$\varepsilon = \frac{k^2}{16\tilde{\lambda}}.$$

Conditions initiales et conditions aux limites La vitesse et l'énergie cinétique turbulente sont supposées nulles au temps $t = 0$ par défaut d'observations disponibles. Les conditions initiales de cholorophylle-a, de température et de salinité jusqu'à $200m$ sont quant à elles issues des données FRONTAL — programme d'étude des zones frontales — de l'année 1985 à l'année 1988 incluses. Pour les valeurs entre $200m$ et $400m$, il s'agit d'une interpolation linéaire entre la valeur à $200m$ issue des données FRONTAL et de la valeur à $600m$ estimée d'après [JT86] et [Min69]. Les conditions aux limites sont reportées dans l'Annexe B.

8.1.3 Description du modèle biologique

Le modèle biologique est du type NPZD — Nutriments, Phytoplanctons, Zooplanctons et Détritus — et possède douze variables d'état répertoriées dans la Table 8.1.

L'unité utilisée pour les quantifier est la concentration en azote ($mmolNm^{-3}$), qui est le principal élément limitant pour le développement phytoplanctonique. Les équations différentielles qui servent à modéliser le modèle biologique sont reportées dans la Table 8.3. Auparavant, nous donnons un rapide descriptif des douzes variables d'état et de leurs interactions (voir également la Figure 8.2).

classe	variable d'état	caractéristique	notation
nutriments	ammonium		nh4
	nitrate		no3
phytoplanctons	picophytoplancton	taille : $[0.2\mu m, 2\mu m]$	pp
	nanophytoplancton	taille : $[2\mu m, 20\mu m]$	np
	microphytoplancton	taille : $[20\mu m, 200\mu m]$	mp
zooplanctons	nanozooplancton	taille : $[2\mu m, 20\mu m]$	nz
	microzooplancton	taille : $[20\mu m, 200\mu m]$	miz
	mésozooplancton	taille : $[200\mu m, 2mm]$	mez
bactéries	bactéries		bac
matières organiques particulaires (m.o.p.) (i.e. détritus)	m.o.p. de classe 1	issue des pico-, nano-, microplancton et bactéries	mop1
	m.o.p. de classe 2	issue du mésozooplancton	mop2
azote organique dissous	azote organique dissous		nod

TABLE 8.1 – Variables d'état du modèle biologique.

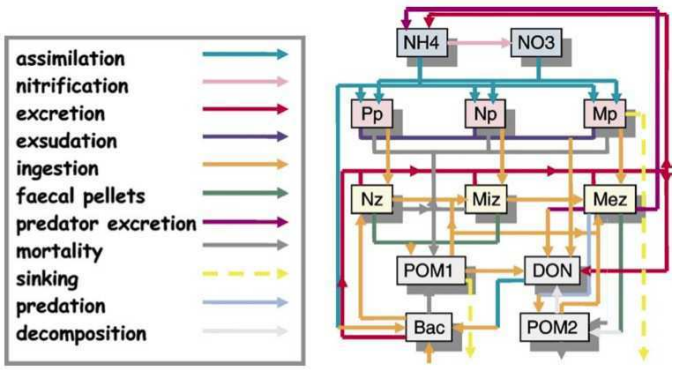


FIGURE 8.2 – Modèle biogéochimique. NH4 : Ammonium ; NH3 : nitrate ; Nz, Miz, Mez : nano-, micro-, mésozooplancton ; Pp, Np, Mp : pico-, nano-, microphytoplancton ; PON1, PON2 : matière organique particulaire de type 1 et 2 ; Bac : bactéries ; DON : azote organique dissous.

Nutriments Les deux types d'azote inorganique dissous considérés, l'ammonium et le nitrate, permettent la croissance du phytoplancton. Le nitrate provient de la diffusion hivernale, et de la nitrification de l'ammonium. Ce dernier provient, quant à lui, de l'excrétion des différents zooplanctons et des bactéries.

Phytoplanctons Le taux de croissance des différents phytoplanctons est défini en prenant en compte les effets limitants dus à l'insolation, la température, à la concentration en nitrate et en ammonium. Les limitations dues aux deux derniers sont régies par une équation du type Michaelis-Menten, en particulier l'équation relative au nitrate prend en compte l'effet inhibiteur de l'ammonium [FDM90]. Les phytoplanctons subissent des pertes par exsudation — suintement — qui alimentent le compartiment d'azote organique dissous, des pertes par mortalité qui alimentent la matière organique particulaire de classe 1, et des pertes par broutage du zooplancton dont la taille est juste supérieure à la leur — e.g. le nanophytoplancton est consommé par le microzooplancton. Le microphytoplancton subit également des pertes par sédimentation.

Zooplanctons Les différents zooplanctons ingèrent les phytoplanctons possédant une taille juste inférieure à la leur. En outre, le nanozooplancton ingère des bactéries, le microzooplancton ingère de la matière organique particulaire de classe 1, et le mésozooplancton ingère de la matière organique

variables	valeur d'initialisation	variables	valeur d'initialisation
nh4	0.2	mop1	0.005
nz	0.02	mop2	0.005
miz	0.02	nod	0.005
mez	0.02	bac	0.02

TABLE 8.2 – Valeurs d'initialisations des variables d'état du système biologique ne dépendant pas des données FRONTAL.

particulaire de classes 1 et 2. L'ingestion se fait par filtration et est modélisée par une équation du type Michaelis-Menten. Seul environ 80% du volume filtré contribue à la croissance du zooplancton, le reste alimente les compartiments de matière organique particulaire de classes 1 et 2. Le zooplancton subit également des pertes par excrétion et par mortalité alimentant respectivement les compartiments d'azote organique dissous et d'ammonium, et la matière organique particulaire de classes 1 et 2. Enfin, le mésozooplancton subit des pertes par prédation — non explicitée dans les équations — qui contribuent à alimenter les compartiments mop2, nod et nh4.

Matière organique particulaire Comme nous l'avons remarqué dans les paragraphes précédents, les deux compartiments de matière organique particulaire sont alimentés par les processus d'ingestion du zooplancton et de mortalité du plancton. Ils subissent en outre des pertes par sédimentation et par dissolution, cette dernière alimentant le compartiment de l'azote organique dissous.

Azote organique dissous Comme nous l'avons vu précédemment, l'azote organique dissous provient de la dissolution de la matière organique particulaire de classes 1 et 2, de l'exsudation du phytoplancton et de l'excrétion du zooplancton. Il n'est assimilé que par les bactéries.

Bactéries L'ingestion par les bactéries de l'azote organique dissous et de l'ammonium est modélisé par une équation du type Michaelis-Menten. La totalité de ce qu'elles assimilent contribuent à leur croissance. Elles subissent ensuite des pertes par excrétion, par mortalité et par ingestion du nanozooplancton.

Concernant les conditions aux limites du système biologique, les échanges en surface et au fond sont supposés nuls. Pour les conditions initiales, celles du nitrate jusqu'à 200 mètres de profondeur sont issues des données FRONTAL entre les années 1985 et 1988 incluses ; et entre -200m et -400m, elles relèvent d'une interpolation linéaire entre la donnée FRONTAL à -200m et une valeur estimée à -600m [JT86, Min69]. Pour les autres variables, les conditions initiales consistent en des profils constants (voir Table 8.2) sauf pour les différents phytoplanctons qui sont basés sur les données FRONTAL de chlorophylle-a :

$$\begin{aligned} \texttt{pp} &= 0.01 * [\texttt{chla}]_{\text{FRONTAL}} \\ \texttt{np} &= 0.75 * [\texttt{chla}]_{\text{FRONTAL}} \\ \texttt{mp} &= 0.24 * [\texttt{chla}]_{\text{FRONTAL}} \end{aligned}$$

où $[\texttt{chla}]_{\text{FRONTAL}}$ est la donnée FRONTAL de chlorophylle-a convertie en $mmolNm^{-3}$ en notant que

$$1\ mgChlm^{-3} = rCChl * rmmolmgC * rNC\ mmolNm^{-3}$$

avec $rCChl = 50$ (rapport carbone/chlorophylle), $rmmolmgC = 1/12$ (rapport carbone $[mg]$/ carbone $[mmol]$) et $rNC = 8/53$ (rapport azote/carbone).

8.1.4 Couplage des deux modèles

Le couplage entre les modèles physique et biologique est un couplage faible : seul le modèle physique agit sur le modèle biologique. Cette action se traduit par l'ajout d'un terme de diffusion

dans chacune des équations contrôlant les variables d'état biologiques :

$$\frac{\partial}{\partial z}\left(\lambda\frac{\partial \mathtt{bio}}{\partial z}\right)$$

où `bio` désigne l'une des variables biologiques et $\lambda = \tilde{\lambda} * 1.1 * \sqrt{1-R_f}$ (voir les équations dans la Table 8.3).

8.1.5 Détail des processus biologiques

Le détail des processus est donné dans les paragraphes qui suivent ; les valeurs nominales des différents paramètres du modèle sont donnés dans la Table 8.4.

Croissance du phytoplancton Le taux de croissance du phytoplancton est donné par les équations suivantes, où A désigne `pp`, `np` ou `mp`.

$$\begin{aligned}
\mu_A &= \mu_{\mathtt{no3}A} + \mu_{\mathtt{nh4}A} \\
\mu_{\mathtt{no3}A} &= \mu_{max\mathrm{A}} * \mathrm{lim_{IA}} * \mathrm{lim_{TA}} * \mathrm{lim_{no3A}} \\
\mu_{\mathtt{nh4}A} &= \mu_{max\mathrm{A}} * \mathrm{lim_{IA}} * \mathrm{lim_{TA}} * \mathrm{lim_{nh4A}} \\
\mathrm{lim_{no3A}} &= \left(\frac{\mathtt{no3}}{\mathtt{no3} + \mathrm{K_{no3A}}}\right) * \exp(-\Psi \mathtt{nh4}) \\
\mathrm{lim_{nh4A}} &= \frac{\mathtt{nh4}}{\mathtt{nh4} + \mathrm{K_{nh4A}}} \\
\mathrm{lim_{IA}} &= \frac{2 * (1+\beta_{\mathrm{IA}}) * \frac{\mathrm{PAR}}{I_{opt\mathrm{A}}}}{\left(\frac{\mathrm{PAR}}{I_{opt\mathrm{A}}}\right)^2 + 2 * \beta_{\mathrm{IA}} * \frac{\mathrm{PAR}}{I_{opt\mathrm{A}}} + 1} \\
\mathrm{lim_{TA}} &= \max\left(\frac{2 * (1+\beta_{\mathrm{TA}}) * \frac{T - T_{let\mathrm{A}}}{T_{opt\mathrm{A}} - T_{let\mathrm{A}}}}{\left(\frac{T - T_{let\mathrm{A}}}{T_{opt\mathrm{A}} - T_{let\mathrm{A}}}\right)^2 + 2 * \beta_{\mathrm{IA}} * \frac{T - T_{let\mathrm{A}}}{T_{opt\mathrm{A}} - T_{let\mathrm{A}}} + 1}, 0\right)
\end{aligned}$$

avec

μ_A	taux de croissance de A (se dédouble en $\mu_{\mathtt{no3}A} + \mu_{\mathtt{nh4}A}$)
$\mu_{max\mathrm{A}}$	taux de croissance maximum de A
$\mathrm{lim_{no3A}}$	limitation par `no3` pour A
$\mathrm{lim_{nh4A}}$	limitation par `nh4` pour A
$\mathrm{K_{no3A}}$	coefficient de demi-saturation en `no3` pour A
$\mathrm{K_{nh4A}}$	coefficient de demi-saturation en `nh4` pour A
Ψ	coefficient d'inhibition pour `nh4`
$\mathrm{lim_{IA}}$	limitation par l'insolation pour A
β_{IA}	facteur de pente pour la courbe de photo-inhibition
$I_{opt\mathrm{A}}$	insolation optimale pour A
PAR	radiations actives pour la photosynthèse (photosynthetic active radiations)
$\mathrm{lim_{TA}}$	limitation par la température pour A
β_{TA}	facteur de pente pour la courbe de thermo-inhibition
$T_{opt\mathrm{A}}$	température optimale pour A
$T_{let\mathrm{A}}$	température léthale basse pour A
T	température.

Croissance des bactéries Le taux de croissance des bactéries est donné par les équations suivantes.

$$\begin{aligned}
\mu_{\mathtt{bac}} &= \mu_{max\mathtt{bac}} * (lim_{\mathtt{nodbac}} + lim_{\mathtt{nh4bac}}) \\
sub &= \min(\mathtt{nh4}, \eta * \mathtt{nod}) \\
lim_{\mathtt{nodbac}} &= \frac{\mathtt{nod}}{sub + \mathtt{nod} + K_{\mathtt{bac}}} \\
lim_{\mathtt{nodbac}} &= \frac{sub}{sub + \mathtt{nod} + K_{\mathtt{bac}}}
\end{aligned}$$

$$
\begin{aligned}
\frac{\partial \mathtt{no3}}{\partial t} &= \frac{\partial}{\partial z}\left(\lambda \frac{\partial \mathtt{no3}}{\partial z}\right) + nitr_{\mathtt{nh4}} * \mathtt{nh4} - \mu_{\mathtt{no3pp}} * \mathtt{pp} - \mu_{\mathtt{no3np}} * \mathtt{np} - \mu_{\mathtt{no3mp}} * \mathtt{mp} \\
\frac{\partial \mathtt{nh4}}{\partial t} &= \frac{\partial}{\partial z}\left(\lambda \frac{\partial \mathtt{nh4}}{\partial z}\right) - nitr_{\mathtt{nh4}} * \mathtt{nh4} - \mu_{\mathtt{nh4pp}} * \mathtt{pp} - \mu_{\mathtt{nh4np}} * \mathtt{np} - \mu_{\mathtt{nh4mp}} * \mathtt{mp} - \mu_{\mathtt{nh4bac}} * \mathtt{bac} \\
&\quad + (1 - excr_{\mathtt{nod}}) * excr_{\mathtt{nz}} * \mathtt{nz} + (1 - excr_{\mathtt{nod}}) * excr_{\mathtt{miz}} * \mathtt{miz} + (1 - excr_{\mathtt{nod}}) * excr_{\mathtt{mez}} * \mathtt{mez} \\
&\quad + (1 - excr_{\mathtt{nod}}) * excr_{pred} * pred * \mathtt{mez} + excr_{\mathtt{bac}}\mathtt{bac} \\
\frac{\partial \mathtt{pp}}{\partial t} &= \frac{\partial}{\partial z}\left(\lambda \frac{\partial \mathtt{pp}}{\partial z}\right) + (1 - exud_{\mathtt{pp}}) * (\mu_{\mathtt{no3pp}} + \mu_{\mathtt{nh4pp}}) * \mathtt{pp} - mort_{\mathtt{pp}} * \mathtt{pp} - brout_{\mathtt{ppnz}} * \mathtt{nz} \\
\frac{\partial \mathtt{np}}{\partial t} &= \frac{\partial}{\partial z}\left(\lambda \frac{\partial \mathtt{np}}{\partial z}\right) + (1 - exud_{\mathtt{np}}) * (\mu_{\mathtt{no3np}} + \mu_{\mathtt{nh4np}}) * \mathtt{np} - mort_{\mathtt{np}} * \mathtt{np} - brout_{\mathtt{npmiz}} * \mathtt{miz} \\
\frac{\partial \mathtt{mp}}{\partial t} &= \frac{\partial}{\partial z}\left(\lambda \frac{\partial \mathtt{mp}}{\partial z}\right) + (1 - exud_{\mathtt{mp}}) * (\mu_{\mathtt{no3mp}} + \mu_{\mathtt{nh4mp}}) * \mathtt{mp} - mort_{\mathtt{mp}} * \mathtt{mp} - brout_{\mathtt{mpmez}} * \mathtt{mez} \\
&\quad - sed_{\mathtt{mp}} \frac{\partial \mathtt{mp}}{\partial z} \\
\frac{\partial \mathtt{nz}}{\partial t} &= \frac{\partial}{\partial z}\left(\lambda \frac{\partial \mathtt{nz}}{\partial z}\right) + assim_{\mathtt{nz}} * (ing_{\mathtt{ppnz}} + ing_{\mathtt{bacnz}}) * \mathtt{nz} - excr_{\mathtt{nz}} * \mathtt{nz} - mort_{\mathtt{nz}} * \mathtt{nz} \\
&\quad - brout_{\mathtt{nzmiz}} * \mathtt{miz} \\
\frac{\partial \mathtt{miz}}{\partial t} &= \frac{\partial}{\partial z}\left(\lambda \frac{\partial \mathtt{miz}}{\partial z}\right) + assim_{\mathtt{miz}} * (ing_{\mathtt{npmiz}} + ing_{\mathtt{nzmiz}}) * \mathtt{miz} - excr_{\mathtt{miz}} * \mathtt{miz} - mort_{\mathtt{miz}} * \mathtt{miz} \\
&\quad - brout_{\mathtt{mizmez}} * \mathtt{mez} + assim_{\mathtt{mopmiz}} * ing_{\mathtt{mop1miz}} * \mathtt{miz} \\
\frac{\partial \mathtt{mez}}{\partial t} &= \frac{\partial}{\partial z}\left(\lambda \frac{\partial \mathtt{mez}}{\partial z}\right) + assim_{\mathtt{mez}} * (ing_{\mathtt{mpmez}} + ing_{\mathtt{mizmez}}) * \mathtt{mez} - excr_{\mathtt{mez}} * \mathtt{mez} - mort_{\mathtt{mez}} * \mathtt{mez} \\
&\quad - pred * \mathtt{mez} + assim_{\mathtt{mopmez}} * (ing_{\mathtt{mop1mez}} + ing_{\mathtt{mop2mez}}) * \mathtt{mez} \\
\frac{\partial \mathtt{bac}}{\partial t} &= \frac{\partial}{\partial z}\left(\lambda \frac{\partial \mathtt{bac}}{\partial z}\right) + assim_{\mathtt{bac}} * (\mu_{\mathtt{nodbac}} + \mu_{\mathtt{nh4bac}}) * \mathtt{bac} - excr_{\mathtt{bac}} * \mathtt{bac} - mort_{\mathtt{bac}} * \mathtt{bac} \\
&\quad - brout_{\mathtt{bacnz}} * \mathtt{nz} - brout_{\mathtt{bacmiz}}\mathtt{miz} \\
\frac{\partial \mathtt{mop1}}{\partial t} &= \frac{\partial}{\partial z}\left(\lambda \frac{\partial \mathtt{mop1}}{\partial z}\right) + mort_{\mathtt{pp}} * \mathtt{pp} + mort_{\mathtt{np}} * \mathtt{np} + mort_{\mathtt{mp}} * \mathtt{mp} + mort_{\mathtt{nz}} * \mathtt{nz} + mort_{\mathtt{miz}} * \mathtt{miz} \\
&\quad + mort_{\mathtt{bac}} * \mathtt{bac} + (1 - assim_{\mathtt{miz}}) * (ing_{\mathtt{ppmiz}} + ing_{\mathtt{bacmiz}} + ing_{\mathtt{nzmiz}} + ing_{\mathtt{npmiz}}) * \mathtt{miz} \\
&\quad + (1 - assim_{\mathtt{nz}}) * (ing_{\mathtt{ppnz}} + ing_{\mathtt{bacnz}}) * \mathtt{nz} + (1 - assim_{\mathtt{bac}}) * \mu_{\mathtt{nodbac}} * \mathtt{bac} - sed_{\mathtt{mop1}} \frac{\partial \mathtt{mop1}}{\partial z} \\
&\quad - assim_{\mathtt{mopmiz}} * ing_{\mathtt{mop1miz}} * \mathtt{miz} - assim_{\mathtt{mopmez}} * ing_{\mathtt{mop1mez}} * \mathtt{mez} - diss_{\mathtt{mop1}} * \mathtt{mop1} \\
\frac{\partial \mathtt{mop2}}{\partial t} &= \frac{\partial}{\partial z}\left(\lambda \frac{\partial \mathtt{mop2}}{\partial z}\right) + (1 - assim_{\mathtt{mez}}) * (ing_{\mathtt{npmez}} + ing_{\mathtt{nzmez}} + ing_{\mathtt{mpmez}} + ing_{\mathtt{mizmez}}) * \mathtt{mez} \\
&\quad + mort_{\mathtt{mez}} * \mathtt{mez} - assim_{\mathtt{mopmez}} * ing_{\mathtt{mop2mez}} * \mathtt{mez} - diss_{\mathtt{mop2}} * \mathtt{mop2} - sed_{\mathtt{mop2}} \frac{\partial \mathtt{mop2}}{\partial z} \\
&\quad + (1 - excr_{pred}) * pred * \mathtt{mez} \\
\frac{\partial \mathtt{nod}}{\partial t} &= \frac{\partial}{\partial z}\left(\lambda \frac{\partial \mathtt{nod}}{\partial z}\right) + diss_{\mathtt{mop1}} * \mathtt{mop1} + diss_{\mathtt{mop2}} * \mathtt{mop2} + exud_{\mathtt{pp}} * (\mu_{\mathtt{no3pp}} + \mu_{\mathtt{nh4pp}}) * \mathtt{pp} \\
&\quad + exud_{\mathtt{np}} * (\mu_{\mathtt{no3np}} + \mu_{\mathtt{nh4np}}) * \mathtt{np} + exud_{\mathtt{mp}} * (\mu_{\mathtt{no3mp}} + \mu_{\mathtt{nh4mp}}) * \mathtt{mp} - \mu_{\mathtt{nodbac}} * \mathtt{bac} \\
&\quad + excr_{\mathtt{nod}} * (excr_{\mathtt{nz}} * \mathtt{nz} + excr_{\mathtt{miz}} * \mathtt{miz} + excr_{\mathtt{mez}} * \mathtt{mez} + excr_{pred} * \mathtt{mez})
\end{aligned}
$$

TABLE 8.3 – Équations régissant le modèle biologique

avec

$\mu_{\texttt{bac}}$	taux de croissance des bactéries
$\mu_{max\texttt{bac}}$	taux de croissance maximum des bactéries
$\text{lim}_{\texttt{nodbac}}$	limitation par l'azote organique dissous
$\text{lim}_{\texttt{nh4bac}}$	limitation par l'ammonium
$\text{K}_{\texttt{bac}}$	coefficient de demi-saturation pour les bactéries
sub	concentration du substrat azoté
η	rapport d'assimilation `nh4`/`nod`.

Ingestion par les hétérotrophes Le processus d'ingestion par les hétérotrophes est donné par les équations suivantes.

$$\begin{aligned} brout_{BA} &= vol * \varepsilon_{BA} * Biom_A \\ vol &= \frac{ing_A}{biom_{BA}} \\ biom_{BA} &= \sum_{proies} \varepsilon_{BA} * biom_A \\ ing_A &= \begin{cases} 0 & \text{si } biom_{BA} \leq seuil_A \\ I_{maxA} * \left(\frac{biom_{BA} - seuil_A}{biom_{BA} - seuil_A + K_A}\right) & \text{sinon} \end{cases} \end{aligned}$$

avec

$brout_{BA}$	taux de broutage sur la proie B par le prédateur A
vol	volume exploré
$biom_{BA}$	biomasse de proies capturables par le prédateur A
ε_{BA}	efficacité de capture de la proie B par le prédateur A
$biom_A$	biomasse du prédateur A
ing_Ab	taux d'ingestion de A
I_{maxA}	taux d'ingestion maximal pour A
$seuil_A$	seuil de nutrition minimal pour A
K_A	coefficient de demi-saturation pour A.

Taux d'excrétion des hétérotrophes Le taux d'excrétion des hétérotrophes est donné par l'équation suivante.

$$excr_A = \alpha_A * \beta_A^T$$

avec

$excr_A$	taux d'excrétion de A
α_A	taux d'excrétion de A à 0°C
β_A	facteur de pente pour la courbe d'excrétion
T	température.

8.2 Analyse de sensibilité de MODECOGeL

Le travail d'analyse de sensibilité globale que nous présentons est essentiellement introductif, et constitue un point de départ à des travaux futurs de plus grande envergure. Comme nous l'avons remarqué jusqu'à présent, les travaux existants sur MODECOGeL en lien avec l'analyse de sensibilité se résument principalement à étudier les variations d'une sortie d'intérêt en faisant varier à tour de rôle les valeurs des paramètres à plus ou moins 20% autour de leur valeur nominale. Ceci revient à effectuer moins de deux cents simulations d'un modèle qui possède plus de quatre-vingt paramètres, et ne peut probablement pas produire une information fiable pour appréhender la sensibilité des paramètres.

Notre travail se restreint à une sortie d'intérêt principale : la concentration de chlorophylle-a, qui constitue un indicateur de l'activité biologique générale. Elle est étudiée sous trois versions distinctes :

croissance des autotrophes	unité	pp	np	mp	
μ_{maxA}	j^{-1}	3	2.5	2	
$K_{\mathtt{no3}A}$	$mmolNm^{-3}$	0.5	0.7	1	
$K_{\mathtt{nh4}A}$	$mmolNm^{-3}$	0.3	0.5	0.7	
Ψ	$(mmolNm^{-3})^{-1}$	1.46			
$\beta_{\mathrm{I}A}$	sans dim.	-0.8	-0.7	-0.6	
I_{optA}	Wm^{-2}	10	15	20	
$\beta_{\mathrm{T}A}$	sans dim.	-0.5	-0.5	-0.55	
T_{letA}	°C	9	9	9	
T_{optA}	°C	15	15	15	
$mort_A$	j^{-1}	0.06	0.05	0.04	
$exud_A$	%	6	5	4	
croissance des hétérotrophes	unité	nz	miz	mez	bac
η	sans dim.				0.6
μ_{maxA}	j^{-1}				2
I_{maxA}	j^{-1}	3	2	1.5	
K_A	$mmolm^{-3}$	0.5	0.75	1	0.5
$assim_A$	sans dim.	0.8	0.8	0.8	1
$assim_{mopA}$	sans dim.		0.5	0.5	
$seuil_A$	$mmolm^{-3}$	0.05	0.03	0.01	
$mort_A$	j^{-1}	0.06	0.05	0.03	0.06
α_A	j^{-1}	0.15	0.1	0.05	0.15
β_A	sans dim.	1.05	1.05	1.02	1.04
efficacité de capture des hétérotrophes	unité	bac	pp	np	nz
$\varepsilon_{B\mathtt{nz}}$	sans dim.	0.8	1		
$\varepsilon_{B\mathtt{miz}}$	sans dim.			1	0.8
		np	nz	mp	miz
$\varepsilon_{B\mathtt{mez}}$	sans dim.			0.8	0.8
		mop1	mop2		
$\varepsilon_{\mathtt{mopmiz}}$	sans dim.	0.2			
$\varepsilon_{\mathtt{mopmez}}$	sans dim.	0.2	0.2		
prédation par les carnivores	unité				
$seuil_{pred}$	$mmolm^{-3}$	0.02			
$pred_{max}$	j^{-1}	1			
K_{pred}	$mmolm^{-3}$	1			
$excr_{pred}$	%	33			
azote organique et inorganique	unité	mop1	mop2	nh4	nod
$sed_{\mathtt{mop}A}$	mj^{-1}	1.5	95		
$diss_{\mathtt{mop}A}$	j^{-1}	0.065	0.06		
$nitr_{\mathtt{nh4}}$	j^{-1}			0.03	
$excr_{\mathtt{nod}}$	%				25
autres paramètres	unité				
$sed_{\mathtt{mp}}$	mj^{-1}	1			
PAR_0	%	50			
K_{eau}	m^{-1}	0.04			
K_{chl1}	$mgChl^{-1}m^2$	0.0088			
K_{chl2}	$mgChl^{-2/3}m$	0.054			

TABLE 8.4 – Valeurs nominales des différents paramètres du modèle MODECOGeL.

- maximum annuel de la concentration de chlorophylle-a en surface
- moyenne annuelle de la concentration de chlorophylle-a en surface
- maximum annuel de la concentration moyenne de chlorophylle-a entre -20m et -50m.

Notre étude se divise principalement en deux phases distinctes. La première ne considère que six paramètres incertains, les autres étant fixés à leur valeur nominale, et consiste essentiellement à prendre la mesure de la complexité — intercations entre plusieurs paramètres, non-linéarités, etc. — des sorties précédemment citées en mettant en œuvre une analyse par les dérivées, et en calculant les indices des Sobol' d'ordre 1 et 2 à des résolutions très faibles — i.e. avec un nombre de simulations inférieur à 2000 — puis à des résolutions moyennes — i.e. avec un nombre de simulations inférieur à 10000 — à l'aide de la méthode RBD.

Dans la seconde phase, nous incorporons deux nouvelles variables incertaines au modèle précédent et nous effectuons des calculs d'indices de Sobol' à des résolutions plus élevées — i.e. avec un nombre de simulations entre 50000 et 100000. Dans cette seconde phase, au regard des conclusions de la première phase qui laisse apparaître des sorties de modèle fortement non-linéaires et difficiles à appréhender à l'aide d'une méthode basée sur l'analyse harmonique telle que RBD — voir Chapitre 5, nous privilégions des techniques d'estimation des indices de Sobol' basée sur la méthode de Monte Carlo. Plus précisément, nous appliquons la méthode de Sobol' dans la version que nous avons introduite au Chapitre 6 de cette thèse — qui réduit le coût dans ce cas précis d'un facteur 5 par rapport à la méthode classique — ainsi que l'approche par quasi-régression suivant une base de polynômes orthogonaux, qui comme nous l'avons remarqué au Chapitre 2 de cette thèse permet de dériver des intervalles de confiance pour les indices de Sobol'.

8.2.1 Première phase

Dans cette première étude, nous avons uniquement considéré comme sortie du modèle, le maximum annuel de la concentration de chlorophylle-a en surface. De plus, sur les 85 paramètres du modèle biologique, nous n'en considérons que 6 comme incertains, les autres étant fixés à leur valeur nominale. La concentration en chlorophylle est définie à partir de la somme des concentrations des différents phytoplanctons à un changement d'unité près :

$$\texttt{chla} = (\texttt{pp} + \texttt{np} + \texttt{mp}) * 1.59$$

où $1.59 = 1/(rCChl * rNC * rmmolmgC)$ est la constante prenant en compte le changement d'unité, avec $rCChl = 50$ (rapport carbone/chlorophylle), $rmmolmgC = 1/12$ (rapport carbone $[mg]$/ carbone $[mmol]$) et $rNC = 8/53$ (rapport azote/carbone). Les 6 paramètres considérés comme incertains dont on veut analyser la sensibilité sont les taux de croissance maximaux de chacun des 3 phytoplanctons notés *mumaxpp*, *mumaxnp* et *mumaxmp*, et l'insolation optimale pour chacun des 3 phytoplanctons notés *paroptpp*, *paroptnp* et *paroptmp* ; nous les avons sélectionnés car nous pensons a priori qu'ils jouent un rôle important sur les sorties du modèle que l'on souhaite étudier.

Estimation des indices de Sobol' La première expérience consiste à évaluer indépendamment à 15 reprises les indices de Sobol' d'ordre 1 des 6 paramètres à l'aide de la méthode RBD, et de donner une représentation graphique des résultats sous forme de diagrammes en "boîte à moustaches" [1]. Les 6 paramètres d'entrée sont supposés indépendants et naïvement uniformément distribués sur un intervalle $[x_{nom}/3, x_{nom} \times 3]$ où x_{nom} désigne leur valeur nominale (voir Table 8.5). Chacune des 15 estimations est menée sur un plan d'expérience — spécifique à la méthode RBD — contenant 1200 points, et les indices de Sobol' d'ordre 1 sont évalués en prenant en compte les 12 premières harmoniques relatives à chacun des paramètres d'intérêt (voir Figure 8.3). Les résultats montrent une variance importante des estimateurs, les différences entre les valeurs minimale et maximale pouvant atteindre près de 0.05. En outre, on remarque que la borne supérieure de la somme des 6 indices d'ordre 1 reste en deça des 50%, ce qui laisse présager un modèle fortement non-additif.

1. Dans les "boîtes à moustaches", le trait rouge au milieu de la boîte représente la médiane, les délimiteurs inférieurs et supérieurs de la boîte représentent respectivement les quantiles q et Q à 25% et 75%, les "moustaches" représentent les valeurs $q - 1.5 \times (Q - q)$ et $Q + 1.5 \times (Q - q)$, et les croix rouges éventuelles représentent les valeurs en dehors des deux moustaches — considérées comme statistiquement aberrantes.

paramètre	valeur nominale
mumaxpp	3
mumaxnp	2.5
mumaxmp	2
paroptpp	10
paroptnp	15
paroptmp	20

TABLE 8.5 – Valeurs nominales des paramètres d'entrée.

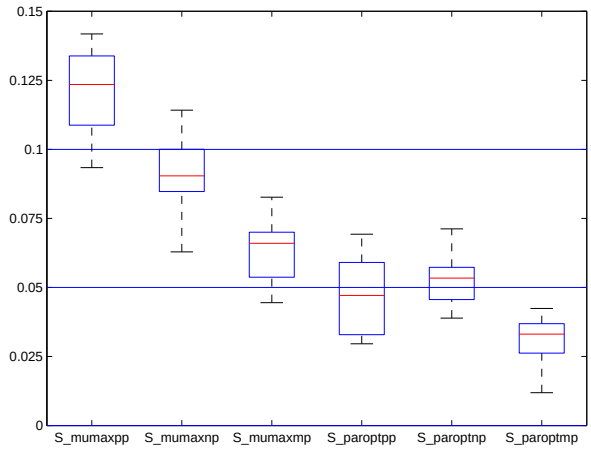


FIGURE 8.3 – Graphique en "boîte à moustaches" de 15 évaluations indépendantes des 6 indices de Sobol' d'ordre 1 par la méthode RBD effectuée avec une taille d'échantillon de 1200.

La seconde expérience consiste à reprendre la précédente en appliquant cette fois-ci une méthode RBD basée sur un tableau orthogonal de force 2 — voir Chapitre 5 — afin d'évaluer également les indices de Sobol' d'ordre 2 et ainsi d'avoir un premier aperçu du comportement non-additif du modèle. Les estimations des indices de Sobol' d'ordre 1 et 2 sont répétées indépendamment 10 fois sur un plan d'expériences contenant $1681 (= 41^2)$ points. Les indices de Sobol' d'ordre 2 sont évalués en prenant en compte les harmoniques relatives à chaque paire de paramètres d'intérêt suivant une croix de Zaremba — ou croix hyperbolique — de paramètre 12 (voir Chapitre 5 Section 5.4.1). Les résultats sont représentés graphiquement suivant des diagrammes en "boîte à moustaches" sur la Figures 8.4. Les résultats montrent une diminution de la variance des estimateurs des indices d'ordre 1 par rapport à l'expérience précédente, due à la fois à l'augmentation de la taille d'échantillon — 1681 contre 1200 précédemment — et à la stratification de force supérieure. Quant aux estimations des indices d'ordre 2, elles présentent une variance élevée. On note également que la borne supérieure de la somme des indices d'ordre 1 et 2 atteint approximativement les 100%, mais la variance importante des estimateurs ne peut pas permettre de conclure catégoriquement que le modèle considéré possède une dimension effective par superposition de 2 — i.e. pas d'effets d'ordre supérieur à 2.

La troisième expérience consiste à considérer que les paramètres ne sont plus naïvement supposés distribués suivant des lois uniformes, mais suivant des lois plus localisées autour de la valeur nominale, comme les log-normales ou gamma (voir Figure 8.5). L'expérience est toujours menée en appliquant une méthode RBD basée sur un tableau orthogonal de force 2 afin d'évaluer également les indices de Sobol' d'ordre 2 et ainsi d'avoir un premier aperçu du comportement non-additif du modèle. Les estimations des indices de Sobol' d'ordre 1 et 2 sont répétées indépendamment 10 fois sur un plan d'expériences contenant $1681 (= 41^2)$ points. Les paramètres des lois log-normales et gamma sont fixés de telle sorte que la moyenne soit la valeur nominale du paramètre considéré et que le quantile à 99% corresponde à trois fois cette valeur nominale. Les plans d'expériences ad hoc pour une telle

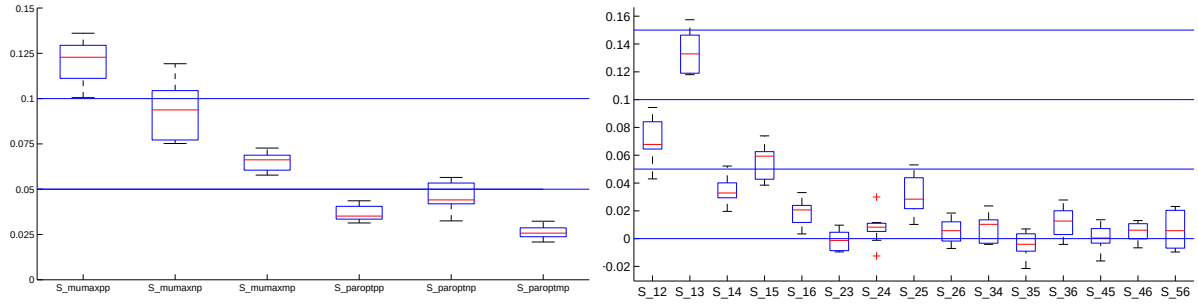


FIGURE 8.4 – Graphique en "boîte à moustaches" de 10 évaluations indépendantes des 6 indices de Sobol' d'ordre 1 (à gauche) et des 15 indices d'ordre 2 après correction de biais (à droite) par la méthode RBD effectuée avec une taille d'échantillon de 1681 et des lois uniformes en entrée. Légende en abscisse : 1 (mumaxpp), 2 (mumaxnp), 3 (mumaxmp), 4 (paroptpp), 5 (paroptnp), 6 (paroptmp).

application de la méthode RBD à des distributions non-uniformes sont obtenus par transformation inverse. Les résultats sont représentés graphiquement suivant des diagrammes en "boîte à moustaches" sur les Figures 8.6 et8.7.

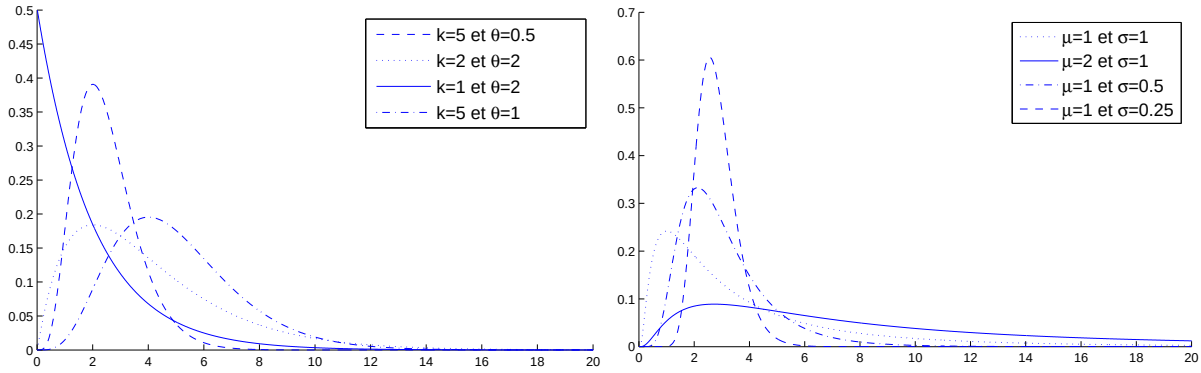


FIGURE 8.5 – Tracé des densités des distribution gamma $f(x;k,\theta) = x^{k-1}\exp(-x/\theta)/(\Gamma(\theta)\theta^k)$ (à gauche) et des distributions log-normales $f(x;\mu,\sigma) = \exp(((\ln(x)-\mu)/\sigma)^2/2)/(x\sigma\sqrt{2\pi})$ (à droite) pour quelques jeux de paramètres.

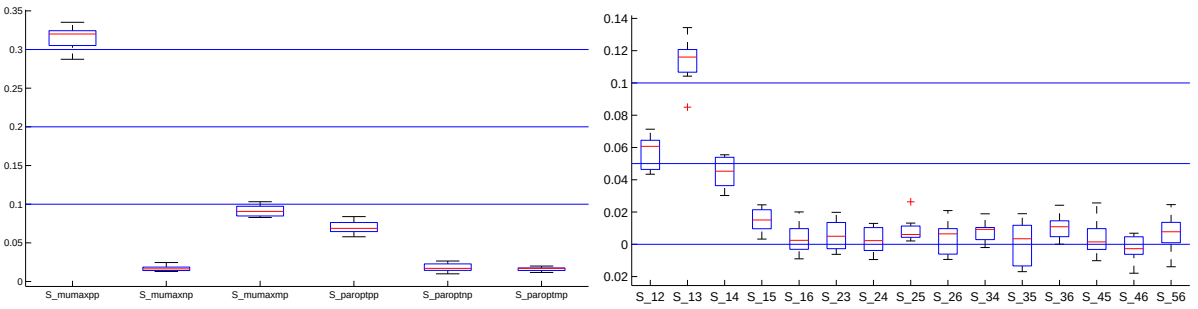


FIGURE 8.6 – Graphique en "boîte à moustaches" de 10 évaluations indépendantes des 6 indices de Sobol' d'ordre 1 (à gauche) et des 15 indices d'ordre 2 après correction de biais (à droite) par la méthode RBD effectuée avec une taille d'échantillon de 1681 et des lois log-normales en entrée. Légende en abscisse : 1 (mumaxpp), 2 (mumaxnp), 3 (mumaxmp), 4 (paroptpp), 5 (paroptnp), 6 (paroptmp).

On constate que les indices de Sobol' ne sont pas robustes au changement de loi effectué, au sens

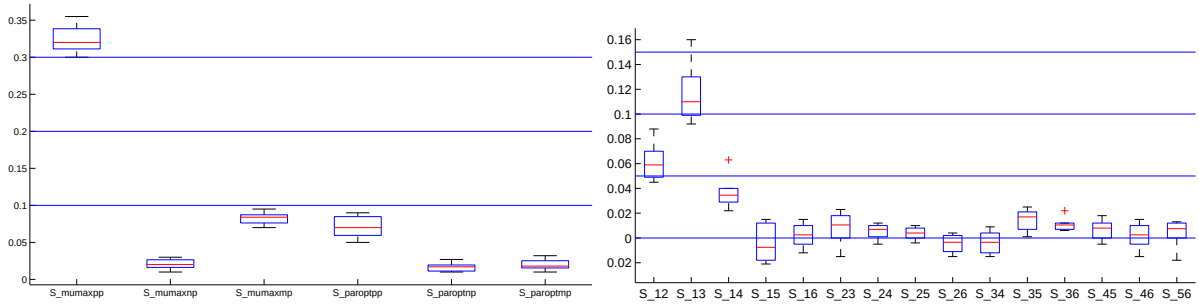


FIGURE 8.7 – Graphique en "boîte à moustaches" de 10 évaluations indépendantes des 6 indices de Sobol' d'ordre 1 (à gauche) et des 15 indices d'ordre 2 après correction de biais (à droite) par la méthode RBD effectuée avec une taille d'échantillon de 1681 et des lois gamma en entrée. Légende en abscisse : 1 (mumaxpp), 2 (mumaxnp), 3 (mumaxmp), 4 (paroptpp), 5 (paroptnp), 6 (paroptmp).

où la classification des différents indices d'ordre 1 est modifiée suivant qu'on utilise des lois uniformes en entrée ou des lois non-uniformes. On remarque que le fait d'utiliser des lois recentrées autour de la valeur nominale a pour effet de "simplifier" la structure du modèle, au sens où certains paramètres qui apparaîssaient comme sensibles dans l'analyse vis-à-vis d'une loi uniforme sont considérés comme non-actifs vis-à-vis d'une distribution plus centrée autour de leur valeur nominale. Cette différence d'appréciation pour les lois uniformes est due à l'attribution d'un même poids aux valeurs proches de la valeur nominale et à celles en queue de distribution lorsqu'on s'éloigne de la valeur nominale par valeur supérieure. Il nous semble donc plus judicieux de ne considérer par la suite que des distributions "moins grossières" qui pondèrent la valeur du paramètre suivant sa proximité avec sa valeur nominale. Nous retenons la loi gamma qui permet d'accéder à une grande variété de profils à l'aide de seulement deux paramètres k et θ, et dont la moyenne est simplement donnée par le produit $k\theta$.

Enfin, dans la quatrième expérience, nous reprenons les caractéristiques de la troisième expérience en nous cantonnant à des entrées distribuées suivant des lois gamma, en augmentant le nombre de points de simulations — ici 10000 — et en analysant les trois sorties différentes décrites dans l'introduction de cette section. En outre, les plages de variation des paramètres sont élargies : les paramètres des gamma sont fixés de telle sorte que la moyenne soit la valeur nominale du paramètre considéré et le quantile à 99% corresponde à quatre fois cette valeur nominale. Les estimations des différents indices de Sobol' sont représentés graphiquement en fonction du nombre d'harmoniques prises en compte (voir Figures 8.8 à 8.14).

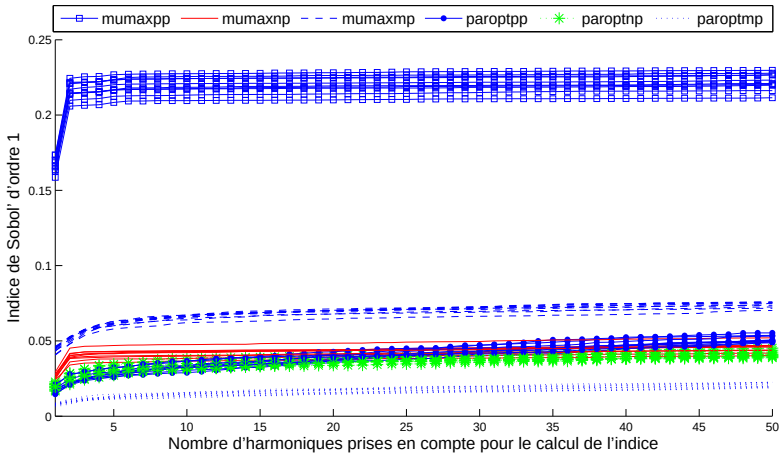


FIGURE 8.8 – Tracé des indices de Sobol' d'ordre 1 en fonction du nombre d'harmoniques prises en compte dans leur calcul. Sortie : maximum annuel de la concentration de chlorophylle-a en surface.

Concernant l'estimation des indices de Sobol' d'ordre 1, on remarque que même après avoir pris en

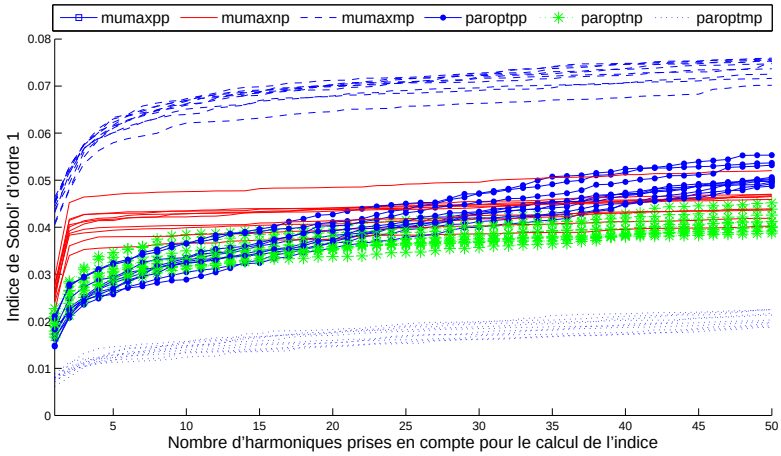


FIGURE 8.9 – Agrandissement du tracé des indices de Sobol' d'ordre 1 en fonction du nombre d'harmoniques prises en compte dans leur calcul. Sortie : maximum annuel de la concentration de chlorophylle-a en surface.

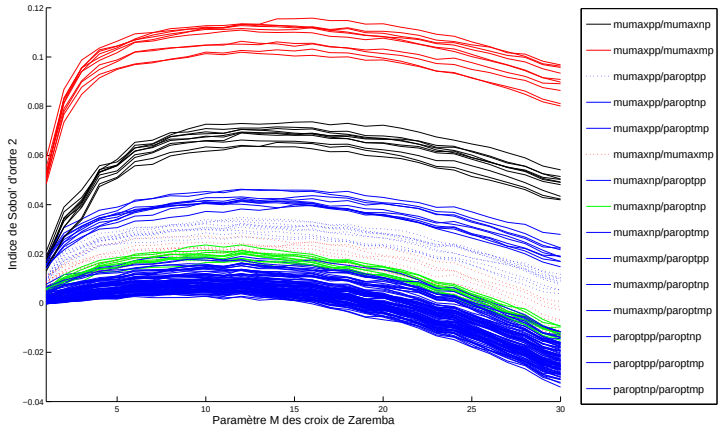


FIGURE 8.10 – Tracé des indices de Sobol' d'ordre 2 en fonction du nombre d'harmoniques prises en compte dans leur calcul. Sortie : maximum annuel de la concentration de chlorophylle-a en surface.

compte 50 harmoniques, certains calculs d'indices ne semblent pas avoir convergé (voir e.g. *paroptpp* sur la Figure 8.9). Ceci laisse penser que le spectre trigonométrique analysé ne décroît pas suffisamment à l'infini, et donc que la décomposition en série de Fourier n'est pas adaptée à ce modèle. Cela doit nous inciter à travailler sur une décomposition spectrale relative aux polynômes de Laguerre généralisés qui, eux, forment un système orthonormé relativement à des lois gamma. La seconde remarque concerne l'estimation des indices de Sobol' d'ordre 2. On remarque principalement que la correction de biais effectuée produit des estimations d'indices qui diminuent avec l'augmentation du nombre d'harmoniques prises en compte dans le calcul. Une telle dérive dans l'estimation corrigée des indices ne doit pas apparaître comme surprenante étant donnée l'expression des biais et des biais résiduels après correction dans la méthode RBD (voir Chapitre 5). Néanmoins, pour pouvoir conclure clairement sur les valeurs d'indices, il faudrait que la convergence se fasse très rapidement, jusqu'à atteindre un niveau bien marqué — i.e. l'estimation apparaîsse comme constante par rapport à l'augmentation du nombre d'harmoniques prises en compte — avant que l'estimation de l'indice diminue. Dans notre cas, ce n'est généralement pas le cas au sens où la convergence est plutôt lente vers l'estimation la plus grande, cette valeur maximale est peu marquée et l'estimation diminue immédiatement ce niveau maximal atteint. Une fois encore, la nécessité d'un spectre de Fourier à décroissance rapide

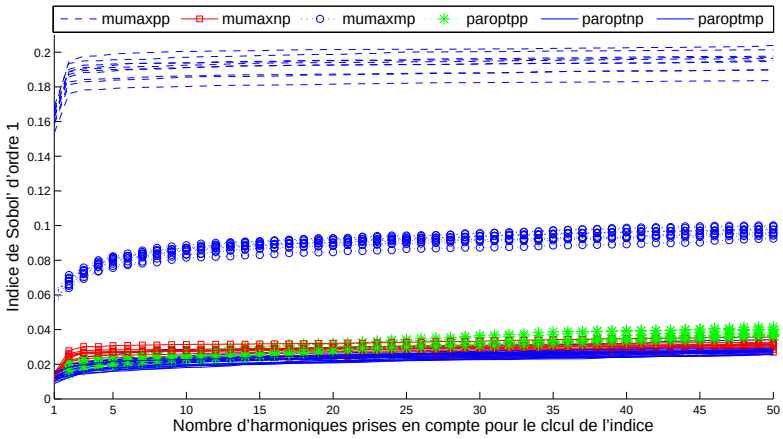


FIGURE 8.11 – Tracé des indices de Sobol' d'ordre 1 en fonction du nombre d'harmoniques prises en compte dans leur calcul. Sortie : moyenne annuelle de la concentration de chlorophylle-a en surface.

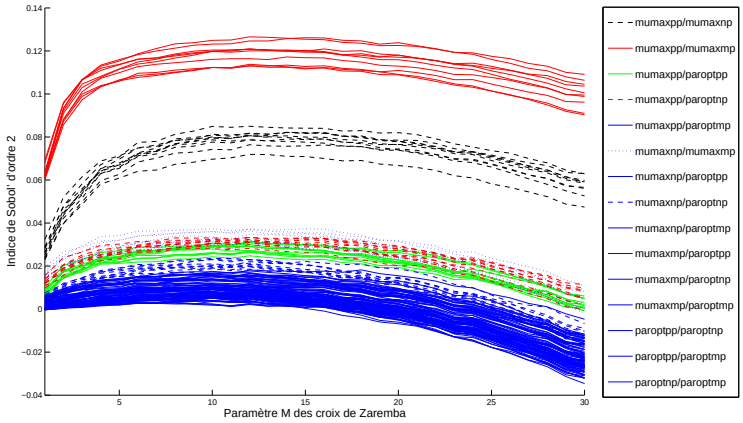


FIGURE 8.12 – Tracé des indices de Sobol' d'ordre 2 en fonction du nombre d'harmoniques prises en compte dans leur calcul. Sortie : moyenne annuelle de la concentration de chlorophylle-a en surface.

est centrale dans la méthode RBD, et le modèle ne semble pas remplir cette hypothèse.

De manière plus générale sur l'ensemble des expériences menées, on remarque que le paramètre *mumaxpp* apparaît comme le plus actif sur les sorties considérées. Le modèle apparaît en outre comme assez complexe, ne présentant pas de paramètre totalement inactif. Ceci peut être attribué aux plages de valeurs relativement grandes qu'on considère pour chacun des paramètres, ceux-ci pouvant varier du simple au triple, voire au quadruple dans la dernière expérience.

Analyse par les dérivées L'analyse de sensibilité précédente est complétée par une analyse plus qualitative en considérant les dérivées partielles d'ordre 1 suivant chacun des paramètres d'entrée. Ces dérivées partielles sont évaluées en 1681 points répartis sur un tableau orthogonal de force 2 construit à partir de lois marginales log-normales — i.e. l'un des 10 plans utilisés dans l'expérience précédente — en utilisant un schéma aux différences finies (avec $\varepsilon = 10^{-3}$). Les résultats sont représentés graphiquement en histogramme dans la Figure 8.15.

On constate à la vue des résultats que le modèle ne présente aucune dépendance linéaire par rapport à l'un des paramètres, et plus généralement aucune monotonie. De manière plus fine, nous cherchons à localiser ces valeurs de gradients afin de comprendre si les changements de signes sont répartis sur tout le domaine de variation des paramètres (phénomène oscillatoire) ou si ils sont plus localisés (effet sur la sortie très marqué). Pour cela rappelons que le plan d'expérience où les dérivées

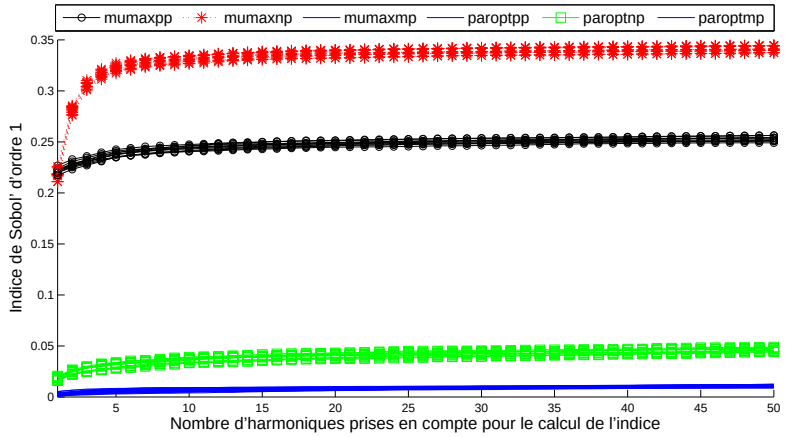


FIGURE 8.13 – Tracé des indices de Sobol' d'ordre 1 en fonction du nombre d'harmoniques prises en compte dans leur calcul. Sortie : maximum annuel de la concentration moyenne de chlorophylle-a entre -20m et -50m.

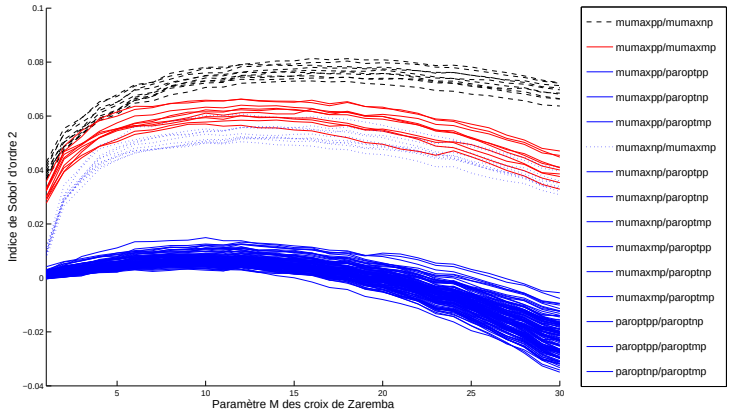


FIGURE 8.14 – Tracé des indices de Sobol' d'ordre 2 en fonction du nombre d'harmoniques prises en compte dans leur calcul. Sortie : maximum annuel de la concentration moyenne de chlorophylle-a entre -20m et -50m.

partielles sont calculées est un tableau orthogonal de force 2 à 41 niveaux (=1681 points) — notons le $(\mathbf{x}_j)_{j=1..1681}$ où $\mathbf{x}_j = (x_{j1}, \ldots, x_{j6})$ est un vecteur de dimension 6. Par conséquence, pour $i \in \{1, \ldots 6\}$, les $(x_{ji})_{j=1..1681}$ ne prennent que 41 valeurs distinctes et chacune est prise 41 fois ; notons $(x_i^k)_{k=1..41}$ ces valeurs. Alors nous considérons les quantités

$$\text{dérivée moyenne}(x_i^k) = \frac{1}{41} \sum_{\mathbf{x}_j \mid x_{ij}=x_i^k} \frac{f(\mathbf{x}_j + \mathbf{1}_i * 0.001) - f(\mathbf{x}_j)}{0.001}$$

où $\mathbf{1}_i = (0, \ldots, 0, 1, 0, \ldots, 0)$ avec la valeur 1 sur la i-ème composante, i.e. la valeur de la dérivée partielle relative à la i-ème variable obtenue en moyenne lorsque les 5 autres variables parcourent leur domaine de définition, et que la i-ème variable a pour valeur x_i^k. On peut alors représenter les courbes formées par ces 41 valeurs moyennes pour chacun des paramètres. Les résultats sont présentés à la Figure 8.16.

Le fait de considérer des valeurs moyennes lisse l'information et ne fait ressortir que les phénomènes importants. Le seul phénomène intéressant qui ressort est que la valeur moyenne de la dérivée partielle par rapport au paramètre *mumaxpp* change de signe de façon très marquée lorsque *mumaxpp* est de part et d'autre de sa valeur nominale égale à 3. La valeur nominale de *mumaxpp* apparaît ainsi

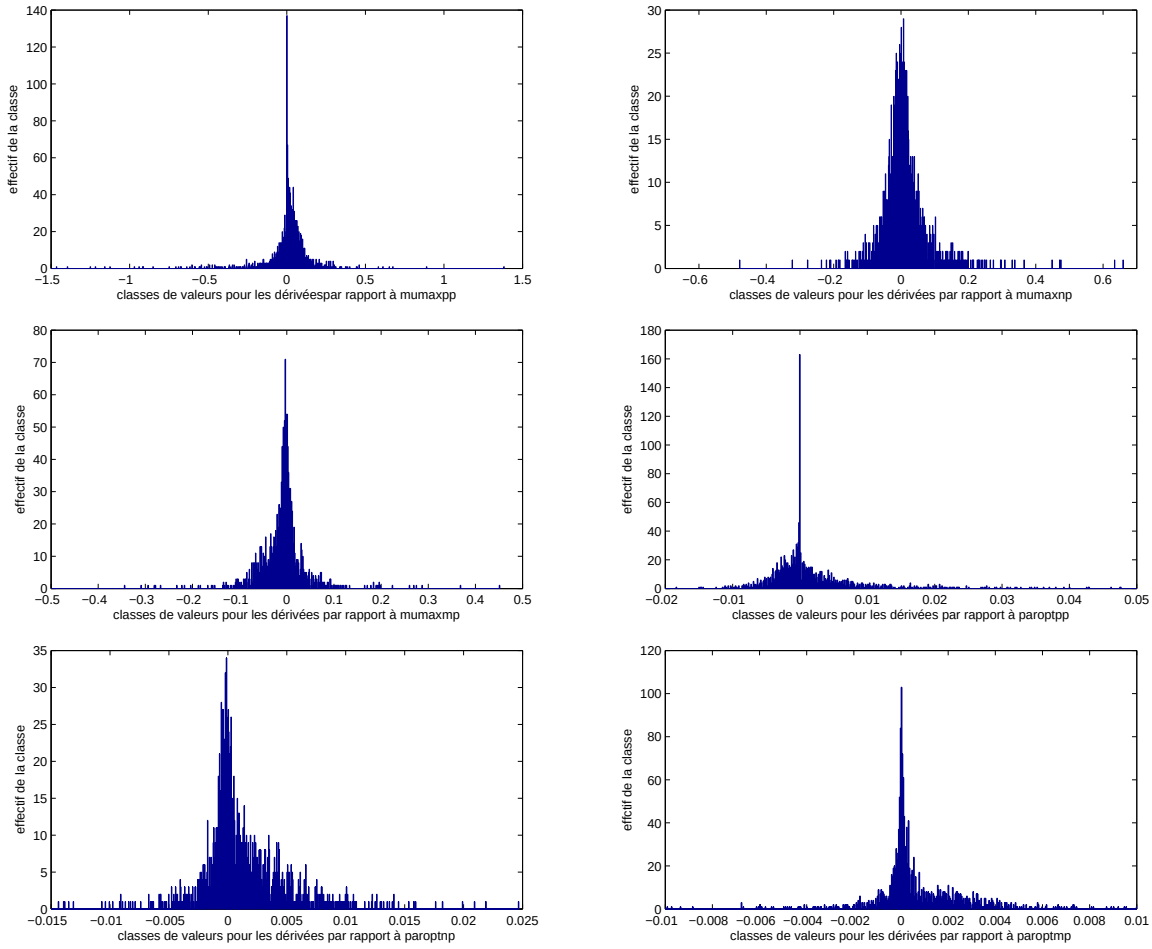


FIGURE 8.15 – Histogrammes représentant la répartition de 1681 dérivées partielles relatives à chacun des paramètres.

comme un maximum local relatif pour la sortie considérée et corrobore le fait que *mumaxpp* est identifié comme un paramètre actif dans le calcul des indices de Sobol' qui précède.

8.2.2 Seconde phase

Au regard des résultats de la première phase, nous nous orientons vers le calcul des indices de Sobol' par des méthodes autres que la méthode RBD ; en particulier des méthodes qui permettent d'estimer l'erreur commise lors de l'estimation des indices de Sobol'. Ainsi, les 10 ou 15 estimations indépendantes que nous effectuions avec la méthode RBD, afin d'avoir un aperçu de l'erreur d'estimation, ne sont plus nécessaires, et donc nous pouvons effectuer par la suite une seule estimation à l'aide d'un plan d'expérience de taille plus grande. Ainsi nous mettons en œuvre la méthode de Sobol' dans sa version que nous avons introduite au Chapitre 6 qui en réduit le coût original, et nous appliquons également une quasi-régression relativement à une décomposition spectrale suivant les polynômes de Laguerre généralisés — adaptés aux distributions gamma. Les calculs par la méthode de Sobol' sont faits sur deux hypercubes latins répliqués ($2 * 65536$ points) pour les indices d'ordre 1, et à l'aide de deux hypercubes latins répliqués basés sur un tableau orthogonal de force 2 ($2 * 66049$ points) pour les indices d'ordre 2. L'analyse par quasi-régression est menée à l'aide d'un des deux hypercubes latins répliqués précédemment cités. Les 3 sorties du modèle étudiées restent inchangées, mais nous ajoutons deux paramètres incertains dont nous souhaitons connaître le comportement : la vitesse de sédimentation de la matière organique de classe 1, *sedmop*1, normalement fixée à sa valeur nominale 1.5, et le taux de croissance maximal des bactéries, *mumaxbac*, normalement fixé à

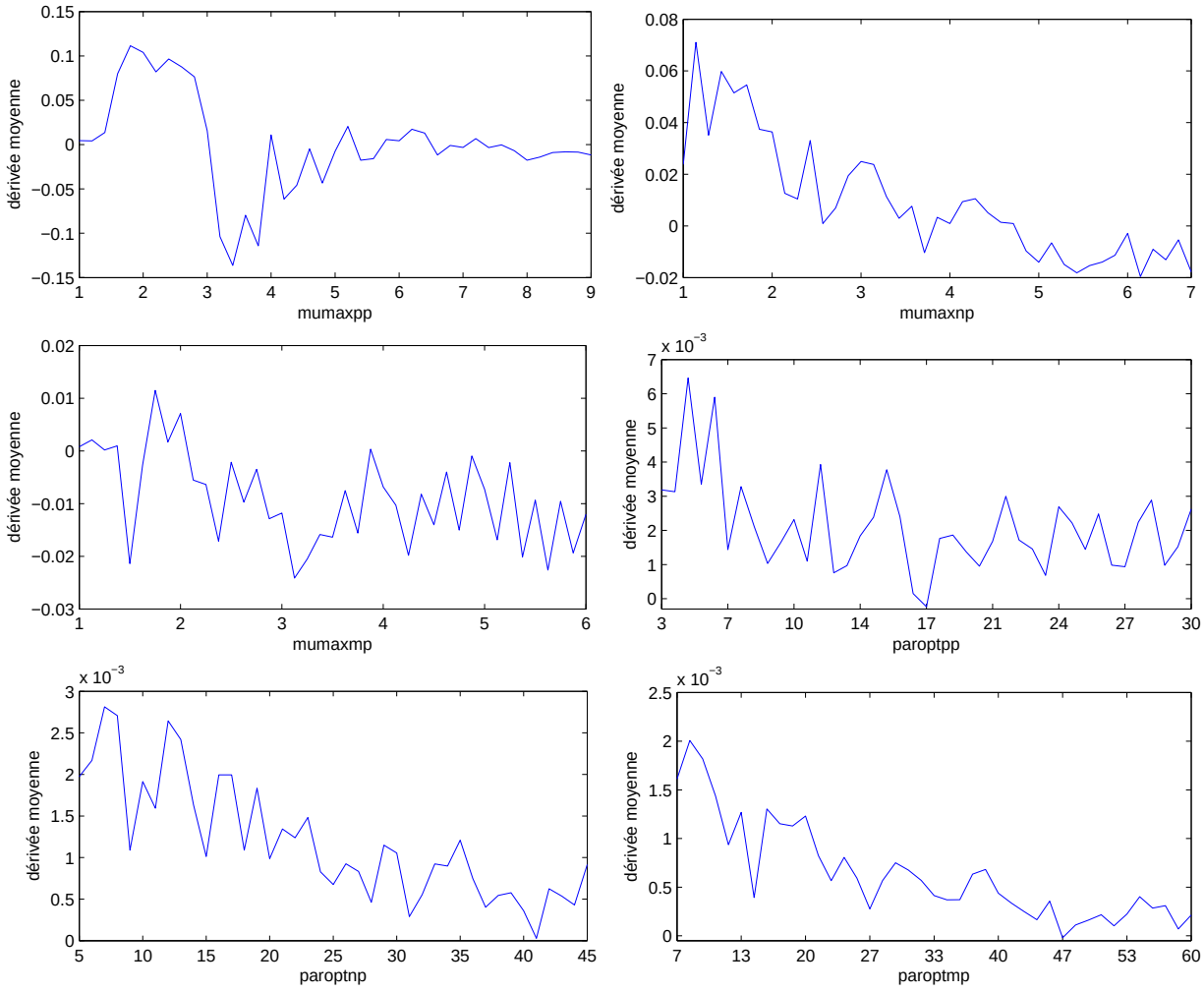


FIGURE 8.16 – Graphique des valeurs moyennes des dérivées partielles de chacun des paramètres d'entrée en fonction de la valeur où elles sont calculées.

sa valeur nominale 2. Les 8 paramètres incertains sont tous modélisés par des lois gamma centrées sur leur valeur nominale et dont le quantile à 99% coïncide avec le triple de cette valeur nominale.

Nous commençons l'étude par l'analyse d'un certain nombre de distributions évaluées à partir des données issues d'un des deux hypercubes latins répliqués mentionnés ci-avant. Plus précisément, nous étudions la répartition des valeurs et la localisation du maximum annuel en surface de concentration en picophytoplancton, en nanophytoplancton et en mésophytoplancton, et également du maximum annuel de leur concentration moyenne entre -20m et -50m ; ceci dans le but de mieux comprendre pourquoi le paramètre *mumaxpp* ressort tellement parmi les variables actives par rapport aux autres paramètres dans l'étude de la première phase. Les histogrammes sont présentés dans les Figures 8.17 à 8.22.

On constate que seul le maximum annuel relatif au picophytoplancton est bien localisé. En outre les nanophytoplancton et mésophytoplancton montrent, dans une proportion importante, un maximum annuel qui est localisé au début de l'année, qui laisse penser qu'aucune floraison printanière n'a eu lieu et que la concentration dérive au cours de l'année. Enfin, la plus forte concentration maximale annuelle est clairement celle du picophytoplancton. Par suite on comprend mieux pourquoi l'analyse du maximum annuel de chlorophylle-a en surface, ou en profondeur, fait apparaître le paramètre *mumaxpp* comme incontournable ; rappelons à cet effet que la concentration en chlorophylle-a est définie dans le modèle comme la somme des concentrations en pico-, nano- et mésophytoplancton, à une constante multiplicative près.

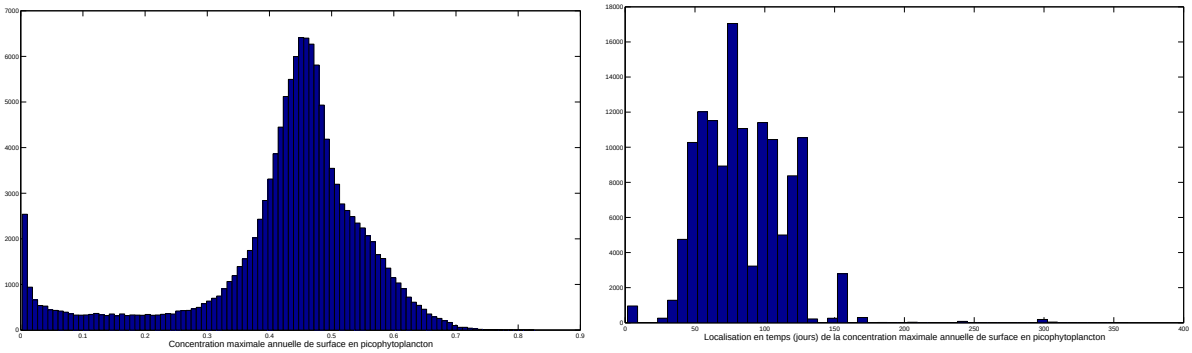


FIGURE 8.17 – Histogrammes représentant les valeurs (concentration en azote) (à gauche) et la localisation (jours) (à droite) du maximum annuel en surface de concentration en picophytoplancton.

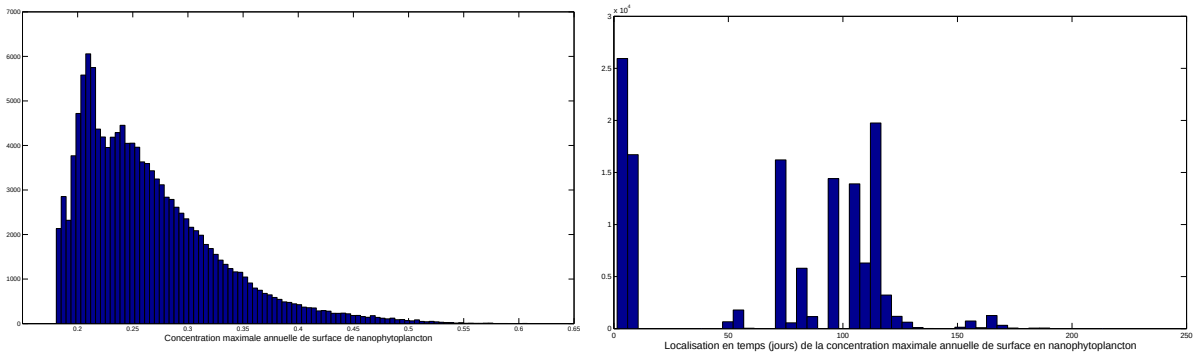


FIGURE 8.18 – Histogrammes représentant les valeurs (concentration en azote) (à gauche) et la localisation (jours) (à droite) du maximum annuel en surface de concentration en nanophytoplancton.

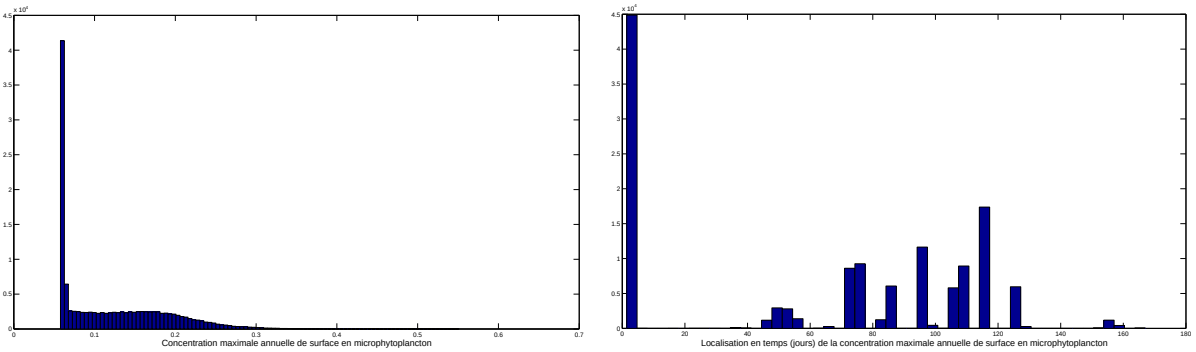


FIGURE 8.19 – Histogrammes représentant les valeurs (concentration en azote) (à gauche) et la localisation (jours) (à droite) du maximum annuel en surface de concentration en mésophytoplancton.

Dans un second temps, nous estimons les indices de Sobol' des 3 sorties : maximum annuel en surface, maximum annuel en profondeur (entre -20m et -50m) et moyenne annuelle de la concentration en chlorophylle-a. Les deux premières sont présentées dans le cadre du Chapitre 6, où nous mettons en œuvre la méthode de Sobol' conjuguée aux hypercubes latins répliqués, et nous renvoyons le lecteur à la Section 6.5.2 pour les détails. Les conclusions quant à l'estimation elle-même est que les différents estimateurs des indices d'ordre 1 et 2 convergent clairement et que les erreurs d'estimations calculées analytiquement sont de l'ordre de $+/-1\%$ sur chacun des indices — pour des échantillons de taille 65536 et 66049. Même avec cette précision plutôt grossière, les estimations réalisées ne laissent

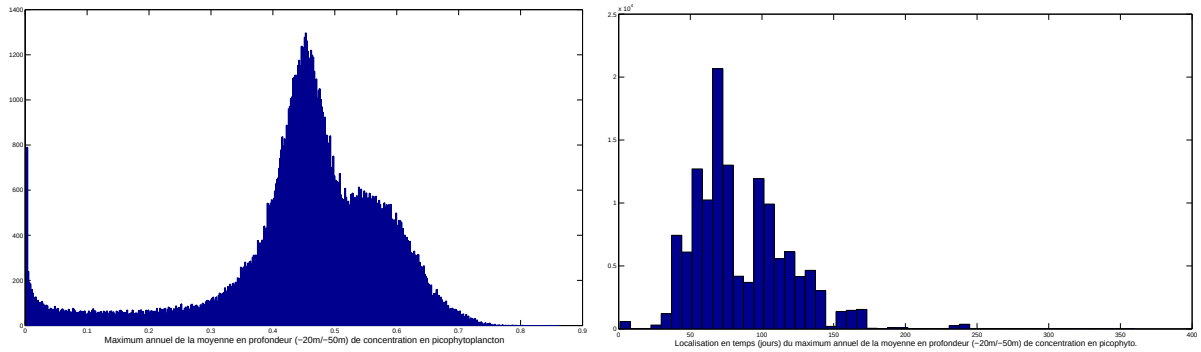


FIGURE 8.20 – Histogrammes représentant les valeurs (concentration en azote) (à gauche) et la localisation (jours) (à droite) du maximum annuel de concentration moyenne en picophytoplancton entre -20m et -50m.

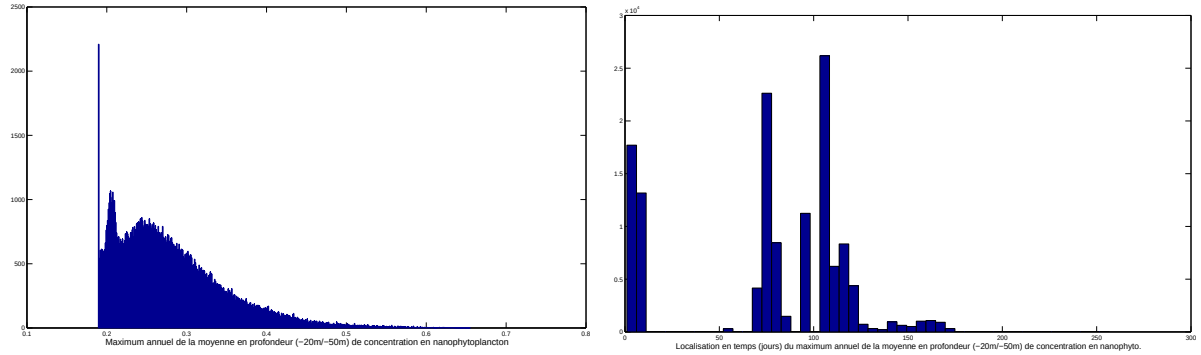


FIGURE 8.21 – Histogrammes représentant les valeurs (concentration en azote) (à gauche) et la localisation (jours) (à droite) du maximum annuel de concentration moyenne en nanophytoplancton entre -20m et -50m.

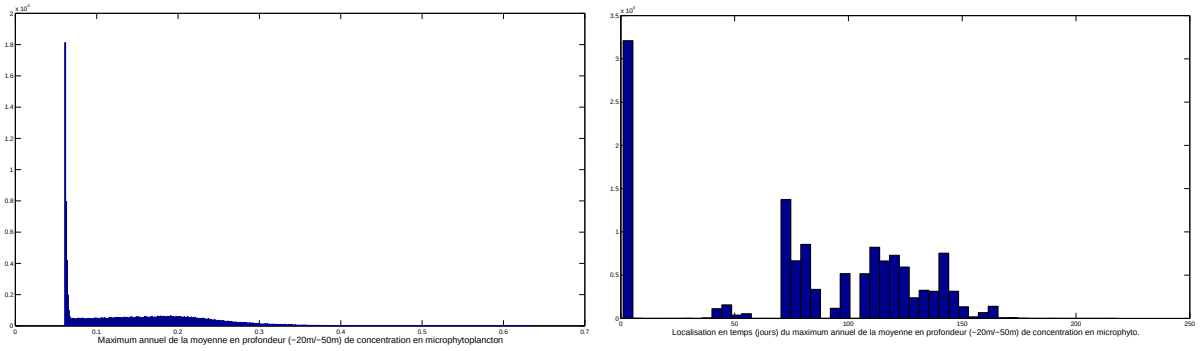


FIGURE 8.22 – Histogrammes représentant les valeurs (concentration en azote) (à gauche) et la localisation (jours) (à droite) du maximum annuel de concentration moyenne en mésophytoplancton entre -20m et -50m.

absolument aucun doute sur le rôle prépondérant du paramètre *mumaxpp*. Notons en particulier que la somme de son indice d'ordre 1, et des indices d'ordre 2 qui lui sont relatifs, atteint près de 60% pour les deux sorties considérées. Se pose alors la question de savoir si ce déséquilibre a déjà été identifié dans ce modèle, et quelles peuvent être ses conséquences.

La dernière sortie (moyenne annuelle) est étudiée en calculant les indices de Sobol' par quasi-régression suivant une base de polynômes de Laguerre généralisés qui forment un système orthonormé

relativement aux distributions gamma utilisées en entrée. L'analyse va montrer que cette sortie est essentiellement additive — i.e. près de 90% de la variance est expliquée par les effets d'ordre 1 — et nous nous contentons donc de calculer les indices de Sobol' d'ordre 1. Ceux-ci sont estimés à l'aide des 20 premières harmoniques, ayant noté que ces dernières sont déjà négligeables au-delà de la 15ème. Les résultats sont présentés à la Figure 8.23.

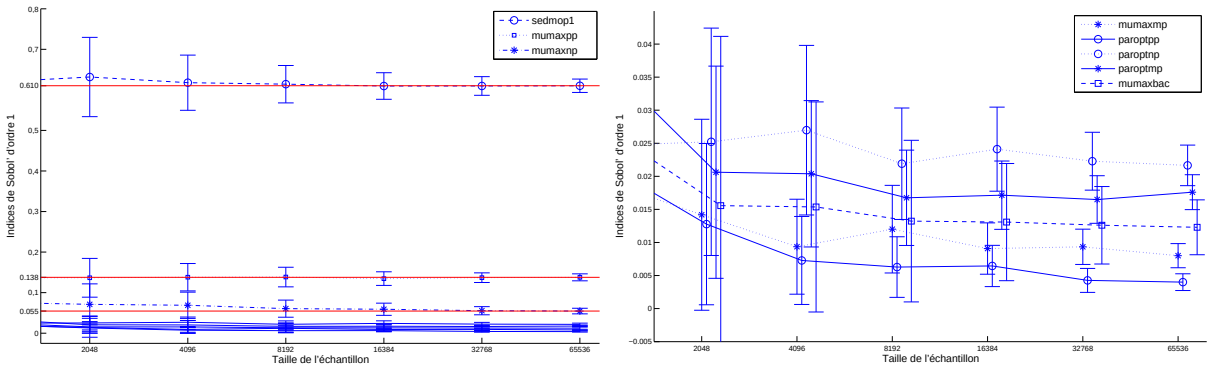


FIGURE 8.23 – Tracé des indices de Sobol' d'ordre 1 en fonction de la taille de l'échantillon (agrandissement à droite). Sortie : moyenne annuelle de la concentration de chlorophylle-a en surface.

On note d'abord que les barres d'erreurs obtenues analytiquement sont comparables à celles dérivées dans la méthode de Sobol' appliquée à la Section 6.5.2. Comme mentionné ci-avant, on remarque que la somme des indices de Sobol' d'ordre 1 atteint près de 90% démontrant que cette sortie du modèle est essentiellement la somme de fonctions d'une seule variable. Le plus frappant reste néanmoins le fait que l'indice de Sobol' du paramètre *sedmop*1 dépasse les 60%. Dans les deux sorties étudiées précédemment, les deux paramètres incertains ajoutés, *sedmop*1 et *mumaxbac*, n'avaient pas fondamentalement bouleversé les analyses précédentes, mais dans le cas de la moyenne annuelle de la concentration de chlorophylle-a en surface, le constat est tout autre. Une tentative d'explication simpliste de ce phénomène peut être de penser qu'une vitesse de sédimentation excessive va réduire le processus de dissolution MOP1→NOD qui permet via le processus d'assimilation puis d'excrétion des bactéries de regénérer des nutriments ($NH4$). Par suite, les $NH4$ et $NO3$ présents initialement sont consommés jusqu'à épuisement, bloquant ainsi la croissance du phytoplancton. Cette hypothèse est corroborée par le graphique de l'effet principal dû à sedmop1 (voir Figure 8.24) qui montre que l'augmentation de sa valeur conduit à une diminution de la sortie étudiée. On trace en outre les effets principaux — i.e. les fonctions d'une variable dans la décomposition ANOVA — des trois paramètres prépondérants (voir Figures 8.24 à 8.26). On peut alors s'interroger sur la courbe de l'effet principal de mumaxpp. On aurait pu croire que l'augmentation de sa valeur conduirait à une augmentation de la concentration en picophytoplancton et donc à une augmentation de la sortie étudiée. Mais la courbe présentée montre exactement le contraire. Est-ce qu'un taux de croissance maximal trop élevé prive les autres phytoplanctons de nutriments et fait ainsi diminuer la concentration en chlorophylle-a ?

Enfin, nous terminons l'étude dans cette seconde phase en évaluant une borne pour les indices de Sobol' ascendants des paramètres *paroptmp* et *mumaxbac* pour les trois sorties étudiées en calculant des indices basés sur les dérivées (voir Chapitre 3), ceux-ci apparaissant comme inactifs dans toutes les analyses faites jusqu'à présent. Ces calculs se font via un schéma aux différences finies en considérant un incrément de 10^{-5}, et un échantillon en hypercube latin respectant les distributions marginales suivant les lois gamma, de taille 65536. Les majorations obtenues sont

$$S_{paroptpnp} \leq 0.15 \quad \text{et} \quad S_{mumaxbac} \leq 4.6$$

pour la sortie : maximum annuel de la concentration moyenne de chlorophylle-a entre -20m et -50m,

$$S_{paroptpnp} \leq 0.19 \quad \text{et} \quad S_{mumaxbac} \leq 6.2$$

pour la sortie : maximum annuel de la concentration de chlorophylle-a en surface, et

$$S_{paroptpnp} \leq 0.38 \quad \text{et} \quad S_{mumaxbac} \leq 450$$

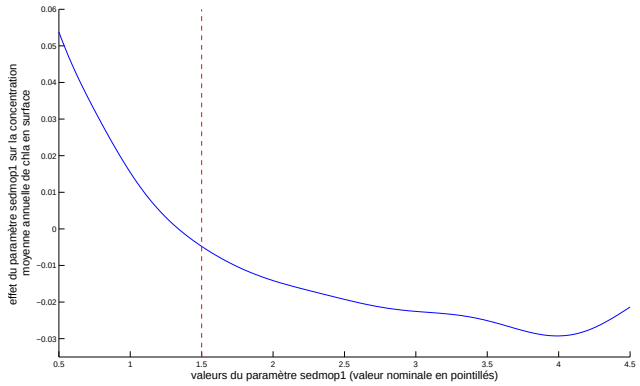


FIGURE 8.24 – Effet principal du paramètre *sedmop*1. Valeur de la sortie du modèle en fonction du paramètre *sedmop*1, et en moyenne par rapport à tous les autres paramètres.

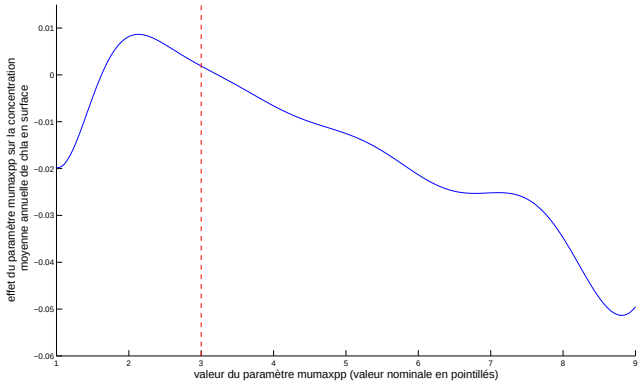


FIGURE 8.25 – Effet principal du paramètre *mumaxpp*.Valeur de la sortie du modèle en fonction du paramètre *mumaxpp*, et en moyenne par rapport à tous les autres paramètres.

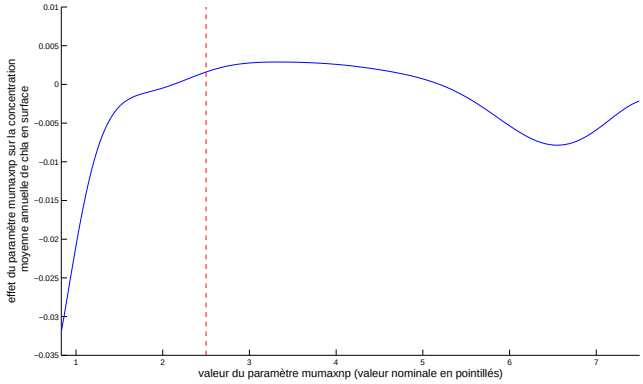


FIGURE 8.26 – Effet principal du paramètre *mumaxnp*. Valeur de la sortie du modèle en fonction du paramètre *mumaxnp*, et en moyenne par rapport à tous les autres paramètres.

pour la sortie : moyenne annuelle de la concentration de chlorophylle-a en surface. Elles ne présentent malheureusement que peu d'intérêt au sens où l'indice de Sobol' ascendant est essentiellement utilisé pour trier les paramètres inactifs dont les indices sont petits, typiquement inférieurs à 0.01. Les bornes obtenues ici étant bien au-delà, nous sommes pessimistes quant à l'utilisation de cette approche sur des modèles non-analytiques de grande complexité.

Remarque 8.1. *Le fait que les méthodes basées sur les dérivées ne soient pas adaptées à la quantification des différentes parts de variance d'une fonction, tient par exemple au fait qu'elles ne sont pas capables d'identifier un bruit fortement oscillatoire. En effet, considérons la variable aléatoire*

$$Y = \frac{1}{k}\sin(kX), \quad , k \in \mathbb{N}^*$$

où X est uniformément distribuée sur $[0, 1]$. Alors on peut constater que la variance de Y tend vers 0 lorsque k tend vers $+\infty$ — en effet, on a $Var[Y] = 1/k^2$ — alors que la distribution de sa dérivée reste inchangée quelque soit k.

8.3 Conclusions et perspectives

Au cours de cette étude, nous avons pu mettre en œuvre un certain nombre de méthodes d'analyse de sensibilité basées sur les dérivées et sur la décomposition ANOVA. Pour les approches basées sur les dérivées, une fois l'hypothèse d'effets linéaires additifs écartés (voir Figure 8.15), l'information obtenue est plutôt pauvre et difficilement interprétable (voir Figure 8.16). De plus, les derniers développements théoriques visant à majorer les indices de Sobol' ascendants d'ordre 1 par les indices de sensibilité basés sur les dérivées [LIPG12] ne permettent aucunement de conclure sur la négligeabilité de certains paramètres dans notre cas. Concernant l'approche basée sur la décomposition ANOVA, l'information qu'elle fournit est extrêmement riche et semble fiable pour la plupart des méthodes appliquées. Néanmoins, la méthode RBD ne donne pas entière satisfaction et devra être appliquée par la suite sur de plus grands échantillons, afin de réduire les problèmes liés au biais. Pour la suite, l'étude de sous-modèles prenant en compte plus de paramètres incertains ne paraît présenter aucune difficulté si ce n'est la nécessité de disposer d'une ressource en calcul plus importante (machines parallèles par exemple), et peut être d'une plus grande expertise pour fixer les distributions a priori des paramètres considérés comme incertains.

Conclusions et perspectives

Conclusions et perspectives

Conclusions

Dans cette thèse, nous nous sommes principalement intéressés à la décomposition ANOVA qui constitue un objet mathématiquement rigoureux et pertinent pour l'analyse de sensibilité globale, et l'objectif principal a été d'analyser le problème de l'estimation des indices de Sobol' — qui se définissent naturellement à partir de la décomposition ANOVA — au seul regard de la théorie de l'intégration numérique. Dans cette démarche, en prenant appui sur des théories classiques et des développements plus récents, notre travail a permis d'établir rigoureusement les bases de méthodes existantes, de les développer, ainsi que de proposer de nouvelles pistes de réflexion dans le but de les optimiser.

Méthode FAST. En nous appuyant sur des théories existantes liées à l'intégration numérique, nous avons explicité le fonctionnement de la méthode FAST qui avait été introduite initialement par Cukier et al. [CFS+73]. Nous expliquons en effet que cette méthode qui reposait sur une approximation du théorème ergodique de Weyl [Wey38] doit être appréhendée au regard de la théorie de la dualité sur les sous-groupes cycliques du tore unité. Ce point de vue permet d'expliciter l'erreur commise lors de l'estimation des indices de Sobol' en s'appuyant sur la formule de Poisson généralisée (voir, e.g., [Bou67]). Comme conséquence directe, et en se référant à la théorie de l'intégration numérique relative aux sous-groupes finis du tore unité — désignée en anglais par l'expression *lattice rule* — on peut montrer que les estimateurs des parts de variances partielles σ^2_{u} et de la variance totale σ^2 peuvent atteindre une vitesse de convergence en $O(n^{-\alpha})$ avec $\alpha > 1/2$, i.e. supérieure à celle de la méthode de Monte Carlo. Néanmoins, ce résultat n'est théoriquement vérifiable que dans des espaces de fonctions dont le spectre de Fourier décroît suffisamment vite à l'infini — i.e. comme on le rencontre souvent dans la littérature des espaces de Korobov pondérés ou des espaces de Sobolev pondérés.

De manière pratique, la méthode FAST suppose de disposer d'un "bon" générateur du groupe cyclique sur lequel est effectuée l'intégration numérique. Le critère original permettant de construire un tel générateur a été introduit par Schaibly et Shuler [SS73]. Nous l'avons affaibli en nous basant sur la formule sommatoire de Poisson généralisée, et comme nous le détaillons dans l'Annexe A, ce critère est un cas particulier du critère destiné à se prémunir des *confusions d'effets* lors d'une analyse de la variance effectuée sur un plan factoriel construit à partir d'un morphisme de groupe. Néanmoins ce nouveau critère, comme le critère original, du fait d'un coût algorithmique trop grand — il s'agit d'une recherche exhaustive — ne permet de construire que des générateurs faiblement robustes aux interférences et uniquement en petite dimension. Pour contourner ce problème, nous proposons d'utiliser les algorithmes "composante-par-composante" récemment introduits par Nuyens [Nuy07] et ses co-auteurs [CKN06, CKN10]. Ceci permet de lever la difficulté algorithmique liée à la construction des générateur des groupes cycliques, mais ne change rien sur le fait que cette méthode FAST ne peut être efficace que pour des fonctions bien particulières possédant un spectre de Fourier à décroissance rapide.

Méthode RBD. La méthode RBD, initialement introduite par Tarantola et al. [TGM06], repose essentiellement sur les mêmes bases que la méthode FAST, et donc sur l'approximation originale du théorème de Weyl. Par conséquent, elle souffre tout comme FAST d'une faible connaissance théorique ; ces estimateurs possèdent en particulier un biais non négligeable pour des échantillons réduits. Dans un premier temps, à l'aide d'un argumentaire encore insuffisant pour comprendre la méthode RBD de

manière totalement rigoureuse, nous avons proposé une correction de biais de ces estimateurs. Dans un second temps, en nous appuyant sur la nouvelle manière d'appréhender la méthode FAST que nous avons introduite, et également en se référant à des notions classiques d'intégration numérique, nous explicitons le fonctionnement de la méthode RBD, et la traduisons comme une quasi-régression [AO01] suivant la base des polynômes trigonométriques, relativement à un tableau orthogonal de force 1 [HSS99]. Comme noté récemment par Xu et Gertner [XG11b], on remarque alors que la méthode RBD, qui avait été introduite initialement pour estimer les indices de Sobol' élémentaires d'ordre 1, peut être utilisée pour estimer les indices de Sobol' élémentaires d'ordres plus élevés. En outre, nous généralisons son application à la quasi-régression relativement à des tableaux orthogonaux de force supérieure à 1 — dans la pratique, de force 2 avec la construction de Bose [Bos38].

Au regard de cette nouvelle introduction de la méthode RBD, il est dès lors possible d'expliciter le biais de ses estimateurs. Nous le faisons en nous appuyant sur les travaux d'Owen [Owe94] relatifs aux tableaux orthogonaux randomisés — i.e. des tableaux orthogonaux dont on permute aléatoirement les niveaux. En revisitant l'une des formules d'Owen en terme d'analyse harmonique, il nous est possible d'expliciter les biais de RBD et de proposer des méthodes de corrections pour les différents estimateurs. Néanmoins, comme pour la méthode FAST, ces résultats théoriques de correction de biais ne sont performants que pour des fonctions possédant un spectre de Fourier à décroissance rapide.

Méthode RBD-FAST. Avant l'étude théorique approfondie sur les méthodes RBD et FAST effectuée au Chapitre 5, nous avons proposé une technique pour estimer tous les indices d'ordres 1 et 2 d'une fonction donnée à l'aide de RBD-FAST, ainsi qu'une correction de biais de ces estimateurs. Cette méthode souffre comme FAST et RBD d'une faible compréhension théorique. Au regard de notre travail consistant à revisiter rigoureusement FAST et RBD, on peut voir RBD-FAST comme une quasi-régression suivant des polynômes trigonométriques relativement à un supercube latin (construit à partir d'un sous-groupe cyclique) — pour les supercubes latins, voir [Owe98]. Dans la pratique, il est difficile de voir quels avantages on peut tirer de ce type de plans d'expériences extrêmement complexes.

Méthode de Sobol' La méthode de Sobol' bénéficie d'un avantage inestimable face à toutes les autres existantes : ses hypothèses d'applications ne contiennent aucune contrainte relative à la régularité de la fonction étudiée, ou à sa structure spectrale relativement à une base particulière. Chacun de ses estimateurs bénéficie en outre d'une vitesse de convergence indépendante de la dimension d de la fonction étudiée. Malheureusement sur ce dernier point, l'estimation de tous les indices de Sobol' descendants d'ordre 1, ou de tous ceux d'ordres 1 et 2, requiert un nombre d'échantillons distincts dépendant linéairement de la dimension d ; et par conséquent la vitesse de convergence se dégrade avec l'augmentation de la dimension. Pour pallier cet inconvénient, nous avons introduit une nouvelle manière d'utiliser cette méthode à l'aide d'hypercubes latins répliqués [McK95]. Cette nouvelle méthode permet, pour tout $k \leq d$, d'estimer tous les indices de Sobol' descendants d'ordre k à l'aide de seulement 2 échantillons. Cette technique est complétée par des résultats théoriques pour le cas $k = 1$; pour les indices d'ordre 1, l'estimateur ainsi introduit est fortement consistant, normalement asymptotique et possède un biais en $O(n^{-1})$ dont nous explicitons un majorant de la constante.

Conclusion générale Dans le paysage de l'analyse de sensibilité globale, la décomposition ANOVA nous apparaît comme l'objet mathématique le mieux posé et le plus pertinent au regard des méthodes de screening et des méthodes basées sur les dérivées. Et par conséquent, nous nous sommes uniquement attachés à développer des aspects de l'estimation des indices de Sobol'. De plus, sur des modèles tels que MODECOGeL dont le coût d'une simulation est inférieur à 5 secondes, il devient possible à l'aide de machines de calcul parallèle d'envisager des analyses dépassant le million de simulations. Dans ce cadre, la méthode de Sobol' telle que nous l'avons introduite — i.e. indépendante de la dimension — peut devenir un outil puissant.

Concernant l'intégration des dérivées en analyse de sensibilité globale, nous constatons qu'elles permettent clairement, d'un point de vue qualitatif, d'identifier si un effet est linéaire ou non. A contrario, nous remarquons que les indices de sensibilité basés sur les dérivées ne permettent pas de quantifier clairement l'influence des différentes variables en terme de variance. En effet, nous avons

remarqué que les bornes des indices de sensibilités ascendants d'ordre 1 fournis par les indices basés sur la variance ne permettaient de tirer aucune conclusion. La raison de ce constat tient essentiellement au fait que, contrairement aux indices basés sur la variance, les indices basés sur les dérivées ne sont pas capables d'identifier un bruit fortement oscillatoire comme nous l'avons noté dans la Remarque 8.1.

Perspectives

Au regard de notre travail, on peut envisager des perspectives à la fois d'ordre théorique et d'ordre pratique.

D'abord, il nous paraît important d'analyser en détail l'article de Loh [Loh08] énonçant les propriétés asymptotiques de l'intégration numérique relativement à des hypercubes latins basés sur des tableaux orthogonaux de force 2. Le théorème central limit (TCL) qu'il énonce est en effet conditionné à des hypothèses de régularité de la fonction étudiée, alors que son analogue pour les hypercubes latins ne requiert qu'une simple hypothèse d'intégrabilité. Il est donc nécessaire de comprendre pourquoi la stratification à l'ordre 2 restreint le champ d'application du TCL, et éventuellement d'élargir ce champ si cela est possible. En particulier, ce travail en amont permettra de préciser les propriétés asymptotiques de la méthode de Sobol' construite avec les hypercubes latins répliqués basés sur des tableaux orthogonaux de force 2 (voir Chapitre 6).

Concernant la méthode de Sobol' construite avec les hypercubes latins répliqués, il reste une question ouverte sur l'estimateur de la variance asymptotique. En effet, dans le Chapitre 6, nous nous contentons d'utiliser l'estimateur introduit dans la méthode de Sobol' classique, mais nous savons qu'il n'est pas optimal et produit donc un intervalle de confiance plus large que l'intervalle théorique.

Nous avons constaté que la Delta méthode s'applique naturellement aux estimateurs des indices de Sobol' basés sur une décomposition ANOVA spectrale. En conséquence, il est possible d'avoir asymptotiquement une estimation de la variance de chacun des coefficients spectraux. Par suite, il devient possible d'envisager une méthode de seuillage afin de réduire la variance des estimateurs des indices de Sobol' spectraux. Un point de départ éventuel pour ce travail est l'article de Jiang et Owen [JO03]. Notons que la sélection automatique de coefficients spectraux dans le cadre du calcul des indices de Sobol' existe déjà ; citons par exemple la méthode de régression adaptative de Blatman et Sudret [BS10].

D'un point de vue applicatif, la perspective principale consiste à notre avis à généraliser l'utilisation de l'analyse de la variance, via l'estimation des indices de Sobol' par la méthode de Monte Carlo que nous avons développée au Chapitre 6, à des modèles de grande dimension — i.e. supérieure à 30 — en s'appuyant sur une puissance de calcul importante. Comme nous l'avons noté précédemment, MODECOGeL comme de nombreux autres simulateurs d'écosystèmes, de pêcherie ou de chimie, ont un coût de calcul très faible qui laisse entrevoir la possibilité de mener des intégrations numériques prenant en compte des échantillons dont la taille peut excéder le million.

Annexes

Annexe A

Lien entre les plans d'expérience de FAST et les plans factoriels fractionnaires

Notations

Soit $\mathcal{E}$ un ensemble. On note classiquement l'indicatrice

$$\mathbf{1}_{\mathcal{E}}(e) = \begin{cases} 1 & \text{si } e \in \mathcal{E} \\ 0 & \text{sinon.} \end{cases}$$

Si de plus $\mathcal{E}$ est fini,

$$|\mathcal{E}| \quad \text{et} \quad \mathcal{P}(\mathcal{E})$$

désignent respectivement le cardinal et l'ensemble des parties de $\mathcal{E}$. Enfin, on note $\overline{z}$ le nombre complexe conjugué de z, $p \wedge q$ le pgcd des entiers p et q, et $\{x\}$ la partie fractionnaire du réel x. Dans ce qui suit, $\mathbb{C}^\times$ désigne classiquement le groupe des éléments inversibles de $\mathbb{C}$.

A.1 Plans factoriels, modèle linéaire et ANOVA

1.1. Notons $\mathcal{U}$ l'ensemble des *unités expérimentales* et $\mathcal{T}$ l'ensemble des *traitements* — également appelé *domaine expérimental*. Considérons l'application

$$\begin{array}{rccc} d: & \mathcal{U} & \longrightarrow & \mathcal{T} \\ & u & \longmapsto & t = d(u) \end{array}$$

qui associe à toute unité $u \in \mathcal{U}$ le traitement $d(u) \in \mathcal{T}$ effectué sur celle-ci. On suppose que $\mathcal{U}$ est fini, on note $|\mathcal{U}| = n$. On suppose, par ailleurs, que l'ensemble $\mathcal{T}$ se décompose comme un produit cartésien fini $\mathcal{T} = \mathcal{T}_1 \times \cdots \times \mathcal{T}_m$ et on note $\mathfrak{m} = \{1, \ldots, m\}$. Chaque traitement t s'écrit donc comme un m-uplet $(t_1, \ldots, t_m)$ dans lequel t_i est le *niveau* du i-ème *facteur*. L'application d est appelée *plan factoriel*. Lorsque les $\mathcal{T}_i$ sont finis, si les $\mathcal{T}_i$ ont tous même cardinal, le plan factoriel est dit *symétrique*, et *asymétrique* sinon. Si l'application d est injective, le plan est dit *régulier*, et si elle est surjective, le plan est dit *complet*.

1.2. La réponse sur l'unité expérimentale $u \in \mathcal{U}$ est modélisée par une variable aléatoire (v.a.) notée Y_u dont la valeur observée après l'expérience est notée y_u. On suppose que la v.a. Y_u se décompose comme suit

$$Y_u = \tau\big(d(u)\big) + Z_u \tag{A.1}$$

où $\tau\big(d(u)\big)$ est un terme dépendant du traitement $d(u)$ appliqué à l'unité u — τ étant une fonction réelle définie sur $\mathcal{T}$ — et Z_u est une variable aléatoire dépendant de l'unité u. $\tau\big(d(u)\big)$ décrit l'effet *déterministe* du traitement $d(u) \in \mathcal{T}$ sur la sortie, tandis que Z_u permet d'intégrer, de manière additive, l'effet *aléatoire* de l'unité $u \in \mathcal{U}$, elle-même, dû à la variabilité de l'expérimentation.

1.3. Dans le cas où les $\mathcal{T}_i$ sont **finis**, il est aisé de munir chacun d'entre eux d'une structure d'espace probabilisé à l'aide de la tribu triviale $\mathcal{P}(\mathcal{T}_i)$ et d'une mesure de probabilité sur cette dernière, notée μ_i. Pour chaque $1 \leq i \leq m$, considérons alors $\mathbb{R}^{\mathcal{T}_i}$ l'espace vectoriel des fonctions réelles définies sur $\mathcal{T}_i$ [1], ainsi que ses deux sous-espaces vectoriels $\Theta_{\mu_i,0}$ et $\Theta_{\mu_i,1}$ des fonctions, respectivement, constantes et de moyennes nulles par rapport à μ_i. Puis, en notant μ la mesure produit des μ_i, $1 \leq i \leq m$, on considère l'espace probabilisé $(\mathcal{T}, \mathcal{P}(\mathcal{T}), \mu)$ ainsi que $\mathbb{R}^{\mathcal{T}}$ l'espace vectoriel des fonctions réelles définies sur $\mathcal{T}$. Enfin, en identifiant l'espace vectoriel $\mathbb{R}^{\mathcal{T}}$ au produit tensoriel $\otimes_{i=1}^{m}\mathbb{R}^{\mathcal{T}_i}$ (voir Corollaire 1. p84 dans [CO68]), on introduit pour tout $\mathfrak{u} \subset \mathfrak{m}$, le sous-espace vectoriel de $\mathbb{R}^{\mathcal{T}}$,

$$\Theta_{\mu,\mathfrak{u}} = \bigotimes_{i=1}^{m} \Theta_{\mu_i,\mathbf{1}_{\mathfrak{u}}(i)} \ .$$

1.4. En définissant le produit scalaire sur $\mathbb{R}^{\mathcal{T}}$

$$< f, g >_{\mu} = \sum_{t \in \mathcal{T}} f(t)g(t)\mu(t) \ ,$$

on obtient la décomposition orthogonale de $\mathbb{R}^{\mathcal{T}}$ en somme directe de sous-espaces vectoriels de "complexité" croissante, sous la forme du

Théorème A.1. *L'espace $\mathbb{R}^{\mathcal{T}}$ vérifie*

$$\mathbb{R}^{\mathcal{T}} = \bigoplus_{\mathfrak{u} \subset \mathfrak{m}} \Theta_{\mu,\mathfrak{u}} \tag{A.2}$$

En outre, la décomposition est orthogonale au sens où, pour tous sous-ensembles distincts $\mathfrak{u}$ et $\mathfrak{v}$ de $\mathfrak{m}$,

$$\forall f \in \Theta_{\mu,\mathfrak{u}},\ \forall g \in \Theta_{\mu,\mathfrak{v}},\quad < f, g >_{\mu} = 0. \tag{A.3}$$

Démonstration. On vérifie aisément qur $\mathbb{R}^{\mathcal{T}_i} = \Theta_{\mu_i,0} \oplus \Theta_{\mu_i,1}$, et on conclut à la Formule (A.2) par un résultat élémentaire relatif aux produits tensoriels :

$$\bigotimes_{i=1}^{m} \mathbb{R}^{\mathcal{T}_i} \simeq \bigoplus_{\mathfrak{u} \subset \mathfrak{m}} \Theta_{\mu,\mathfrak{u}} \qquad \text{(voir Théorème 8.10. p94–95 dans [CO68])}. \tag{A.4}$$

La preuve de la Formule (A.3) est essentiellement calculatoire ; on la trouvera dans [Col70] (voir Démonstration du Théorème 3.2. p43). □

En notant,

$$\tau = \sum_{\mathfrak{u} \subset \mathfrak{m}} \tau_{\mathfrak{u}}$$

la décomposition de la fonction τ sur la somme directe des $\Theta_{\mu,\mathfrak{u}}$, on a aisément la décomposition de la variance

$$\mathrm{Var}_{\mu}[\tau] = \sum_{\emptyset \neq \mathfrak{u} \subset \mathfrak{m}} \mathrm{Var}_{\mu}[\tau_{\mathfrak{u}}] \ .$$

Les composantes $\tau_{\{i\}}$ sont appelées *effets principaux* de τ par rapport aux variables t_i, et les $\tau_{\mathfrak{u}}$, avec $|\mathfrak{u}| > 1$, sont qualifiées d'*interactions d'ordre* $|\mathfrak{u}|$.

1. Les fonctions de cet espace sont toutes, trivialement, mesurables par rapport à la tribu $\mathcal{P}(\mathcal{T}_i)$ et intégrables par rapport à μ_i. Lorsque $\mathcal{T}_i$ n'est pas un ensemble fini, on ne peut éluder ces questions de mesurabilité et d'intégrabilité (voir §**1.5.**).

1.5. (Remarques)

1.5.1. Dans le cadre des plans factoriels, la décomposition orthogonale précédente est généralement exploitée sous l'hypothèse que les μ_i sont des mesures d'équiprobabilité, d'où

$$\forall t \in \mathcal{T},\ \mu(t) = \frac{1}{|\mathcal{T}|}.$$

1.5.2. La décomposition de la Formule (A.2) est généralisable à des ensembles $\mathcal{T}_i$ qui ne sont pas finis, typiquement $\mathbb{R}$. En effet, il suffit de munir chacun des $\mathcal{T}_i$ de sa tribu borélienne et d'une mesure de probabilité μ_i sur cette dernière ; puis de considérer les espaces de fonctions intégrables $L^p(\mathcal{T}_i, \mu_i)$ et $L^p(\mathcal{T}, \mu)$ — avec $p \in [1, +\infty[$ — en lieu et place des $\mathbb{R}^{\mathcal{T}_i}$ et $\mathbb{R}^{\mathcal{T}}$. Dans ce cas, l'identification entre $L^p(\mathcal{T}, \mu)$ et $\otimes_{i=1}^m L^p(\mathcal{T}_i, \mu_i)$ est possible (voir Théorème VI-20 dans [Vil08] et Corollaire I-55 pour les hypothèses d'application du Théorème VI-20), les définitions des $\Theta_{\mu_i,0}$ et $\Theta_{\mu_i,1}$ sont les mêmes et la Formule (A.4) reste valide.

1.5.2. a) Pour $p = 1$ et des mesures μ_i uniformes sur $[0, 1]$, on retrouve la décomposition L^1 donnée par Sobol' en 1990 [Sob93].

1.5.2. b) Pour $p = 2$, la Formule (A.3) du Théorème A.1 est valide, et on obtient une décomposition orthogonale pour le produit scalaire

$$< f, g >_\mu = \int_{\mathcal{T}} f(t)g(t)\mu(t)dt\ .$$

Pour la suite, sauf mention contraire, les $\mathcal{T}_i$ sont supposés **finis** et les μ_i sont des mesures **d'équiprobabilité**.

A.2 Effets factoriels

Dorénavant $< f, g >_\mu$ désigne le produit scalaire hermitien

$$< f, g >_\mu = \sum_{t \in \mathcal{T}} f(t)\overline{g(t)}\mu(t)\ .$$

2.1. Chacun des $\mathcal{T}_i$ est dorénavant muni d'une structure de groupe cyclique d'ordre $q_i = |\mathcal{T}_i|$,

$$\mathcal{T}_i \simeq \mathbb{Z}/q_i\mathbb{Z}$$

et par conséquent, $\mathcal{T}$ est muni d'une structure de groupe abélien fini, produit des $\mathbb{Z}/q_i\mathbb{Z}$,

$$\mathcal{T} \simeq \mathbb{Z}/q_1\mathbb{Z} \times \cdots \times \mathbb{Z}/q_m\mathbb{Z}\ . \tag{A.5}$$

Notons $\mathcal{T}^* = \mathrm{Hom}(\mathcal{T}, \mathbb{C}^\times)$ l'ensemble des homomorphismes du groupe $\mathcal{T}$ dans le groupe multiplicatif des nombres complexes $\mathbb{C}^\times$. Pour tous χ et χ' dans $\mathcal{T}^*$, on définit le produit $\chi\chi'$, l'inverse χ et le neutre $e_{\mathcal{T}^*}$, respectivement comme les homomorphismes qui à tout élément t de $\mathcal{T}$, associent $\chi(t)\chi'(t)$, $\big(\chi(t)\big)^{-1}$ et 1. Par suite, $\mathcal{T}^*$ possède trivialement une structure de groupe ; c'est le *groupe dual* de $\mathcal{T}$ et ses éléments, sont appelés *caractères* de $\mathcal{T}$.

2.2. L'ensemble des caractères de $\mathcal{T}$ vérifie le

Théorème A.2. *$\mathcal{T}^*$ forme une base orthonormale de $\mathbb{C}^{\mathcal{T}}$.*

Démonstration. Notons d'abord que $\mathcal{T}$ est d'exposant [2] fini et qu'il est égal à $q = \text{ppcm}(q_1, \ldots, q_m)$, et que les caractères de $\mathcal{T}$ prennent leurs valeurs dans $\mathbb{U}_q$, le sous-groupe de $\mathbb{C}^\times$ constitué des racines q-ièmes de l'unité. Par suite, on a

$$\chi^{-1}(g) = \big(\chi(g)\big)^{-1} = \overline{\chi(g)}\ .$$

Puis, pour χ, $\chi' \in \mathcal{T}^*$ il vient

$$< \chi, \chi' >_\mu = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \chi(t)\overline{\chi'(t)} = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \chi\chi'^{-1}(t),$$

et on obtient aisément que les caractères forment une famille orthonormale — et par conséquent libre — de $\mathbb{C}^{\mathcal{T}}$ en notant que pour tout caractère χ non trivial, on a

$$\sum_{t \in \mathcal{T}} \chi(t) = 0 \tag{A.6}$$

et que pour l'élément neutre $e_{\mathcal{T}^*}$, on a

$$\sum_{t \in \mathcal{T}} e_{\mathcal{T}^*}(t) = |\mathcal{T}|.$$

On conclut en remarquant que le cardinal de cette famille de caractères est égal à la dimension de l'espace $\mathbb{C}^{\mathcal{T}}$ par la formule suivante,

$$\sum_{\chi \in \mathcal{T}^*} \chi(e_{\mathcal{T}^*}) = |\mathcal{T}|, \tag{A.7}$$

où $e_{\mathcal{T}^*}$ est l'élément neutre de $\mathcal{T}$. Pour les formules (A.6–A.7), voir Théorème (somme de caractères) p65 dans [Fre01]. □

Cette propriété peut être rendue plus explicite en exhibant un isomorphisme — non canonique — entre le groupe $\mathcal{T}$ et son groupe dual $\mathcal{T}^*$. Pour cela, on définit pour chaque $1 \leq i \leq m$,

$$\begin{array}{rccc} \chi_i : & \mathcal{T} & \longrightarrow & \mathbb{C}^\times \\ & (t_1, \ldots, t_m) & \longmapsto & \mathrm{e}^{2i\pi t_i/q_i} \end{array} \tag{A.8}$$

où, suivant la Formule (A.5), les t_j sont dans $\mathbb{Z}/q_j\mathbb{Z}$ [3], et on a la

Proposition A.1. *L'application*

$$\begin{array}{rccc} \Psi : & \mathcal{T} & \longrightarrow & \mathcal{T}^* \\ & (t_1, \ldots, t_m) & \longmapsto & \chi_1^{t_1} \cdots \chi_m^{t_m} \end{array} \tag{A.9}$$

est un isomorphisme de groupe.

Démonstration. D'abord posons $e_1 = (1, 0, \ldots, 0)$,…,$e_m = (0, \ldots, 0, 1)$. La famille $(e_i)_{1 \leq i \leq m}$ constitue une base de $\mathcal{T}$. Maintenant, Ψ étant clairement un homomorphisme, montrons qu'elle est à la fois injective et surjective.

Ψ est injective. Soit $(k_1, \ldots, k_m)$ un élément du noyau de Ψ i.e. pour tout $t \in \mathcal{T}$, $\chi_1^{k_1} \cdots \chi_m^{k_m}(t) = e_{\mathcal{T}^*}(t) = 1$. En particulier, pour tout $1 \leq i \leq m$, on a

$$\chi_1^{k_1} \cdots \chi_m^{k_m}(e_i) = e^{2i\pi \frac{k_i}{q_i}} = 1$$

et par suite, on a que $k_1 = \cdots = k_m = 0$, et on conclut que le noyau de Ψ est trivial et donc que Ψ est injective.

2. L'exposant d'un groupe G est le plus petit entier strictement positif n, s'il existe, tel que $\forall g \in G,\ g^n = e$, où e désigne l'élément neutre de G.

3. Ici, et par la suite également, on ne distingue pas les éléments t_j du groupe quotient $\mathbb{Z}/q_j\mathbb{Z}$ — qui sont des classes d'équivalence — de leurs représentants dans $\mathbb{Z}$.

Ψ *est surjective.* Soit χ un caractère de $\mathcal{T}$ et $1 \leq i \leq m$, on a $\chi(e_i)^{q_i} = 1$, et on en déduit qu'il existe $0 \leq p_i < q_i$ tel que $\chi(e_i) = \chi_i^{p_i}(e_i)$. Par suite, il vient que pour tout $t = (t_1, \ldots, t_m)$,

$$\begin{aligned}
\chi(t) = \chi(t_1 e_1 + \cdots + t_m e_m) &= \prod_{i=1}^{m} \chi^{t_i}(e_i) \\
&= \prod_{i=1}^{m} \chi_i^{p_i t_i}(e_i) \\
&= \prod_{i=1}^{m} \chi_i^{p_i}(t_i e_i) \\
&= \prod_{i=1}^{m} \chi_i^{p_i}(t) \\
&= \Psi(p_1, \ldots, p_m)
\end{aligned}$$

ce qui montre que Ψ est surjective. □

Par suite, on déduit du Théorème A.2 et de la Proposition A.1, que tout $f \in \mathbb{C}^{\mathcal{T}}$ se décompose comme

$$\begin{aligned}
f(t) &= \sum_{\chi \in \mathcal{T}^*} \widehat{f}(\chi)\chi(t) \quad , \quad t = (t_1, \ldots, t_m) \in \mathcal{T} \\
&= \sum_{k_1=0}^{q_1-1} \cdots \sum_{k_m=0}^{q_m-1} \widehat{f}(k_1, \ldots, k_m)\chi_1^{k_1}(t_1) \cdots \chi_m^{k_m}(t_m) \\
&= \sum_{k_1=0}^{q_1-1} \cdots \sum_{k_m=0}^{q_m-1} \widehat{f}(k_1, \ldots, k_m) \mathrm{e}^{2i\pi(k_1 t_1/q_1 + \cdots + k_m t_m/q_m)} \qquad \text{(A.10)}
\end{aligned}$$

où $\widehat{f}(\chi) = < f, \chi >_\mu$ et la notation $\widehat{f}(k_1, \ldots, k_m)$ désigne $< f, \chi_1^{k_1} \cdots \chi_m^{k_m} >_\mu$.

2.3. La décomposition (A.10) s'applique en particulier à la fonction τ qui décrit l'effet des traitements sur la sortie dans le modèle linéaire (A.1). La quantité $\widehat{\tau}(\chi) = \widehat{\tau}(k_1, \ldots, k_m)$ peut être appelée *paramètre canonique*, ou *effet factoriel*, ou plus simplement *contraste* associé au traitement de niveaux $k_1, \ldots, k_m$ (voir [Kob95]). Plus généralement, la Formule (A.10) constitue la *décomposition en série de Fourier* de la fonction f et les $\widehat{f}(k_1, \ldots, k_m)$ sont les *coefficients de Fourier*.

2.4. Remarquons que la base orthonormale constituée des caractères de $\mathcal{T}$ — quelquefois appelée *base de Yates* — est *adaptée* à la décomposition orthogonale de la Section 1.4., i.e. pour tout $\mathfrak{u} \subset \mathfrak{m}$, toute fonction de $\Theta_{\mu,\mathfrak{u}}$ se décompose sur

$$\mathcal{B}_\mathfrak{u} = \{\chi_1^{k_1} \cdots \chi_m^{k_m},\ 0 < k_i < q_i,\ \text{si } i \in \mathfrak{u} \ \text{ et } \ k_i = 0 \text{ sinon}\}$$

et

$$\mathcal{T}^* = \bigsqcup_{\mathfrak{u} \subset \mathfrak{m}} \mathcal{B}_\mathfrak{u} \ .$$

Par suite en notant, pour tout $\mathfrak{u} \subset \mathfrak{m}$,

$$\mathbb{K}_\mathfrak{u} = \{(k_1, \ldots, k_m),\ 0 < k_i < q_i,\ \text{si } i \in \mathfrak{u} \ \text{ et } \ k_i = 0 \text{ sinon}\},$$

on déduit de la formule de Parseval que pour tout $\mathfrak{u}$ non vide inclu $\mathfrak{m}$,

$$\mathrm{Var}_\mu[\tau_\mathfrak{u}] = \sum_{\mathbf{k} \in \mathbb{K}_\mathfrak{u}} \left|\widehat{\tau}(\mathbf{k})\right|^2 ,$$

et

$$\mathrm{Var}_\mu[\tau] = \sum_{k_1=0}^{q_1-1} \cdots \sum_{k_m=0}^{q_m-1} \left|\widehat{\tau}(\mathbf{k})\right|^2 .$$

2.5. (Remarques)

2.5.1. Contrairement aux effets principaux et aux interations qui sont des objets canoniques, les effets factoriels sont définis relativement à une base.

2.5.2. Si $\mathcal{T}$ n'est pas fini, il est possible de dériver le formalisme précédent. Si les $\mathcal{T}_i$ sont les intervalles $[0,1]$ et μ_i des mesures uniformes — c.f. Remarque 1.5.2 a) —, on munit $\mathcal{T}$ d'une structure de tore $\mathbb{T}^m \simeq (\mathbb{R}/\mathbb{Z})^m$ et toutes les propriétés énoncées précédemment se dérivent aisément et restent vraies (voir [TP12c]). Notons néanmoins que la Proposition A.1 ne tient pas dans ce cas, le tore n'étant pas isomorphe à son dual — l'hypothèse de finitude est nécessaire —. L'isomorphisme exhibant une base de $(\mathbb{T}^m)^*$ est ici

$$\begin{array}{rccc} \Psi : & \mathbb{Z}^m & \longrightarrow & (\mathbb{T}^m)^* \\ & (k_1,\ldots,k_m) & \longmapsto & \chi_1^{k_1}\cdots\chi_m^{k_m} \end{array}$$

où

$$\begin{array}{rccc} \chi_i : & \mathbb{T}^m & \longrightarrow & \mathbb{C}^\times \\ & (t_1,\ldots,t_m) & \longmapsto & \mathrm{e}^{2i\pi t_i} \ . \end{array}$$

A.3 Confusion d'effets (préliminaires)

Cette section est un préambule à la question de la confusion des effets factoriels que nous aborderons à la section suivante. Nous renvoyons à l'Annexe A.A. pour la notion de classe à gauche et de groupe quotient.

3.1. Soient $f \in \mathbb{C}^{\mathcal{T}}$, $\mathcal{S}$ un sous-groupe de $\mathcal{T}$, $i : \mathcal{S} \hookrightarrow \mathcal{T}$ l'injection canonique, $\mathcal{S}^*$ le groupe dual de $\mathcal{S}$ et $\mu_{\mathcal{S}}$ la mesure d'équiprobabilité sur $\mathcal{S}$, et considérons la restriction de f à $\mathcal{S}$

$$\begin{array}{rccc} f_{\mathcal{S}} : & \mathcal{S} & \longrightarrow & \mathbb{R} \\ & s & \longmapsto & f(s) \ . \end{array}$$

D'après le Théorème A.2, $\mathcal{S}^*$ forme une base de $\mathbb{C}^{\mathcal{S}}$, d'où la décomposition

$$\forall s \in \mathcal{S},\ f_{\mathcal{S}}(s) = \sum_{\xi \in \mathcal{S}^*} \widehat{f}_{\mathcal{S}}(\xi)\xi(s)$$

où les $\widehat{f}_{\mathcal{S}}(\xi) =< f_{\mathcal{S}}, \xi >_{\mu_{\mathcal{S}}}$ peuvent être reliés aux $\widehat{f}(\chi) =< f, \chi >_{\mu}$, $\chi \in \mathcal{T}^*$, par le biais de la formule Sommatoire de Poisson généralisée (voir [Bou67] p127). On a d'abord la

Proposition A.2. *Soit $\chi_0 \in \mathcal{T}^*$, alors on a l'égalité suivante*

$$\widehat{f}_{\mathcal{S}}(\chi_0 \circ i) = \sum_{\chi \in \mathcal{S}^\perp} \widehat{f}(\chi\chi_0)$$

où $\mathcal{S}^\perp = \{\chi \in \mathcal{T}^* \mid \forall s \in \mathcal{S},\ \chi(s) = 1\}$ est le sous-groupe de $\mathcal{T}^*$ orthogonal à $\mathcal{S}$.

puis en définissant l'homomorphisme dual de i

$$\begin{array}{rccc} i^* : & \mathcal{T}^* & \longrightarrow & \mathcal{S}^* \\ & \chi & \longmapsto & \chi \circ i \ . \end{array}$$

il vient le

Corollaire A.1. *Soit $\xi \in \mathcal{S}^*$, on a l'égalité suivante*

$$\widehat{f}_{\mathcal{S}}(\xi) = \sum_{\chi \in \mathcal{S}^\perp} \widehat{f}(\chi\widetilde{\xi})$$

où $\widetilde{\xi}$ est un antécédent quelconque de ξ par i^*.

Démonstration. (de la Proposition A.2) Si χ_0 est l'élément neutre $e_{\mathcal{T}^*}$, on a

$$\begin{aligned}\widehat{f_{\mathcal{S}}}(e_{\mathcal{T}^*} \circ i) &= \widehat{f_{\mathcal{S}}}(e_{\mathcal{S}^*}) \\ &= < f_{\mathcal{S}}, e_{\mathcal{S}^*} >_{\mu_{\mathcal{S}}} \\ &= \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} f_{\mathcal{S}}(s)\overline{e_{\mathcal{S}^*}(s)} \\ &= \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} f_{\mathcal{S}}(s)\end{aligned}$$

et la Proposition A.2 suit par la formule sommatoire de Poisson généralisée — admis, voir [Bou67] — qui établit que

$$\frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} f_{\mathcal{S}}(s) = \sum_{\chi \in \mathcal{S}^\perp} \widehat{f}(\chi).$$

Dans le cas contraire, on considère l'application

$$\begin{array}{rccc} f^{\chi_0} : & \mathcal{T} & \longrightarrow & \mathbb{C} \\ & t & \longmapsto & f(t)\overline{\chi_0(t)} \end{array}$$

et sa restriction à $\mathcal{S}$, $f_{\mathcal{S}}^{\chi_0} = f^{\chi_0} \circ i$. On a alors

$$\begin{aligned}\widehat{f_{\mathcal{S}}}(\chi_0 \circ i) &= \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} f_{\mathcal{S}}(s)\overline{\chi_0(s)} \\ &= \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} f_{\mathcal{S}}^{\chi_0}(s) \\ &= \widehat{f_{\mathcal{S}}^{\chi_0}}(e_{\mathcal{S}^*})\end{aligned}$$

puis d'après ce qui précède,

$$\begin{aligned}\widehat{f_{\mathcal{S}}^{\chi_0}}(e_{\mathcal{S}^*}) &= \sum_{\chi \in \mathcal{S}^\perp} \widehat{f^{\chi_0}}(\chi e_{\mathcal{S}^*}) \\ &= \sum_{\chi \in \mathcal{S}^\perp} \widehat{f}(\chi\chi_0)\ ,\end{aligned}$$

et la conclusion suit. □

Démonstration. (du Corollaire A.1) Il suffit de montrer que i^* est surjectif. On note d'abord que

$$\text{Ker } i^* = \mathcal{S}^\perp$$

puis par le théorème de factorisation des homomorphismes de groupes — voir Théorème 1.1.4.3. p18 dans [Fre01] — on déduit

$$\mathcal{T}^*/\mathcal{S}^\perp \simeq \text{Im } i^*\ .$$

Enfin, le résultat classique suivant

$$\mathcal{T}^*/\mathcal{S}^\perp \simeq \mathcal{S}^* \tag{A.11}$$

— voir preuve dans l'Annexe A.B. — permet de conclure. □

3.2. Soit $\mathcal{C} = t_0 + \mathcal{S}$, $t_0 \in \mathcal{T}$, une classe à gauche de $\mathcal{T}$ suivant $\mathcal{S}$, et considérons $f_{\mathcal{C}}$ la restriction de f à $\mathcal{C}$ vue comme une fonction définie sur $\mathcal{S}$,

$$\begin{array}{rccc} f_{\mathcal{C}} : & \mathcal{S} & \longrightarrow & \mathbb{R} \\ & s & \longmapsto & f(t_0 + s)\ . \end{array}$$

D'après le Théorème A.2, on a la décomposition

$$\forall s \in \mathcal{S},\ f_{\mathcal{C}}(s) = \sum_{\xi \in \mathcal{S}^*} \widehat{f}_{\mathcal{C}}(\xi)\xi(s) \tag{A.12}$$

où les $\widehat{f}_{\mathcal{C}}(\xi) =< f(t_0 + \cdot), \xi >_{\mu_{\mathcal{S}}}$ peuvent être reliés aux $\widehat{f}(\chi)$, $\chi \in \mathcal{T}^*$ par le

Corollaire A.2. *Soient $\mathcal{C} = t_0 + \mathcal{S}$ et $\xi \in \mathcal{S}^*$, on a l'égalité suivante*

$$\widehat{f}_{\mathcal{C}}(\xi) = \sum_{\chi \in \mathcal{S}^\perp} \chi\widetilde{\xi}(t_0)\widehat{f}(\chi\widetilde{\xi})$$

où $\widetilde{\xi}$ est un antécédent quelconque de ξ par i^*.

Démonstration. En notant $g : t \mapsto f(t_0 + t)$, le Corollaire A.1 donne

$$\widehat{f}_{\mathcal{C}}(\xi) = \sum_{\chi \in \mathcal{S}^\perp} \widehat{g}(\chi\widetilde{\xi})$$

et la conclusion suit en notant que

$$\begin{aligned}
\widehat{g}(\chi\widetilde{\xi}) &= \frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}} g(t)\overline{\chi\widetilde{\xi}(t)} \\
&= \frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}} g(t-t_0)\overline{\chi\widetilde{\xi}(t-t_0)} \\
&= \overline{\chi\widetilde{\xi}(-t_0)}\,\frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}} f(t)\overline{\chi\widetilde{\xi}(t)} \\
&= \chi\widetilde{\xi}(t_0)\widehat{f}(\chi\widetilde{\xi})\ .
\end{aligned}$$

□

3.3. (Remarque) Si les $\mathcal{T}_i$ sont les intervalles $[0,1]$ — voir Remarques 1.5.2. et 2.5.2. —, les résultats précédents restent vrais. Pour la Formule de Poisson généralisée on renvoie à [Bou67] et pour la Formule (A.11), on se réfère à [Rud62].

A.4 Confusion d'effets (suite) et méthode FAST

L'ensemble des unités expérimentales $\mathcal{U}$ est maintenant supposé muni d'une structure de groupe abélien.

4.1. Nous considérons les plans factoriels d donnés par

$$d(u) = t_0 + \Phi(u)$$

où Φ est un homomorphisme du groupe $\mathcal{U}$ vers le groupe $\mathcal{T}$, et t_0 un traitement dans $\mathcal{T}$. Dans ce cas l'ensemble des traitements qui sont effectivement expérimentés correspond à l'ensemble

$$\mathcal{C} = t_0 + \mathrm{Im}\Phi$$

où $\mathrm{Im}\Phi = \{\Phi(u),\ u \in \mathcal{U}\}$ désigne l'image de l'homomorphisme Φ. $\mathcal{C}$ est une classe à gauche de $\mathcal{T}$ suivant le sous-groupe $\mathrm{Im}\Phi$. L'ensemble des unités qui reçoivent le même traitement qu'une unité $u_0 \in \mathcal{U}$ est

$$u_0 + \mathrm{Ker}\Phi$$

où $\mathrm{Ker}\Phi = \{u \in \mathcal{U},\ \Phi(u) = e_{\mathcal{T}^*}\}$ désigne le noyau de l'homomorphisme Φ. Cet ensemble constitue une classe à gauche de $\mathcal{U}$ suivant le sous-groupe $\mathrm{Ker}\Phi$.

4.2. Une application directe de (A.12) et du Corollaire A.2 donne la décomposition

$$\forall u \in \mathcal{U},\ \tau\big(d(u)\big) = \sum_{\xi \in (\mathrm{Im}\Phi)^*} \widehat{\tau_{\mathcal{C}}}(\xi)\xi\big(\Phi(u)\big) \tag{A.13}$$

avec

$$\widehat{\tau_{\mathcal{C}}}(\xi) = \sum_{\chi \in (\mathrm{Im}\Phi)^\perp} \chi\widetilde{\xi}(t_0)\widehat{\tau}(\chi\widetilde{\xi}) \ , \tag{A.14}$$

où $\widetilde{\xi}$ est un antécédent quelconque de ξ par $i^* : \mathcal{T}^* \to (\mathrm{Im}\Phi)^*$, l'application duale de l'injection canonique $i : \mathrm{Im}\Phi \hookrightarrow \mathcal{T}$. Puis en notant Φ^* l'homomorphisme dual de Φ,

$$\begin{array}{rccc} \Phi^* : & \mathcal{T}^* & \longrightarrow & \mathcal{S}^* \\ & \chi & \longmapsto & \chi \circ \Phi \ , \end{array}$$

et en remarquant que

$$\begin{array}{rcl} \{\chi\widetilde{\xi} \mid \chi \in (\mathrm{Im}\Phi)^\perp\} & = & \{\chi\widetilde{\xi} \mid \forall s \in \mathrm{Im}\Phi,\ \chi\big(i(s)\big) = 1\} \\ & = & \big\{\chi \mid \forall s \in \mathrm{Im}\Phi,\ \chi\big(i(s)\big) = \widetilde{\xi}\big(i(s)\big)\big\} \\ & = & \{\chi \mid \forall u \in \mathcal{U},\ \chi\big(\Phi(u)\big) = \widetilde{\xi}\big(\Phi(u)\big)\} \\ & = & \{\chi \mid \Phi^*(\chi) = \xi \circ \Phi\} \end{array}$$

la Formule (A.14) se réécrit simplement

$$\widehat{\tau_{\mathcal{C}}}(\xi) = \sum_{\chi : \Phi^*(\chi) = \xi\circ\Phi} \chi(t_0)\widehat{\tau}(\chi) \ . \tag{A.15}$$

Finalement les Formules (A.13) et (A.15) donnent

$$\forall u \in \mathcal{U},\ \tau\big(d(u)\big) = \sum_{\xi \in (\mathrm{Im}\Phi)^*} \Big(\sum_{\chi : \Phi^*(\chi) = \xi\circ\Phi} \chi(t_0)\widehat{\tau}(\chi) \Big) \xi \circ \Phi(u)$$

i.e.

$$\forall u \in \mathcal{U},\ \tau\big(d(u)\big) = \sum_{\xi \in \mathcal{U}^*} \Big(\underbrace{\sum_{\chi : \Phi^*(\chi) = \xi} \chi(t_0)\widehat{\tau}(\chi)}_{\widehat{\tau \circ d}(\xi)} \Big) \xi(u).$$

On en déduit la notion de confusion d'effets

Définition A.1. *Soit $\xi \in \mathcal{U}^*$. Les effets factoriels $\widehat{\tau}(\chi)$ tels $\Phi^*(\chi) = \xi$ sont dits* confondus (*avec $\widehat{\tau \circ d}(\xi)$) ; on dit également que les caractères χ sont confondus (avec ξ).*

La confusion d'effets est caractérisée par la proposition suivante

Proposition A.3. *Deux effets factoriels $\widehat{\tau}(\chi_1)$ et $\widehat{\tau}(\chi_2)$ sont confondus si et seulement si $\chi_1\chi_2^{-1} \in \mathrm{Ker}\Phi^*$.*

Démonstration. Immédiat par la Définition A.1. □

4.3. De la même manière que pour $\mathcal{T}$ dans (A.5), posons

$$\mathcal{U} \simeq \mathbb{Z}/p_1\mathbb{Z} \times \cdots \times \mathbb{Z}/p_{m'}\mathbb{Z}$$

et comme pour $\mathcal{T}$ dans (A.8), posons pour chaque $1 \leq j \leq m'$

$$\begin{array}{rccc} \xi_j : & \mathcal{U} & \longrightarrow & \mathbb{C}^\times \\ & (u_1, \ldots, u_{m'}) & \longmapsto & \mathrm{e}^{2i\pi t_j/p_j} \end{array} \ .$$

Par suite, tout comme pour chaque $\chi \in \mathcal{T}^*$, il existe $(k_1, \ldots, k_m)$, $0 \leq k_i < q_i$, tel que $\chi = \chi_1^{k_1} \times \cdots \times \chi_m^{k_m}$; pour chaque $\xi \in \mathcal{U}^*$, il existe $(h_1, \ldots, h_{m'})$, $0 \leq h_i < p_i$, tel que $\xi = \xi_1^{h_1} \times \cdots \times \xi_{m'}^{h_{m'}}$. Rappelons alors — voir §**2.3.** — que $\widehat{\tau}(\chi) = \widehat{\tau}(k_1, \ldots, k_m)$ est l'effet factoriel de τ associé au traitement de niveaux $k_1, \ldots k_m$, et que c'est une quantité qu'on souhaite connaître. Et d'autre part, si $Z_u = 0$ — voir §**1.2.** puis remarque en fin de section — les $\widehat{\tau \circ d}(\xi) = \widehat{\tau \circ d}(h_1, \ldots h_{m'})$ définis par

$$\widehat{\tau \circ d}(h_1, \ldots h_{m'}) = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \tau\big(d(u)\big) e^{-2i\pi \mathbf{h} \cdot u}$$

sont les quantités qu'on connait par expérimentation sur le plan d. Par conséquent, la formule de confusion d'effets nous dit que si parmi les χ confondus avec ξ, un seul possède un effet factoriel $\widehat{\tau}(\chi)$ non négligeable, ce dernier est correctement estimé. Une application directe de ce constat est donc, en pratique, de fixer de manière a priori l'ensemble des χ ayant des effets factoriels $\widehat{\tau}(\chi)$ non négligeables, noté $\mathcal{E}_{NN}$, et de trouver $\mathcal{U}$ et Φ satisfaisant au critère de confusion d'effets, i.e. construire $\Phi : \mathcal{U} \to \mathcal{T}$ tel que $\forall \chi,\ \chi' \in \mathcal{E}_{NN}$, $\chi \neq \chi'$, $\chi \chi'^{-1} \notin \mathrm{Ker}\Phi$.

4.4. D'après la Remarque **3.3.**, ce qui précède s'étend au cas où $\mathcal{T}$ n'est pas fini. En particulier, dans la méthode FAST on a $\mathcal{T}_i = [0, 1]$ et par suite $\mathcal{T}$ est le tore $\mathbb{T}^d$. Dans ce cas, par l'isomorphisme de la Formule (A.9), se donner un sous-ensemble de $\chi \in \mathcal{T}^*$ dont les effets factoriels sont a priori non négligeables revient à se donner $K \subset \mathbb{Z}^d$. Imposons maintenant à $\mathcal{U}$ d'être cyclique d'ordre n. Par suite, la construction d'un plan factoriel sans confusion d'effets revient à construire

$$\begin{array}{rccc} \Phi : & \mathbb{Z}/n\mathbb{Z} & \longrightarrow & \mathbb{T}^d \\ & k & \longmapsto & \Phi(k) \end{array} .$$

Φ est connu explicitement par $\Phi(1)$ car $\Phi(k) = \{k\Phi(1)\}$; et remarquons que $n\Phi(1) \in \mathbb{Z}^d$. Par suite, tout Φ de la forme précédente peut donc s'écrire

$$\begin{array}{rccc} \Phi_{\boldsymbol{\omega}} : & \mathbb{Z}/n\mathbb{Z} & \longrightarrow & \mathbb{T}^d \\ & k & \longmapsto & \left\{\frac{k}{n}\boldsymbol{\omega}\right\} \end{array} .$$

où $\boldsymbol{\omega} \in \mathbb{N}^*$. On en est donc réduit à chercher n — minimal — et $\boldsymbol{\omega}$ qui satisfont le critère de confusion d'effets pour l'ensemble K introduit ci-dessus. On remarque alors que l'homomorphisme dual de $\Phi_{\boldsymbol{\omega}}$ est

$$\begin{array}{rccc} \Phi^*_{\boldsymbol{\omega}} : & \mathbb{Z}^d & \longrightarrow & \mathbb{Z}/n\mathbb{Z} \\ & (k_1, \ldots, k_d) & \longmapsto & k_1\omega_1 + \cdots + k_d\omega_d \end{array} .$$

et par suite le critère pour éviter la confusion d'effets s'écrit

$$\forall \mathbf{k}, \mathbf{k}' \in\ K,\ \mathbf{k} \neq \mathbf{k}',\ (\mathbf{k} - \mathbf{k}') \cdot \boldsymbol{\omega} \not\equiv 0 \ (\mathrm{mod}\ n) \tag{A.16}$$

ce qui est rigoureusement le critère dans la méthode FAST revisitée (voir Formule (5.40) dans le Chapitre 5).

4.5. Le plan obtenu dans le paragraphe précédent est une fraction régulière[4] du groupe fini

$$\mathcal{T} \simeq \mathbb{Z}/\widetilde{\omega_1}\mathbb{Z} \times \cdots \times \mathbb{Z}/\widetilde{\omega_d}\mathbb{Z}$$

où $\widetilde{\omega_i} = \omega_i/(\omega_i \wedge n)$. Vu comme un tableau orthogonal, il est de force 1 et a priori, il n'y a pas de raison qu'il soit de force supérieure.

4. Le fait que $\Phi_{\boldsymbol{\omega}}$ soit injectif est une conséquence de l'optimisation sur n qui doit être choisi le plus petit possible. En effet, considérons $(\boldsymbol{\omega}, n)$ satisfaisant (A.16) et tel que $\Phi_{\boldsymbol{\omega}}$ ne soit pas injectif ; alors pour tout i, ω_i divise n. Par suite, en posant $p = \omega_1 \wedge \cdots \wedge \omega_d \wedge n$, on montre que $(\boldsymbol{\omega}/p, n/p)$ vérifie (A.16) et que $\Phi_{\boldsymbol{\omega}/n}$ est injectif.

4.6. (Remarque) Si Z_u n'est pas identiquement égal à 0, la valeur à laquelle on a accès est

$$\widetilde{\tau \circ d}(h_1, \dots h_{m'}) = \widehat{\tau \circ d}(h_1, \dots h_{m'}) + \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} Z_u e^{-2i\pi \mathbf{h} \cdot u}.$$

Et par conséquent, cette valeur $\widetilde{\tau \circ d}(h_1, \dots h_{m'})$ est aléatoire ; on a

$$\mathbb{E}\big[\widetilde{\tau \circ d}(h_1, \dots h_{m'})\big] = \widehat{\tau \circ d}(h_1, \dots h_{m'}) + B_u$$

où

$$B_u = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \mathbb{E}[Z_u] e^{-2i\pi \mathbf{h} \cdot u}.$$

A.A Classe à gauche et groupe quotient

Les ensembles $t + \mathcal{S}$, $t \in \mathcal{T}$ sont appelés *classes à gauche* de $\mathcal{T}$ suivant $\mathcal{S}$. Deux classes $t_1 + \mathcal{S}$ et $t_2 + \mathcal{S}$ sont égales si et seulement si $(-t_1) + t_2 \in \mathcal{S}$; dans le cas contraire, les deux classes sont disjointes. L'ensemble des classes à gauche est noté $\mathcal{T}/\mathcal{S}$. On définit de manière analogue les classes à droites dont l'ensemble est noté $\mathcal{S}\backslash\mathcal{T}$. Si $\mathcal{S}$ est un *sous-groupe distingué* de $\mathcal{T}$ — autrement dit si les classes à gauche et les classes à droites coïncident — alors les classes à gauche (resp. à droites) forment un groupe pour la loi

$$(t_1 + \mathcal{S}) + (t_2 + \mathcal{S}) = (t_1 + t_2) + \mathcal{S}.$$

Ce groupe est noté $\mathcal{T}/\mathcal{S}$ et est appelé groupe quotient de $\mathcal{T}$ par $\mathcal{S}$. Dans ce cas, on peut considérer l'homomorphisme surjectif

$$\begin{array}{rccc} \Pi : & \mathcal{T} & \longrightarrow & \mathcal{T}/\mathcal{S} \\ & t & \longmapsto & t + \mathcal{S} \end{array}$$

— appelé *projection canonique* — dont le noyau est $\mathcal{S}$.

On rappelle que dans le cas où $\mathcal{T}$ est abélien, tout sous-groupe $\mathcal{S}$ de $\mathcal{T}$ est distingué, et par suite l'ensemble des classes à gauche $\mathcal{T}/\mathcal{S}$ possède automatiquement une structure de groupe.

A.B Preuve de la Formule (A.11)

Dans un premier temps, on considère l'homomorphisme de groupes

$$\begin{array}{rccc} \Psi : & \mathcal{S}^\perp & \longrightarrow & (\mathcal{T}/\mathcal{S})^* \\ & \chi & \longmapsto & \overline{\chi} : t + \mathcal{S} \mapsto \chi(t) \end{array}$$

On vérifie aisément que Ψ est bien défini, que son noyau est réduit à l'élément neutre, et qu'il est surjectif car pour tout $\overline{\chi} \in (\mathcal{T}/\mathcal{S})^*$, on a $\overline{\chi} = \Psi(\chi)$ avec

$$\begin{array}{rccc} \chi : & \mathcal{T} & \longrightarrow & \mathbb{C}^\times \\ & t & \longmapsto & \overline{\chi}(t + \mathcal{S}) \ . \end{array}$$

Par suite, on a

$$(\mathcal{T}/\mathcal{S})^* \simeq \mathcal{S}^\perp \ . \tag{A.17}$$

Dans un second temps, notons que $\mathcal{T}$ étant supposé fini, $(\mathcal{T}/\mathcal{S})^*$ est isomorphe à $\mathcal{T}/\mathcal{S}$ et, en particulier, ils ont même ordre, $|\mathcal{T}|/|\mathcal{S}|$; et donc d'après la Formule (A.17) on a

$$|\mathcal{S}^\perp| \cdot |\mathcal{S}| = |\mathcal{T}| \ . \tag{A.18}$$

De même, on a

$$|\mathcal{S}^{\perp\perp}| \cdot |\mathcal{S}^{\perp\perp\perp}| = |\mathcal{T}| \ . \tag{A.19}$$

Enfin, en notant que $\mathcal{S} \subset \mathcal{S}^{\perp\perp}$ et $\mathcal{S}^\perp \subset \mathcal{S}^{\perp\perp\perp}$, on déduit de (A.18) et (A.19) que

$$\mathcal{S} = \mathcal{S}^{\perp\perp} \ . \tag{A.20}$$

Finalement, (A.17) et (A.20) donnent

$$\mathcal{T}^*/\mathcal{S}^\perp \simeq \mathcal{S} \ .$$

Annexe B

Conditions aux limites du modèle hydrodynamique

Le modèle hydrodynamique qu'il n'y a aucun flux de chaleur et de salinité au bord du domaine. Quant aux autres conditions aux limites, elles sont données par les équations suivantes.

Flux de quantité de mouvement en surface

$$\tilde{\lambda}\frac{\partial(u,v)}{\partial z}\Big|_{surface} = C_0 * (1 + 0.1 * ||(u_{vent}, v_{vent})||) * ||(u_{vent}, v_{vent})|| * (u_{vent}, v_{vent}) \tag{B.1}$$

avec

C_0 : $0.63.10^{-6}$
(u_{vent}, v_{vent}) : vitesses horizontales du vent à 10 mètres d'altitude.

Flux de l'énergie cinétique turbulente en surface

$$\tilde{\lambda}\frac{\partial k}{\partial z}\Big|_{surface} = 3.10^{-3} * C_0 * ||(u_{vent}, v_{vent})||^3. \tag{B.2}$$

Flux de chaleur en surface

$$Q = \rho * C_p * \tilde{\lambda}^T * \frac{\partial T}{\partial z}\Big|_{surface} = Q_I - Q_R - Q_S - Q_L \tag{B.3}$$

avec

ρ : masse volumique de l'eau (kgm^{-3})
C_p : chaleur spécifique de l'eau ($Jkg^{-1\circ}C^{-1}$)
Q : flux de chaleur en surface (Wm^{-2})
Q_I : radiation solaire (Wm^{-2})
Q_R : radiation infrarouge (Wm^{-2})
Q_S : flux de chaleur sensible (Wm^{-2})
Q_L : flux de chaleur latente (Wm^{-2}).

Flux de salinité en surface

$$\tilde{\lambda}^S * \frac{\partial S}{\partial z}\Big|_{surface} = \frac{1}{\rho} * (E_v - P_r) * S_0 \tag{B.4}$$

avec

P_r : précipitations (ms^{-1})
E_v : flux de surface de vaporisation (ms^{-1})
S_0 : salinité en surface

Flux de quantité de mouvement au fond

$$\left|\left|\tilde{\lambda}\frac{\partial(u,v)}{\partial z}\Big|_{fond}\right|\right| = \left[\frac{\kappa}{\ln(z_1/z_0)}\right]^2 \left|\left|\big(u(z_1),v(z_1)\big)\right|\right|^2 \tag{B.5}$$

avec

z_0 : longueur de rugosité (m)
κ : constante de Von Karman $(= 0.4)$
z_1 : hauteur du centre de la première maille (au fond) (m).

Flux d'énergie cinétique turbulente au fond

$$k = 4 * \left[\frac{\kappa}{\ln(z_1/z_0)}\right]^2 \left|\left|\big(u(z_1),v(z_1)\big)\right|\right|^2. \tag{B.6}$$

Bibliographie

Bibliographie

[AN88] V. Andersen and P. Nival. A pelagic ecosystem model simulating production and sedimentation of biogenic particles : Role of salps and copepods. *Mar. Ecol. Prog.*, 44 :37–50, 1988.

[Ant84] A. Antoniadis. Analysis of variance on function spaces. *Math. Oper. Forsch. und Statist., series Statistics*, 15(1) :59–71, 1984.

[AO01] J. An and A. B. Owen. Quasi-regression. *Journal of Complexity*, 17(4) :588–607, 2001.

[ASS97] G. E. B. Archer, A. Saltelli, and I. M. Sobol'. Sensitivity measures, ANOVA-like techniques and the use of bootstrap. *Journal of Statistical Computation and Simulation*, 58 :99–120, 1997.

[Bai96] R. A. Bailey. Orthogonal partitions in designed experiments. *Designs, Codes and Cryptography*, 8 :45–77, 1996.

[Bar07] J. O. Baruch. 10 idées reçues sur le climat. *La Recherche*, 412, 2007.

[BBCG08] D. Bakry, F. Barthe, P. Cattiaux, and A. Guillin. A simple proof of the Poincaré inequality for a large class of probability measures including the log-concave case. *Electron. Commun. Probab.*, 13 :60–66, 2008.

[BCT11] E. Borgonovo, W. Castaings, and S. Tarantola. Moment independent importance measures : New results and analytical test cases. *Risk Analysis*, 31(3) :404–428, 2011.

[Bea75] K. G. Beauchamp. *Walsh functions and their applications.* Academic Press, London, 1975.

[BG04] H. J. Bungartz and M. Griebel. Sparse grids. *Acta Numerica*, 13 :147–269, 2004.

[BGMA10] A. Boukouvalas, J. P. Gosling, and H. Maruri-Aguilar. An efficient screening method for computer experiments. *Preprint disponible à* `https://wiki.aston.ac.uk/foswiki/pub/AlexisBoukouvalas/WebHome/screenReport.pdf`, 2010+.

[Bla09] G. Blatman. Adaptive sparse polynomial chaos expansions for uncertainty propagation and sensitivity analysis. *Thèse de doctorat disponible à* `http://bruno.sudret.free.fr/docs/Thesis%20Blatman.pdf`, 2009.

[BNR00] V. Barthelmann, E. Novak, and K. Ritter. High dimensional polynomial interpolation on sparse grids. *Adv. Comput. Math.*, 12 :273–288, 2000.

[Bor07] E. Borgonovo. A new uncertainty importance measure. *Reliability Engineering and System Safety*, 92(6) :771–784, 2007.

[Bos38] R. Bose. On the application of the theory of galois fields to the problem of construction of hyper-graeco-Latin squares. *Sankhya*, 3 :323–338, 1938.

[Bou67] N. Bourbaki. *Théories spectrales (chap. 1 et 2).* Hermann, 1967.

[BRK01] R. Brun, P. Reichert, and H. R. Kunsch. Practical identifiability analysis of large environmental simulation models. *Water Resources Research*, 3(4) :1015–1030, 2001.

[BS10] G. Blatman and B. Sudret. Efficient computation of global sensitivity indices using sparse polynomial chaos expansions. *Reliability Engineering and System Safety*, 95 :1216–1229, 2010.

[CCS07] F. Campolongo, J. Cariboni, and A. Saltelli. An effective screening design for sensitivity analysis of large models. *Environ. Modell. Software*, 22 :1509–1518, 2007.

[CFS+73] R. I. Cukier, C. M. Fortuin, K. E. Shuler, A. G. Petschek, and J. H. Schaibly. Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients : Theory. *Journal of Chemical Physics*, 59 :3873–3878, 1973.

[CGP12] G. Chastaing, F. Gamboa, and C. Prieur. Generalized Hoeffding-Sobol' decomposition for dependent variables – application to sensitivity analysis. *Preprint disponible à* `http://hal.archives-ouvertes.fr/docs/00/67/77/24/PDF/ps-template.pdf`, 2012+.

[CH53] R. Courant and D. Hilbert. *Methods of Mathematical Physics (Vol I).* John Wiley, New York, 1953.

[CKN06] R. Cools, F. Y. Kuo, and D. Nuyens. Constructing embedded lattice rules for multivariate integration. *SIAM Journal on Scientific Computing*, 28 :2162–2188, 2006.

[CKN10] R. Cools, F. Y. Kuo, and D. Nuyens. Constructing lattice rules based on weighted degree of exactness and worst case error. *Computing*, 87 :63–89, 2010.

[CLS78] R. I. Cukier, H. B. Levine, and K. E. Shuler. Nonlinear sensitivity analysis of multiparameter model systems. *Journal of Computational Physics*, 26 :1–42, 1978.

[CMC03] P. Cheung-Mon-Chan. Réseaux bayésiens et filtres particulaires pour l'égalisation adaptative et le décodage conjoints. *Thèse disponible à* `http://hal.archives-ouvertes.fr/docs/00/49/97/42/PDF/these_pascal_cheung.pdf`, 2003.

[CMO97] R. E. Caflisch, W. Morokoff, and A. B. Owen. Valuation of mortgage backed securities using Brownian bridges to reduce effective dimension. *Journal of Computational Finance*, 1 :27–46, 1997.

[CO68] L. Chambadal and J. L. Ovaert. *Algèbre linéaire et algèbre tensorielle.* Dunod, 1968.

[Col70] D. Collombier. *Plans d'expérience factoriels.* Springer (Collection Mathématiques et Applications, 1970.

[CSS75] R. I. Cukier, J. H. Schaibly, and K. E. Shuler. Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients : Analysis of the approximations. *Journal of Chemical Physics*, 63 :1140–1149, 1975.

[Da-07] S. Da-Veiga. Analyse d'incertitudes et de sensibilité. application aux modèles de cinétique chimique. *Thèse de doctorat disponible à* `http://www.gdr-mascotnum.fr/media/thesedaveiga.pdf`, 2007.

[DJKP95] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard. Wavelet shrinkage : Asymptopia ? *J. Roy. Statist. Soc. Ser. B*, 57 :301–369, 1995.

[DR84] P. J. Davis and P. Rabinowitz. *Methods of Numerical Integration (2nd edition).* Academic, New York, 1984.

[DS89] F. J. Delvos and W. Schempp. *Boolean methods in interpolation and approximation.* Longman Scientific & Technical, Harley, 1989.

[DSWW06] J. Dick, I. Sloan, X. Wang, and H. Woźniakowski. Good lattice rules in weighted korobov spaces with general weights. *Numerische Mathematik*, 103(1) :63–97, 2006.

[ES81] B. Efron and C. Stein. The jackknife estimate of variance. *The Annals of Statistics*, 9(3) :586–596, 1981.

[ET93] B. Efron and R. J. Tibshirani. *An introduction to the bootstrap*. Chapman & Hall, New York, 1993.

[FDM90] M. J. R. Fasham, H. W. Ducklow, and S. M. McKelvie. A nitrogen-based model of plankton dynamics in the oceanic mixed layer. *Journal of Marine Research*, 48 :591–639, 1990.

[Fis25] R. A. Fisher. *Statistical Methods for Research Workers, first edition*. Oliver & Boyd, Edinburgh, 1925.

[Fis35] R. A. Fisher. *The Design of Experiments*. Oliver & Boyd, Edinburgh, 1935.

[Fre01] J. Fresnel. *Groupes*. Hermann, 2001.

[GI12] G. Glen and K. Isaacs. Estimating Sobol' sensitivity indices using correlations. *Environmental Modelling & Software (article in press)*, 2012.

[Gri05] M. Griebel. Sparse grids and related approximation schemes for higher dimensional problems. *Foundations of Computational Mathematics*, 2005.

[Hal59] M. Hall. *Theory of groups*. MacMillan, 1959.

[Ham94] D. M. Hamby. A review of techniques for parameter sensitivity analysis of environmental models. *Environmental Monitoring and Assessment*, 32 :135–154, 1994.

[Hel98] I. S. Helland. A population approach to analysis of variance model. *Scandinavian Journal of Statistics*, 25 :3–15, 1998.

[HI86] S. C. Hora and R. L. Iman. A comparison of maximum bounding and Bayesian Monte Carlo for fault tree uncertainty analysis. *SANDIA National Laboratories Report, SAND85-2839*, 1986.

[Hoc72] U. W. Hochstrasser. Orthogonal polynomials (chap. 22). In M. Abramowitz and I. A. Stegun, editors, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover Publications, New York, 1972.

[Hoe48] W. F. Hoeffding. A class of statistics with asymptotically normal distributions. *Annals of Mathematical Statistics*, 19 :293–325, 1948.

[Hoo07] G. Hooker. Generalized functional ANOVA diagnostics for highdimensional functions of dependent variables. *Journal of Computational and Graphical Statistics*, 16 :709–732, 2007.

[HS96] T. Homma and A. Saltelli. Importance measures in global sensitivity analysis of nonlinear models. *Reliability Engineering and System Safety*, 52(1) :1–17, 1996.

[HSS99] A. S. Hedayat, N. J. A. Sloane, and J. Stufken. *Orthogonal Arrays : Theory and Applications*. Springer-Verlag, New York, 1999.

[IH89] T. Ishigami and T. Homma. An importance quantification technique in uncertainty analysis. *Japan Atomic Energy Research Institute Report, JAERI M 89-111*, 1989.

[IH90a] R. L. Iman and S. C. Hora. A robust measure of uncertainty importance for use in fault tree system analysis. *Risk Analysis*, 10 :401–406, 1990.

[IH90b] T. Ishigami and T. Homma. An importance quantification technique in uncertainty analysis for computers models. *First International Symposium on Uncertainty Modeling and Analysis Proceedings*, pages 398–403, 1990.

[IPB+12] B. Iooss, A. L. Popelin, G. Blatman, C. Ciric, F. Gamboa, S. Lacaze, and M. Lamboni. Some new insights in derivative-based global sensitivity measures. *Preprint disponible à* `http://www.gdr-mascotnum.fr/media/iooss1_esrel12.pdf`, 2012+.

[Jan99] M. J. W. Jansen. Analysis of variance designs for model output. *Comput. Phys. Comm.*, 117 :35–43, 1999.

[JKL+12] A. Janon, T. Klein, A. Lagnoux, M. Nodet, and C. Prieur. Asymptotic normality and efficiency of two Sobol' index estimators. *Preprint disponible à* `http://hal.inria.fr/hal-00665048/en`, 2012.

[JO03] T. Jiang and A. B. Owen. Quasi-regression with shrinkage. *Math. Comput. Simul.*, 62 :231–241, 2003.

[JT86] G. Jacques and P. Treguer. *Écosystèmes pélagiques marins*. Masson et Cie, 1986.

[Kel35] T. L. Kelley. An unbiased correlation ratio measure. *Proceedings of the National Academy of Sciences of the United States of America*, 21(9) :554–559, 1935.

[KK11] L. Kaemmerer and S. Kunis. On the stability of the hyperbolic cross discrete Fourier transform. *Numerische Mathematik*, 117 :581–600, 2011.

[KKP11] L. Kaemmerer, S. Kunis, and D. Potts. Interpolation lattices for hyperbolic cross trigonometric polynomials. *Journal of Complexity*, page In press, 2011.

[Kob95] A. Kobilinsky. Plans d'expériences. *Cours en ligne disponible à* `http://andre.kobilinsky.free.fr/cours/cours2.pdf`, 1995.

[KRFPS09] S. Kucherenko, M. Rodriguez-Fernandez, C. Pantelides, and N. Shah. Monte Carlo evaluation of derivative-based global sensitivity measures. *Reliability Engineering and System Safety*, 94 :1135–1148, 2009.

[KSS12] F. Y. Kuo, C. Schwab, and I. H. Sloan. Quasi- Monte Carlo methods for high dimensional integration – standard (weighted hilbert space) setting and beyond. *ETH Research Report 2012-01*, 2012.

[KSWW12] F. Y. Kuo, I. H. Sloan, G. W. Wasilkowski, and H. Wozniakowski. On decompositions of multivariate functions, mathematics of computation. *Mathematics of Computations*, 79 :953–966, 2012.

[Kuo03] F. Kuo. Component-by-component constructions achieve the optimal rate of convergence for multivariate integration in weighted korobov and Sobol'ev spaces. *Journal of Complexity*, 19 :301–320, 2003.

[Lac90] G. Lacroix. Modélisation 1d d'un système biologique couplé à un modèle hydrodynamique. *Mémoire de DEA, Université de Liège*, 1990.

[Lac98] G. Lacroix. Simulation de l'écosystème pélagique de la mer ligure à l'aide d'un modèle unidimensionnel. étude du bilan de la matière et de la variabilité saisonnière, interrannuelle et spatiale. *Thèse de doctorat*, 1998.

[LD92] G. Lacroix and S. Djenidi. Extending the gher 3d model to the modelling of ecosystems in western mediterranean coastal zones : Results from an exploratory study. *In Proceedings of the third EROS 2000 workshop on research in the North Western Mediterranean Sea, Commission of the European Communities*, pages 89–104, 1992.

[LG02] G. Lacroix and M. Grégoire. Revisited ecosystem model (modecogel) of the ligurian sea : seasonal and interannual variability due to atmospheric forcing. *Journal of Marine Systems*, 37 :229–258, 2002.

[LIPG12] M. Lamboni, B. Iooss, A. L. Popelin, and F. Gamboa. Derivative-based global sensitivity measures : general links with Sobol' indices and numerical tests. *Preprint disponible à* `http://arxiv.org/pdf/1202.0943.pdf`, 2012+.

[LM98] L. Legendre and J. Michaud. Flux of biogenic carbon in oceans : Size-dependent regulation by pelagic food webs. *Mar. Ecol. Prog.*, 164 :1–11, 1998.

[LN98] G. Lacroix and P. Nival. Influence of meteorogical variability on primary production dynamics in the ligurian sea (nw mediterranean sea) with 1d hydrodynamic/biological model. *Journal of Marine Systems*, 37 :229–258, 1998.

[LO02] C. Lemieux and A. B. Owen. Quasi-regression and the relative importance of the ANOVA components of a function. In K. T. Fang, F. J. Hickernel, and H. Niederreiter, editors, *Monte Carlo and Quasi-Monte Carlo Methods*. Springer, Berlin, 2002.

[LO06] R. Liu and A.B. Owen. Estimating mean dimensionality of analysis of variance decompositions. *Journal of the American Statistical Association*, 101(474) :712–721, 2006.

[Loh96] W. L. Loh. On Latin hypercube sampling. *The Annals of Statistics*, 24(5) :2058–2080, 1996.

[Loh08] W. L. Loh. A multivariate central limit theorem for randomized orthogonal array sampling designs in computer experiments. *The Annals of Statistics*, 36 :1983–2023, 2008.

[Loo53] L. H. Loomis. *An Introduction to Abstract Harmonic Analysis*. D. Van Nostrand Company, 1953.

[Mar08] A. Marrel. Mise en œuvre et utilisation du métamodèle processus gaussien pour l'analyse de sensibilité de modèles numériques. *Thèse de doctorat disponible à* `http://eprint.insa-toulouse.fr /archive/00000242/01/Marell.pdf`, 2008.

[Mar09] T. A. Mara. Extension of the rbd-FAST method to the computation of global sensitivity indices. *Reliability Engineering and System Safety*, 94 :1274–1281, 2009.

[Mau02] W. Mauntz. Global sensitivity analysis of general nonlinear systems. *Master's Thesis, Imperial College. Supervisors : C. Pantelides and S. Kucherenko*, 2002.

[MCB79] M. D. McKay, W. J. Conover, and R. J. Beckman. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 21(2) :239–245, 1979.

[McK95] M. D. McKay. Evaluating prediction uncertainty. *Technical Report NUREG/CR-6311, US Nuclear Regulatory Commission and Los Alamos National Laboratory*, pages 1–79, 1995.

[Mey93] Y. Meyer. *Wavelets : Algorithms and applications*. SIAM, 1993.

[Min69] H. J. Minas. Résultats de la campagne mediprod i du jean charcot (1–14 mars 1969 et 3–17 avril 1969), station marine d'endoume et centre d'océanographie. *Faculté des Sciences de Marseille*, 1969.

[MNM06] 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.

[Mor91] M. D. Morris. Factorial sampling plans for preliminary computational experiments. *Technometrics*, 33(2) :161–174, 1991.

[ND90] J. C. J. Nihoul and S. Djenidi. Introduction to system analysis and mathematical modelling applied to the marine system. *The European Advanced Studies in Oceanography, Intensive course : Modeling of Marine Ecosystems*, 1990.

[Nel65] J. A. Nelder. The analysis of randomized experiments with orthogonal block structure, I. *Proceedings of the Royal Society of London A*, 283 :147–178, 1965.

[Nie92] H. Niederreiter. *Random Number Generation and Quasi-Monte Carlo Methods.* SIAM, Philadelphia, 1992.

[Nih81] J. C. J. Nihoul. *Ecohydrodynamics*, chapter Marine Hydrodynamics at Ecological Scale. Elsevier, 1981.

[Nuy07] D. Nuyens. *FAST construction of good lattice rules.* PhD thesis, Catholic University of Leuven, 2007.

[Owe92] A. Owen. Orthogonal arrays for computer experiments, integration and visualization. *Statistica Sinica*, 2 :439–452, 1992.

[Owe94] A. B. Owen. Lattice sampling revisited : Monte Carlo variance of means randomized orthogonal arrays. *The Annals of Statistics*, 22(2) :930–945, 1994.

[Owe97] A. B. Owen. Monte Carlo variance of scrambled equidistribution quadrature. *SIAM Journal of Numerical Analysis*, 34(5) :1884–1910, 1997.

[Owe98] A. B. Owen. Latin supercube sampling for very high-dimensional simulations. *ACM Trans. Model. Comput. Simul.*, 8(1) :71–102, 1998.

[Owe02] A. B. Owen. Necessity of low effective dimension. *Research Report*, 2002.

[Owe12a] A. B. Owen. Better estimation of small Sobol' sensitivity indices. *Preprint disponible à* `http://www-stat.stanford.edu/~owen/reports/newSobol'.pdf`, 2012+.

[Owe12b] A. B. Owen. Effective dimension for weighted function spaces. *Preprint disponible à* `http://www-stat.stanford.edu/~owen/reports/effdim-may.pdf`, 2012+.

[Owe12c] A. B. Owen. Variance components and generalized Sobol' indices. *Preprint disponible à* `http://statistics.stanford.edu/~ckirby/techreports/GEN/2012/2012-07.pdf`, 2012+.

[PC81] T. H. Pierce and R. I. Cukier. Global nonlinear sensitivity analysis using walsh functions. *Journal of Computational Physics*, 41 :427–443, 1981.

[Pli10] E. Plischke. An effective algorithm for computing global sensitivity indices (EASI). *Reliability Engineering and System Safety*, 95(4) :354–360, 2010.

[PT95] A. Papageorgiou and J. Traub. FASTer evaluation of multidimensional integrals. *Computer in Physics*, 22 :113–120, 1995.

[PTH84] T. R. Parsons, M. Takahashi, and B. Hargrave. *Biological Oceanographic Processes.* Oxford, Butterworth-Heinemann, third edition, 1984.

[Qia09] P. Z. G. Qian. Nested Latin hypercube sampling. *Biometrika*, 96(4) :957–970, 2009.

[RA99] H. Rabitz and O. F. Aliş. General foundations of high-dimensional model representations. *Journal of Mathematical Chemistry*, 25 :197–233, 1999.

[RSG06] C. Raick, K. Soetaert, and M. Grégoire. Model complexity and performance : How far can we simplify ? *Progress in Oceanography*, 70 :27–57, 2006.

[Rud62] W. Rudin. *Fourier Analysis on groups*. Interscience Publishers (a division of John Wiley & Sons, New York - London, 1962.

[SAA+10] A. Saltelli, P. Annoni, I. Azzini, F. Campolongo, M. Ratto, and S. Tarantola. Variance based sensitivity analysis of model output. design and estimator for the total sensitivity index. *Computer Physics Communications*, 181 :259–270, 2010.

[Sal02] A. Saltelli. Making best use of model evaluations to compute sensitivity indices. *Computer Physics Communications*, 145 :280–297, 2002.

[Sat59] F. E. Satterthwaite. Random balance experimentation. *Technometrics*, 1(2) :111–137, 1959.

[SB98] A. Saltelli and R. Bolado. An alternative way to compute Fourier amplitude sensitivity test (FAST). *Computational Statistics and Data Analysis*, 26 :445–460, 1998.

[Sch96] J. L. Schnoor. *Environmental Modelling. Fate and Transport of Pollutants in Water, Air and Soil*. New York, Wiley and Sons, Inc., 1996.

[SCS00] A. Saltelli, K. Chan, and E. M. Scott. *Sensitivity Analysis*. John Wiley & Sons, 2000.

[SG95] I. M. Sobol' and A. Gresham. On an alternative global sensitivity estimators. *Proceedings of SAMO, Belgirate*, pages 40–42, 1995.

[Sin77] Y. G. Sinai. *Introduction to ergodic theory*. Princeton University Press, 1977.

[SJ94] I. H. Sloan and S. Joe. *Lattice Methods for Multiple Integration*. Oxford University Press, 1994.

[SK09] I. M. Sobol' and S. Kucherenko. Derivative based global sensitivity measures and their link with global sensitivity indices. *Mathematics and Computers in Simulation*, 79(10) :3009–3017, 2009.

[SK10] I. M. Sobol' and S. Kucherenko. A new derivative based importance criterion for groups of variables and its link with the global sensitivity indices. *Computer Physics Communications*, 181(7) :1212–1217, 2010.

[SM07] I. M. Sobol' and E. E. Myshetskaya. Monte Carlo estimators for small sensitivity indices. *Monte Carlo methods and their applications*, 13 :455–465, 2007.

[Smo63] S. Smolyak. Quadrature and interpolation formulas for tensor products of certain classes of functions. *Soviet. Math. Dokl.*, 4 :240–243, 1963.

[Sob67] I. M. Sobol'. On the distribution of points in a cube and the approximate evaluation of integrals. *Zh. Vychisl. Mat. Mat. Fiz.*, 7(4) :784–802, 1967.

[Sob69] I. M. Sobol'. Multidimensional quadrature formulas and haar functions. *Nauka, Moscow (in russian)*, 1969.

[Sob93] I. M. Sobol'. Sensitivity analysis for nonlinear mathematical models. *Mathematical Modeling and Computational Experiment*, 1 :407–414, 1993.

[Sob03] I. M. Sobol'. Theorems and examples on high dimensional model representation. *Reliability Engineering and System Safety*, 79 :187–193, 2003.

[SRA+08] A. Saltelli, M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli M. Saisana, and S. Tarantola. *Global Sensitivity Analysis : The Primer*. John Wiley & Sons, 2008.

[SS73] J. H. Schaibly and K. E. Shuler. Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients : Applications. *Journal of Chemical Physics*, 59 :3879–3888, 1973.

[SS93] J. H. Siegenthaler and J. L. Sarmiento. Atmospheric carbon dioxyde and the ocean. *Nature*, 365 :119–125, 1993.

[SS95] A. Saltelli and I. M. Sobol'. About the use of rank transformation in sensitivity analysis model. *Reliability Engineering and System Safety*, 50 :225–239, 1995.

[Sta12] R. P. Stanley. *Enumerative Combinatorics, Volume 1 (2nd edition)*. Cambridge University Press, 2012.

[STC99] A. Saltelli, S. Tarantola, and K. P. S. Chan. A quantitative model-independent method for global sensitivity analysis of model output. *Technometrics*, 41 :39–56, 1999.

[Ste87] M. Stein. Large sample properties of simulations using Latin hypercube sampling. *Technometrics*, 29(2) :143–151, 1987.

[Sto94] C. J. Stone. The use of polynomial splines and their tensor products in multivariate function estimation. *The Annals of Statistics*, 118 :118–171, 1994.

[Sud08] B. Sudret. Global sensitivity analysis using polynomial chaos expansions. *Reliability Engineering and System Safety*, 93 :964–979, 2008.

[Tak83] A. Takemura. Tensor analysis of ANOVA decomposition. *Journal of the American Statistical Association*, 78 :894–900, 1983.

[Tan93] B. Tang. Orthogonal array-based Latin hypercubes. *Journal of the American Statistical Association*, 88(424) :1392–1397, 1993.

[TGM06] S. Tarantola, D. Gatelli, and T. A. Mara. Random balance designs for the estimation of first-order global sensitivity indices. *Reliability Engineering and System Safety*, 91 :717–727, 2006.

[Thi00] E. Thiémard. Sur le calcul et la majoration de la discrépance à l'origine. *Thèse de doctorat disponible à* `http://www.ressources-actuarielles.net/ext/isfa/1226.nsf/769998e0a65ea348c1257052003eb94f/f7a0b6627f4656c8c12576af002dd138/$FILE/TheseET.pdf`, 2000.

[Tju84] T. Tjur. Analysis of variance models in orthognal designs. *International Statistical Review*, 52(1) :33–81, 1984.

[Tju91] T. Tjur. Analysis of variance and design of experiments. *Scandinavian Journal of Statistics*, 18 :273–322, 1991.

[TK10] S. Tarantola and M. Koda. Improving random balance designs for the estimation of first order sensitivity indices. *Procedia - Social and Behavioral Sciences*, 2(6) :7753–7754, 2010.

[TP12a] J. Y. Tissot and C. Prieur. Bias correction for the estimation of sensitivity indices based on random balance designs. *Reliability Engineering and System Safety*, 107 :205–213, 2012.

[TP12b] J. Y. Tissot and C. Prieur. Estimating Sobol' indices combining Monte Carlo estimators and Latin hypercube sampling. *Preprint disponible à* `http://hal.archives-ouvertes.fr/hal-00743964`, 2012+.

[TP12c] J. Y. Tissot and C. Prieur. Variance-based sensitivity analysis using harmonic analysis. *Preprint disponible à* `http://hal.archives-ouvertes.fr/docs/00/68/07/25/PDF/FAST_RBD_revisited.pdf`, 2012+.

[Van98] A. W. Van der Vaart. *Asymptotics Statistics*. Cambridge University Press, 1998.

[Vil08] C. Villani. Cours d'intégration et analyse de Fourier, chapitre 6. *Cours disponible à* `http://math.univ-lyon1.fr/~villani/Cours/PDFFILES/chap6.pdf`, 2008.

[Wey16] H. Weyl. Über die gleichverteilung von zahlen mod. eins. *Mathematische Annalen*, 77 :313–352, 1916.

[Wey38] H. Weyl. Mean motion. *American Journal of Mathematics*, 60 :889–896, 1938.

[XG11a] C. Xu and G. Z. Gertner. Reliability of global sensitivity indices. *Journal of Statistical Computation and Simulation*, 81(12) :1939–1969, 2011.

[XG11b] C. Xu and G. Z. Gertner. Understanding and comparisons of different sampling approaches for the Fourier amplitudes sensitivity test (FAST). *Computational Statistics & Data Analysis*, 55(1) :184–198, 2011.

[Zar68] S. K. Zaremba. Some applications of multidimensional integration by parts. *Annales Polonici Mathematici*, 21 :85–96, 1968.

Résumé : Dans les domaines de la modélisation et de la simulation numérique, les simulateurs développés prennent parfois en compte de nombreux paramètres dont l'impact sur les sorties n'est pas toujours bien connu. L'objectif principal de l'analyse de sensibilité est d'aider à mieux comprendre comment les sorties d'un modèle sont sensibles aux variations de ces paramètres. L'approche la mieux adaptée pour appréhender ce problème dans le cas de modèles potentiellement complexes et fortement non linéaires repose sur la décomposition ANOVA et les indices de Sobol'. En particulier, ces derniers permettent de quantifier l'influence de chacun des paramètres sur la réponse du modèle.

Dans cette thèse, nous nous intéressons à l'étude théorique et au développement de méthodes d'estimation des indices de Sobol'. Dans une première partie, nous réintroduisons de manière rigoureuse des méthodes existantes au regard de l'analyse harmonique discrète sur des groupes cycliques et des tableaux orthogonaux randomisés. Cela nous permet d'étudier les propriétés théoriques de ces méthodes et de les généraliser. Dans un second temps, nous considérons la méthode de Monte Carlo spécifique à l'estimation des indices de Sobol' et nous introduisons une nouvelle approche permettant de l'améliorer. Cette amélioration est construite autour de la notion d'hypercube latin et permet de réduire de manière conséquente le nombre de simulations nécessaires pour estimer les indices de Sobol' par cette méthode.

En parallèle, nous mettons en pratique ces différentes méthodes au travers d'une analyse de sensibilité introductive sur un modèle d'écosystème marin.

Mots-clés : analyse de sensibilité, indices de Sobol', décomposition ANOVA, intégration numérique, analyse harmonique discrète