



École doctorale de sciences mathématiques de Paris centre

Thèse de doctorat

Discipline : Mathématiques

Spécialité : Statistiques

présentée par

Nazih Benoumechiara

Traitement de la dépendance en analyse de sensibilité pour la fiabilité industrielle

---

dirigée par Gérard Biau, Bertrand Michel et Philippe Saint-Pierre

Au vu des rapports établis par Jérôme Morio et Jean-Michel Poggi

Soutenue le 10 juillet 2019 devant le jury composé de :

M Gérard Biau	Sorbonne Université	Co-directeur
M Nicolas Bousquet	Quantmetry	Encadrant
M Bertrand Iooss	EDF R&D	Invité
M^{me} Véronique Maume-Deschamps	Université Claude Bernard	Examinatrice
M. Bertrand Michel	Ecole Centrale de Nantes	Co-directeur
M. Jérôme Morio	ONERA	Rapporteur
M Grégory Nuel	Sorbonne Université	Examinateur
M Philippe Saint-Pierre	Université Paul Sabatier	Co-directeur
M Roman Sueur	EDF R&D	Encadrant

Laboratoire de Probabilités, Statistique et Modélisation (LPSM)
Université Pierre et Marie Curie
Boîte 158, Tours 15-25, 2ème étage
4 place Jussieu
75252 Paris Cedex 05

Département Performance, Risque Industriel, Surveillance pour la Maintenance et l'Exploitation (PRISME)
EDF Recherche & Développement
6, quai Watier
78401 Chatou Cedex

Remerciements

Mes premiers remerciements vont à mes directeurs de thèse Gérard, Bertrand et Philippe qui m'ont toujours soutenu depuis le début de cette thèse. Je vous remercie pour votre investissement, votre bienveillance et vos connaissances sans lesquelles cette thèse n'aurait jamais pu aboutir. Malgré la distance et vos nombreuses responsabilités, vous avez toujours su vous rendre disponible pour échanger, me conseiller et surtout m'encourager. J'ai eu l'immense privilège de vous avoir en tant que directeurs et je garderai de très bons souvenirs de ces dernières années. J'espère avoir l'occasion de continuer à travailler avec vous dans le futur.

Je tiens à remercier très chaleureusement Roman, Nicolas et Bertrand I. pour leurs encadrements à PRISME durant ces dernières années. Je vous remercie la confiance que vous m'avez accordée depuis mon stage. J'ai découvert grâce à vous le monde de la recherche en entreprise et j'y ai pris goût. J'ai surtout énormément appris à vos côtés, autant humainement qu'intellectuellement. J'ai eu la chance d'avoir une équipe encadrante comme la vôtre et je ne saurais jamais vous remercier assez.

Je remercie également Jérôme Morio et Jean-Michel Poggi pour m'avoir fait l'honneur de rapporter ma thèse. Merci pour votre lecture attentive, vos remarques constructives et pour l'intérêt que vous portez à mes travaux. Je remercie également Véronique Maume-Deschamps et Grégory Nuel pour avoir accepté de faire partie de mon jury.

Je tiens à remercier tous mes anciens collègues de PRISME. Merci aux mes « Mesh bros » Xav et Tuan ainsi que Jo pour ces nombreuses années, discutions, voyages, sorties, etc. Ces années n'auraient pas été les mêmes sans vous. Je souhaite une grande réussite à Pablo, Mathieu et Jérôme pour leurs thèses à PRISME. Merci aux anciens doctorants Guillaume, Jeanne, Thomas, Vincent et Mathieu C. Merci bien sûr à tous les autres collègues d'EDF pour ces déjeuners quotidiens dans la cuisine, ces points hebdos, ces conférences, ces sorties, ces pique-niques au fond du site, ces soirées folles, ces voyages, ces karaokés, etc. J'ai vécu de folles aventures avec vous et il me faudrait une page entière pour tous vous citer un par un.

J'adresse également mes remerciements à tous les membres du LPSM qui m'ont accompagné durant cette thèse. Merci à mes co-bureaux du 204, Taieb, Lucie, Nicolas, Adeline et Ugo. Merci aux autres doctorants (ou jeunes docteurs), Félix, Panayotis, Yohann, Émilie, Simon, Qiming. Merci aussi aux anciens, Erwan, Baptiste, Moktar, Matthieu, Assia, Roxane, Quyen. Je remercie aussi tous les membres du labo. Je remercie très chaleureusement Corinne et Louise pour m'avoir aidé à organiser cette soutenance et surtout pour leurs gentillesses, leurs disponibilités

et leurs efficacités depuis mon arrivée au laboratoire. Merci aussi à Mathieu Yves qui m'a permis d'imprimer ce manuscrit.

Merci aussi à mes nouveaux collègues de QRT. J'ai eu la chance de tomber dans une équipe extraordinaire avec qui j'apprends énormément et où chaque jour travaillé est un vrai plaisir. J'espère bien sûr continuer à vous battre au babyfoot.

Je remercie bien évidemment mes amis, la mif, les enfants de Belleville, ceux avec qui j'ai grandi et partagé toutes ces aventures. Dédicace à Hich, Tartare, Wilo, Tony, Drogo, HatHat, Kevin, Fayçal, Mourad, Yuv, Brams, Nad, Leo, Zac et Krim. On en a fait du chemin jusqu'ici ! Merci aussi à Mohamed, Florent, Kamo, Andres, Kevin D., Sarah, Léa, Keke, Jo et Hélène. Sans oublier mes amis du CEMRACS, Kévin, Houssam, Mohamed, Alexandre, Nordine et tous les autres. Un grand merci aussi à Yasmine pour sa reLectUre attentiVe du manUscrit.

Enfin, je remercie surtout ma famille. Merci à mes parents pour tous leurs sacrifices et leurs soutiens inconditionnels dans chacun de mes choix. Merci à mes grandes soeurs qui ont toujours su me conseiller, me soutenir et me guider. Tout n'a pas toujours été facile, mais vous m'avez toujours facilité la tâche et si j'en suis là aujourd'hui c'est surtout grâce à vous.

Table des matières

1 Introduction **7**
- 1.1 Contexte général . . . 7
- 1.2 Cas d'étude industriel . . . 10
- 1.3 Contexte statistique et modélisation . . . 13
- 1.4 Organisation du manuscrit et présentation des contributions . . . 17

2 State of the Art **23**
- 2.1 Global Sensitivity Analysis with independent inputs . . . 23
- 2.2 Global Sensitivity analysis with dependent inputs . . . 33
- 2.3 Modeling Dependencies . . . 38
- 2.4 Surrogate models . . . 44

3 Sensitivity Analysis and Random Forest **49**
- 3.1 Introduction . . . 50
- 3.2 New permutation variable importance values and their relations with the Sobol' indices . . . 52
- 3.3 Sobol indices estimation and comparison . . . 54
- 3.4 Using Random Forest . . . 64
- 3.5 Conclusion . . . 72
- 3.6 Appendix . . . 73

4 Shapley effects for sensitivity analysis **79**
- 4.1 Introduction . . . 80
- 4.2 Sobol' sensitivity indices . . . 81
- 4.3 Shapley effects . . . 86
- 4.4 Examples in Gaussian framework: analytical results and relations between indices 89
- 4.5 Numerical studies . . . 95
- 4.6 Kriging metamodel with inclusion of errors . . . 100
- 4.7 Numerical simulations with kriging model . . . 104
- 4.8 Conclusion . . . 107
- 4.9 Appendix . . . 109

5 Worst-case dependence structures **111**
- 5.1 Introduction . . . 112
- 5.2 Minimization of the quantile of the output distribution . . . 114
- 5.3 A preliminary study of the copula influence on quantile minimization . . . 120
- 5.4 Quantile minimization and choice of penalized correlation structure . . . 124

5.5 Applications 131
5.6 Conclusion and discussion 137
5.7 Proof of the consistency result 138

6 Conclusion **151**
6.1 Summary and main contributions 151
6.2 Perspectives 152

Bibliography **155**

Chapitre 1

Introduction

Sommaire

1.1 Contexte général **7**
1.2 Cas d'étude industriel **10**
1.3 Contexte statistique et modélisation **13**
1.3.1 Analyse de Sensibilité à structure de dépendance connue 14
1.3.2 Mesure conservative du risque à structure de dépendance inconnue 15
1.4 Organisation du manuscrit et présentation des contributions **17**
1.4.1 Chapitre 2 : État de l'art 17
1.4.2 Chapitre 3 : Analyse de sensibilité à partir des forêts aléatoires dans le cas de variables corrélées 17
1.4.3 Chapitre 4 : Shapley effects avec prise en compte des erreurs de krigeage et de Monte Carlo 19
1.4.4 Chapitre 5 : Évaluation pénalisante d'un critère de risque lorsque la structure de dépendance est inconnue 20

1.1 Contexte général

La fiabilité est usuellement définie comme "l'aptitude d'un système à accomplir une fonction requise, dans des conditions données et pendant une durée donnée" [Lemaire et al., 2005]. Dans un contexte industriel, cette notion de fiabilité peut être formalisée de diverses manières, et est reliée à plusieurs enjeux distincts, notamment pour la performance de l'exploitation, la qualité des produits, la sécurité et la sûreté des procédés mis en œuvre et l'acceptabilité des activités industrielles. Pour un producteur électrique comme EDF (Électricité de France), une fiabilité insuffisante se traduit d'abord par une indisponibilité des moyens de production et donc une perte de production. À terme, celle-ci peut mener à une perte de valeur de l'actif industriel si un niveau de fiabilité suffisant n'est pas atteint. Mais pour EDF, qui exploite des installations pouvant générer des risques industriels, l'un des enjeux associé à la fiabilité de ses moyens de

production est celui de la sûreté nucléaire. En particulier, les installations nucléaires font l'objet d'une réglementation contraignante et d'un contrôle strict de la part de l'Autorité de Sûreté Nucléaire (ASN). Investir dans des programmes de maintenance efficaces est alors un des principaux leviers permettant de garantir un haut niveau de sûreté du processus de production. Cependant, l'amélioration de l'efficacité des opérations de maintenance et l'augmentation de la fréquence des maintenances préventives ne sont possibles que pour des composants réparables ou remplaçables. Or, la maîtrise de la sûreté requise de la part d'EDF inclut également, et avant tout, la garantie du maintien de l'intégrité des trois barrières de confinement de la radioactivité [Zinkle and Was, 2013; Commission et al., 1975]. En particulier, certaines structures passives incluant des composants massifs et considérées comme non remplaçables participent au confinement indispensable des éléments ionisants, aussi bien en situation normale d'exploitation qu'en situation accidentelle. L'entreprise EDF est alors tenue de justifier, par des méthodes adéquates, la sûreté de ces composants pour toute la durée prévue d'exploitation des tranches nucléaires. Cette démonstration de sûreté conditionnant l'autorisation d'exploitation de la part de l'ASN est porteuse d'un enjeu majeur de durée de vie des moyens de production.

L'occurrence d'événements accidentels pouvant mener à une éventuelle défaillance de composants de ce type est elle-même extrêmement rare. Il est alors très probable qu'aucune défaillance n'ait été observée depuis la mise en service du composant. L'ingénieur fiabiliste, chargé d'étudier ce risque, ne dispose donc bien souvent d'aucun retour d'expérience permettant une évaluation de la fréquence d'occurrence de l'événement indésirable. La fiabilité des structures est le domaine des sciences de l'ingénieur qui a précisément pour objet de fournir au fiabiliste les méthodes et outils permettant d'évaluer ce risque en l'absence de données statistiques portant directement sur l'événement. Dans ce cadre, le risque de défaillance d'une structure, par exemple le risque de rupture d'un barrage hydroélectrique à la suite d'une crue exceptionnelle, est formalisé et calculé sous forme d'une marge entre la valeur d'une grandeur physique y caractérisant l'intégrité du composant, et un critère de sûreté $s < y$ sur cette grandeur. Ce critère est fixé de manière réglementaire par l'ASN de telle manière que son respect garantisse l'absence de danger lié à une éventuelle défaillance. Par ailleurs, la méthode d'évaluation du risque (calcul de la marge $y - s$) doit elle-même être validée par cette autorité réglementaire.

Le comportement de la structure, dans une situation jamais rencontrée en pratique et ne pouvant faire l'objet d'une expérimentation directe, ne peut être étudié que par le recours à des simulations numériques. Un code de calcul, élaboré par des spécialistes de la physique du composant étudié, fournit une implémentation d'un modèle mathématique, défini par une fonction η représentant la physique sous-jacente au phénomène (par exemple la sollicitation exceptionnelle générée par la crue sur l'ouvrage en béton). Ce code a pour variables d'entrée certaines caractéristiques du composant et de la situation accidentelle étudiée, et pour sortie la grandeur physique y. La méthode requise par les autorités consiste à réaliser des simulations en fixant ces d variables d'entrées $x_1, x_2, \ldots, x_d$ à des valeurs pessimistes par rapport à la réalité, afin de garantir une évaluation conservative de la grandeur physique $y = \eta(x_1, \ldots, x_d)$ et donc de la marge au critère réglementaire.

Cette pratique requiert une validation du caractère conservatif du modèle η lui-même ainsi que des hypothèses sur lesquelles il repose et incite souvent le praticien de la fiabilité à choisir des modélisations simplifiées à même de "couvrir" un grand nombre de cas. Cela permet une

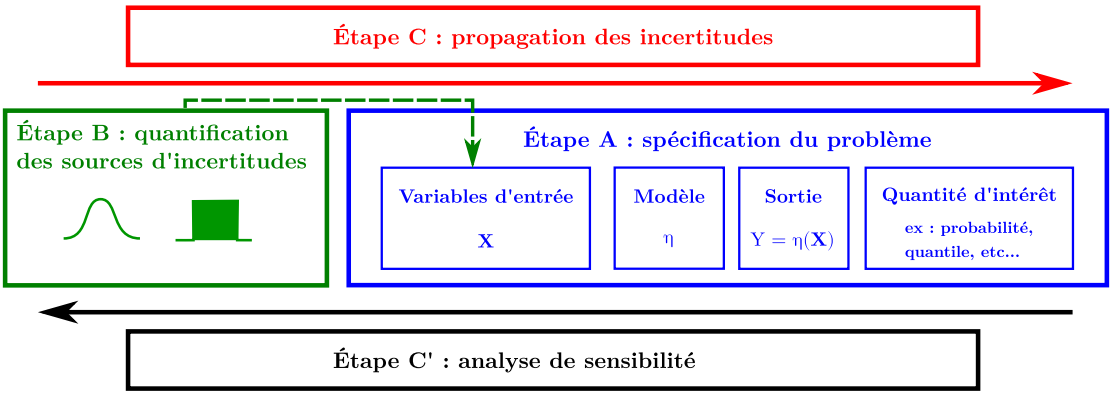


Fig. 1.1 *Méthodologie globale de traitement des incertitudes.*

approximation de la réalité jugée plus sévère au regard des critères réglementaires que chacune des situations particulières ainsi "couvertes". En résulte le calcul d'un majorant du risque réel ou de manière équivalente, une borne inférieure pour la marge existant entre la sortie y et le seuil imposé. On peut qualifier cette démarche d'"hypothético-déductive" au sens où elle consiste à *prouver*, à partir d'hypothèses conservatives, l'existence d'une marge non nulle entre la réalité (non connue) et une potentielle situation à risque (cas $y \leq s$). Cette méthode a l'avantage évident de ne nécessiter qu'un unique calcul de risque pour une situation accidentelle donnée. Des analyses de sensibilité dites "paramétriques" sont souvent réalisées dans le but d'acquérir davantage de renseignements sur l'influence des variables d'entrée du modèle. Pour autant, celles-ci ne donnent lieu qu'à un petit nombre de calculs supplémentaires dans la mesure où c'est chaque variable d'entrée qui est individuellement fixée à des valeurs différentes autour d'une référence. De plus, le recours à des modèles simplifiés facilite les échanges avec les spécialistes de la physique sous-jacente au risque étudié et permet une interprétation aisée des résultats.

En contrepartie, l'éclairage apporté sur le risque réel est très partiel et ne permet pas directement de proposer des leviers de réduction du risque ou une analyse plus approfondie de l'influence des variables du problème. Pour ces raisons, il peut être souhaitable de recourir à une approche probabiliste permettant d'intégrer à l'étude certaines incertitudes du problème. La figure 1.1 représente de manière schématique les différentes étapes d'une étude d'incertitudes, dont la première consiste à spécifier le problème, choisir un modèle numérique et une quantité d'intérêt en sortie de ce modèle (étape A). Les variables d'entrées considérées comme incertaines sont alors modélisées par des variables aléatoires $X_1, X_2, \ldots, X_d$, comme indiqué à l'étape B. Le calcul de la loi de $Y = \eta(\mathbf{X})$ induite par la loi du vecteur aléatoire $\mathbf{X} = (X_1, X_2, \ldots, X_d)$ et le modèle η est appelée *propagation d'incertitudes* (étape C). En général, on ne cherchera pas à caractériser de manière exhaustive la loi de Y, mais on s'intéressera à un indicateur probabiliste de fiabilité défini à partir de cette loi. Par exemple, on s'intéressera à la probabilité de ne pas respecter le critère réglementaire $\mathbb{P}[Y \leq s]$ ou au quantile $Q(\alpha)$ d'ordre α de la loi de Y. Dans le cas du quantile, celui-ci peut-être considéré comme une estimation conservative de la vraie valeur de y avec un niveau de confiance α vis-à-vis des incertitudes connues sur les entrées du modèle. Ce dernier indicateur est souvent utilisé dans le cadre des méthodes appelées BEPU (*Best Estimate Plus Uncertainties*) [D'Auria et al., 2012; Wilson, 2013]. Celles-ci consistent à effectuer un premier calcul de référence à partir d'hypothèses les plus réalistes possible (calcul

"best estimate") auquel on ajoutera la donnée du quantile estimé par une méthode adéquate de propagation d'incertitudes. À la suite d'une étude de type BEPU, il est souvent opportun d'ajouter des analyses de sensibilité qui, selon la méthode choisie, permettront d'objectiver le choix d'éliminer certaines entrées probabilistes et simplifier ainsi le problème de propagation d'incertitudes, ou de quantifier l'influence globale des différentes variables d'entrée sur la sortie Y (étape C' de la figure 1.1) [Saltelli et al., 2008].

Ces méthodes sont destinées à prendre en compte et quantifier les incertitudes sur les grandeurs d'intérêt du problème (risque de non-respect des critères réglementaires). Elles sont plus coûteuses en ressources de calcul et font appel à des outils statistiques souvent plus complexes que ceux habituellement utilisés par les ingénieurs spécialistes de la physique du composant étudié. Elles peuvent être néanmoins d'un grand secours pour la prise de décision lorsque la marge au critère réglementaire $y - s$ est faible. Elles permettent, en effet, une meilleure connaissance des proportions dans lesquelles les différents niveaux de risque sont répartis en fonction des situations couvertes. Elles autorisent ainsi une quantification plus fine du risque réel et peuvent même être employées comme un outil d'exploration d'un code numérique sur le domaine de variation des entrées. Ce dernier usage est, en particulier, d'un intérêt croissant lié à l'utilisation de codes de simulation numérique de plus en plus complexe et réaliste fournissant des résultats qu'un ingénieur, même spécialiste du domaine, peut alors avoir du mal à analyser. Les méthodes de propagation et de traitement d'incertitudes, en se fondant sur une approche que l'on pourra qualifier d'"inductive" (par comparaison avec la méthode déterministe) fournissent de manière générale une palette d'outils permettant de prendre en compte les incertitudes, de caractériser leur impact sur une quantité d'intérêt et apporte ainsi un appui pour la construction d'un argumentaire rigoureux de justification de la sûreté d'une installation.

1.2 Cas d'étude industriel

Ce travail de thèse est en grande partie motivé par les besoins d'études liées à la cuve des réacteurs nucléaires d'EDF. Ce composant en acier forgé est un élément du circuit primaire des réacteurs à eau sous pression (REP) contenant le cœur du réacteur nucléaire et constitue la seconde barrière de confinement des éléments radioactifs. À ce titre, le maintien de son intégrité en toutes circonstances est essentiel pour la sûreté. En particulier, EDF est tenu de justifier l'absence de risque d'amorçage d'une fissure au niveau de potentiels petits défauts situés dans la paroi du composant. L'éventuelle présence de défauts dans l'épaisseur de l'acier est liée à la dépose d'un revêtement en inox en surface interne de la cuve, afin de protéger la structure en acier noir du milieu très corrosif présent dans le circuit primaire dans les conditions normales de fonctionnement du réacteur. Divers phénomènes physico-chimiques liés au procédé de fabrication du composant peuvent amener des micro fissures à se former lors de cette étape de dépose par soudure du revêtement en acier inoxydable. Le risque qu'on étudie ici est celui de la rupture de l'acier de cuve en cas de choc froid pressurisé. Ce type d'évènement serait amené à se produire en cas d'accident de type APRP (accident de perte de réfrigérant primaire), c'est-à-dire en cas de formation d'une brèche dans le circuit primaire du réacteur. La nécessité de maintenir le refroidissement du combustible requerrait alors l'injection d'eau froide dans la cuve provoquant un choc thermique avec le matériau initialement à température nominale (environ $300°$C) d'exploitation

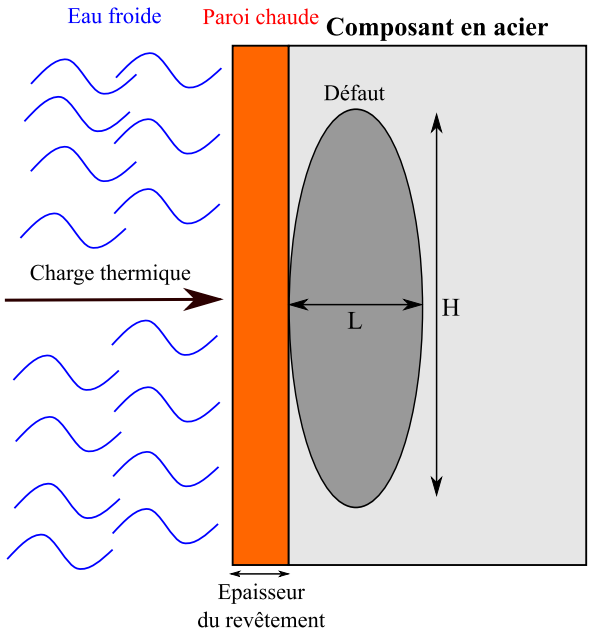


Fig. 1.2 *Illustration d'une charge thermique générée par le contact entre de l'eau à basse température et une paroi à haute température, exposée à un défaut de fabrication.*

avant déclenchement de l'accident considéré. La figure 1.2 illustre ce phénomène. La résistance du matériau à une telle sollicitation n'est pas constante au cours du temps. Le cœur du réacteur nucléaire produisant l'énergie récupérée pour la génération électrique est contenu dans la cuve et provoque une irradiation neutronique importante sur la partie centrale du composant. Cette irradiation a pour effet de diminuer progressivement la résistance de l'acier de cuve à l'amorçage d'une fissure. La durée de vie de la tranche est donc directement liée à cette altération des propriétés mécaniques du matériau.

Il apparaît que l'évaluation d'un tel risque est avant tout conditionnée par l'impossibilité de se fonder sur l'observation directe des phénomènes en jeu. Aucun événement de type choc froid n'est survenu sur un composant tel que la cuve sur le parc nucléaire d'EDF ou d'un autre exploitant de réacteurs similaires. Reproduire expérimentalement ce type de phénomène serait par ailleurs extrêmement complexe et coûteux et donc pratiquement irréalisable. On a donc recours à la simulation numérique pour étudier le phénomène de choc froid et ses éventuelles conséquences sur l'intégrité de la cuve. Cependant, la modélisation à même de représenter correctement ces phénomènes est affectée d'incertitudes. Celles-ci concernent les paramètres physiques décrivant les trois principaux facteurs amenant la potentialité d'un risque : occurrence d'un transitoire sollicitant fortement le matériau, existence d'un petit défaut créé lors de la fabrication de la pièce d'acier revêtu et fragilisation de l'acier au cours de la vie en service de la tranche sous l'effet de l'activité du cœur.

La relative complexité de la modélisation, qui met en jeu différentes physiques (notamment thermiques et mécaniques), limite la possibilité d'explorer formellement le code de calcul (par exemple par différentiation automatique [Rall, 1981]) qui est considéré ici comme une boîte noire.

Nous cherchons alors à caractériser la loi de Y en fonction de celles des entrées incertaines du modèle selon la démarche décrite supra (figure 1.1).

Le modèle Le risque d'amorçage est évalué à travers le calcul d'une valeur de résistance R, d'une part, et d'une sollicitation S d'autre part. Ces grandeurs sont définies pour chaque instant $t \in [0, T]$, où T est un temps maximal de simulation. La résistance du matériau dépend des caractéristiques de celui-ci, de son état de vieillissement, de la teneur en éléments d'alliage de l'acier et de sa résistance initiale avant irradiation. La sollicitation dépend, quant à elle, du type de transitoire accidentel simulé, de la géométrie du défaut et des propriétés de l'acier vis-à-vis de la propagation thermique dans la paroi et des contraintes mécaniques appliquées au niveau du défaut.

Critère de sortie Le critère retenu pour évaluer la sûreté vis-à-vis de ce risque est appelé facteur de marge et défini comme la valeur minimale du rapport entre la résistance et la sollicitation :

$$y = \min_{t \in [0,T]} \frac{R(t)}{S(t)}.$$

L'évènement $y < 1$ est équivalent à l'existence d'un instant au cours du transitoire pour lequel la sollicitation excède la résistance de l'acier. Au contraire, l'intégrité de la structure est assurée tant que $R(t) > S(t)$ à tout instant t. Le risque est jugé d'autant plus faible, que la marge entre y et le seuil $s > 1$ est importante.

La quantité d'intérêt Nous nous intéresserons en particulier au quantile $Q(\alpha)$ définissant la valeur de y associée à un certain niveau de sûreté, ou encore à la probabilité de défaillance $\mathbb{P}[Y \leq s]$.

Les variables d'entrée Une analyse préalable amène à considérer formellement l'incertitude sur six des variables d'entrée du modèle numérique :

- la dimension d'un défaut (considéré de forme semi-elliptique) : hauteur et largeur,
- la position d'un défaut : position verticale et azimutale en surface interne du matériau de base,
- le décalage de température de transition fragile-ductile résumant l'effet de l'irradiation sur le matériau de base,
- la ténacité de la paroi représentant la propension du matériau à résister à l'amorçage d'une fissure sous l'effet d'une contrainte.

La ténacité est intrinsèquement variable du fait du caractère aléatoire du phénomène de rupture.

Problématique L'incertitude sur ces variables d'entrée est définie à l'aide d'une loi de probabilité associée à chacune d'entre elles. Ces lois sont généralement issues de mesures par contrôle

non destructif réalisées en site, d'essais mécaniques réalisés en laboratoire sur des échantillons de matériaux similaires à celui du composant ou encore d'avis d'experts. Quelle qu'en soit l'origine, les connaissances relatives à ces grandeurs physiques incertaines concernent en général une seule de ces variables considérée indépendamment des autres. Dans le cas où des données sont disponibles et peuvent être utilisées comme échantillon statistique d'une des lois d'entrée du modèle, cette loi est le plus souvent connue indépendamment des valeurs des autres entrées malgré leur variabilité également postulée. En d'autres termes, l'éventuelle dépendance entre les entrées incertaines du modèle n'est a priori pas connue. Néanmoins, toute représentation probabiliste du vecteur des entrées comporte une hypothèse, éventuellement implicite, concernant la structure de dépendance entre ces entrées. Faute d'une sensibilisation des ingénieurs et experts du problème physique, cette question de la dépendance tend à être négligée, voire ignorée. Les variables représentant des grandeurs physiques distinctes, déterminées par différents phénomènes et mesurées séparément, sont souvent considérées comme "indépendantes" sans que ce terme recouvre de manière évidente la notion *d'indépendance statistique*. De manière également fréquente, c'est pour des raisons pratiques que l'hypothèse d'indépendance entre les lois des entrées est postulée. En l'absence de données jointes pour les entrées incertaines, de connaissances fines sur les liens qui peuvent exister entre les différentes sources d'incertitude et d'un corpus méthodologique dédié à cette question importante, la spécification du vecteur des entrées incertaines et de sa loi de probabilité se résume en général à la donnée d'un modèle univarié pour chacune des marges du vecteur.

La structure de dépendance définie pour la loi jointe du vecteur des variables d'entrée du code de simulation peut avoir un impact significatif sur les résultats de l'étude et doit, dans le cadre d'une démarche de justification auprès des autorités réglementaires, être explicitée et étayée de manière rigoureuse. L'objet du travail de thèse présenté ici est de proposer des outils méthodologiques pour pallier cette difficulté.

1.3 Contexte statistique et modélisation

Soit $\mathbf{X} = (X_1, \ldots, X_d) \in \mathbb{R}^d$ le vecteur aléatoire représentant les entrées incertaines du modèle, $Y = \eta(\mathbf{X}) \in \mathbb{R}$ la variable aléatoire de sortie du code représenté par la fonction η, et $\mathscr{C}(Y)$ la quantité d'intérêt de l'étude sur la distribution de Y. L'expression générale de η est en pratique inconnue, cette fonction étant alors considérée comme une 'boîte noire". Nous supposons cependant que pour tout $\mathbf{x} \in \mathbb{R}^d$, la quantité $\eta(\mathbf{x})$ peut être calculée moyennant un certain coût en termes de ressources de calcul. La distribution de probabilité de $\mathbf{X}$ peut être définie en distinguant les effets marginaux de chaque variable et leur structure de dépendance. La dépendance est formellement définie par la notion de copule [Sklar, 1959] permettant, avec les distributions marginales, de décrire exhaustivement la loi jointe entre plusieurs variables aléatoires. Une copule est une fonction de répartition multivariée C faisant le lien entre les distributions marginales $F_1, \ldots, F_d$ et la distribution jointe F de $\mathbf{X}$, comme décrit par le théorème de Sklar [Sklar, 1959] :

$$F(\mathbf{x}) = C\left(F_1(x_1), \ldots, F_d(x_d)\right).$$

Les problèmes posés au fiabiliste en matière de prise en compte d'une éventuelle dépendance entre les entrées incertaines peuvent être de deux natures différentes selon le cas de figure. Dans le cas où une dépendance existe et est connue tel que l'on dispose d'une modélisation faisant consensus, il peut toutefois être délicat d'appliquer telles quelles les méthodes usuelles de traitement d'incertitudes et d'analyses de sensibilité. On désignera ce premier contexte comme une situation de *dépendance connue*. La difficulté ne vient alors pas tant de l'estimation du risque qui peut se faire de manière analogue au cas où les entrées seraient indépendantes, à partir des nombreuses méthodes de propagation d'incertitudes pouvant s'appliquer aisément à ce contexte. En revanche, la difficulté intervient en particulier à l'étape d'analyse de sensibilité. En effet, la dépendance des variables rend difficile l'interprétabilité des résultats puisqu'il n'est pas possible pour les méthodes "classiques" de dissocier les effets marginaux, d'interaction et de dépendance. L'utilisation de méthodes d'analyse de sensibilité non adaptées aux variables dépendantes peut significativement altérer les résultats de l'étude et fausser leur interprétation.

Dans le cas où la dépendance n'est pas connue, mais que ne peut être écartée l'hypothèse de « non-indépendance » entre certaines entrées, il est malgré tout nécessaire de tenir compte de cet aspect de la modélisation probabiliste du problème de fiabilité. Dans cette situation, qu'on appellera désormais *cas d'une dépendance inconnue*, la quantification initiale du risque (par propagation des incertitudes appelée étape C dans la figure 1.1) peut poser problème en elle-même. Il peut alors s'avérer judicieux de tenter, par un calcul adapté à une telle démarche, de couvrir le risque associé au manque de connaissance sur la copule.

Ces deux cas de figure distincts (dépendance connue et inconnue) renvoient à des problèmes statistiques qui doivent eux-mêmes être distingués. Les deux principaux thèmes traités dans ce manuscrit sont donc, d'une part, l'analyse de sensibilité pour variables dépendantes et d'autre part l'évaluation d'un risque de fiabilité lorsque la dépendance est inconnue. Ces deux sujets sont présentés plus en détail dans les prochains paragraphes.

1.3.1 Analyse de Sensibilité à structure de dépendance connue

Ce premier cas de figure suppose que la structure de dépendance, définie par la copule C, est connue. Ainsi, la distribution de $\mathbf{X}$ est correctement définie et la quantification du risque peut être établie de façon rigoureuse. De nombreuses méthodes de propagation d'incertitudes permettent d'estimer la quantité d'intérêt $\mathscr{C}(Y)$, même dans le cas de variables aléatoires dépendantes (par exemple les méthodes du type Monte-Carlo). Cependant, ce panel de méthodes est plus restreint pour l'étape d'analyse de sensibilité et peut-être divisé en 3 grandes classes. Les méthodes de criblage, les méthodes locales et les méthodes globales. Les méthodes de criblage étudient les aspects qualitatifs de la sensibilité. Leur objectif est de déterminer les variables d'entrée ayant un effet sur la sortie du modèle, sans en déterminer l'importance. Elles sont généralement utilisées pour réduire le nombre de variables dans les problèmes de grandes dimensions. Ces méthodes ont l'avantage d'être simples à implémenter et peu coûteuses, mais sont d'usage limité. En effet, elles n'apportent qu'une information qualitative sur l'importance des variables et ne prennent pas en compte les effets d'interactions parfois prédominants dans un modèle. Les méthodes locales sont des techniques quantitatives visant à étudier l'impact d'une perturbation autour d'une valeur

nominale $\mathbf{x} \in \mathbb{R}^d$. Elles se basent principalement sur le calcul des dérivées partielles $\frac{\partial \eta}{\partial x_j}(\mathbf{x})$ de chaque composante $j = 1, \ldots, d$. Comparer ces dérivées entre elles permet alors de quantifier leurs influences sur la sortie autour de $\mathbf{x}$. Cependant, l'analyse de ces méthodes est limitée et n'explore pas l'ensemble du domaine de variance des entrées et de surcroît, nécessite la dérivabilité du modèle, ce qui n'est pas toujours le cas en pratique. Les méthodes globales considèrent les variables d'entrée et sortie comme aléatoires. Elles sont plus générales et explorent l'ensemble du domaine de variation des variables. Dans cette thèse, nous nous intéressons en particulier à cette classe de méthodes d'analyse de sensibilité.

De nombreuses sous-classes de méthodes globales existent et se différencient en partie sur la quantité d'intérêt ciblée par l'analyse. Les méthodes les plus répandues sont basées sur l'analyse de la variance, mais d'autres méthodes considèrent différentes mesures [Fort et al., 2016], voire la distribution complète [Borgonovo, 2007]. Nos travaux se concentrent sur les méthodes basées sur la variance. L'objectif est de quantifier comment la variabilité des entrées influe sur la variabilité de la sortie. Ces méthodes prennent leur origine sur l'analyse de la variance (ANOVA) [Fisher, 1925] et plus précisément la décomposition fonctionnelle ANOVA qui consiste à exprimer la variance globale en une somme de variances partielles, composée d'effets principaux et d'interactions. Cependant, une infinité de décompositions existe si aucune condition n'est imposée à ces composantes. Une décomposition unique peut être obtenue en posant des conditions d'orthogonalité entre les composantes. L'indépendance des variables aléatoires d'entrée rend les termes orthogonaux deux à deux et permet alors l'unicité de la décomposition ANOVA. Ainsi, ces composantes correspondent aux indices de Sobol [Sobol, 1993], très utilisés en analyse de sensibilité globale.

Néanmoins, lorsque l'hypothèse d'indépendance n'est pas vérifiée, l'unicité de la décomposition n'est alors plus assurée. Un problème modélisé par plusieurs variables dépendantes offre un cadre plus restrictif de solutions pour l'analyse de sensibilité. De nombreuses méthodes tentent d'étendre l'utilisation des indices de Sobol aux cas de variables dépendantes [Li and Rabitz, 2010; Chastaing et al., 2012]. Dans un contexte similaire, l'utilisation des indices de Shapley peut aussi permettre de traiter la dépendance dans le cas de variables dépendantes [Owen, 2014; Iooss and Prieur, 2017]. Cependant, les indices généralisés et les indices de Shapley demandent un grand nombre d'évaluations du modèle η et sont alors difficilement utilisables lorsque celui-ci est coûteux. Les modèles d'apprentissage tels que les forêts aléatoires [Breiman, 2001] ou le krigeage [Sacks et al., 1989] permettent de réduire ce coût en passant par des modèles de substitution, mais avec une maîtrise de l'erreur d'estimation nettement plus difficile. C'est dans ce contexte que se placent nos contributions visant à étendre les méthodes d'estimations d'indices de sensibilité dans le cas de variables dépendantes à partir de modèles d'apprentissage.

1.3.2 Mesure conservative du risque à structure de dépendance inconnue

Dans une étude de fiabilité, le risque peut être quantifié par différentes quantités d'intérêts $\mathscr{C}(Y)$. Deux mesures du risque sont souvent considérées en pratique : la probabilité de dépassement de seuils (ou probabilité de défaillance) et le quantile. La probabilité de défaillance représente le risque d'avoir la sortie du modèle inférieure (ou supérieure) à un seuil réglementaire t et est

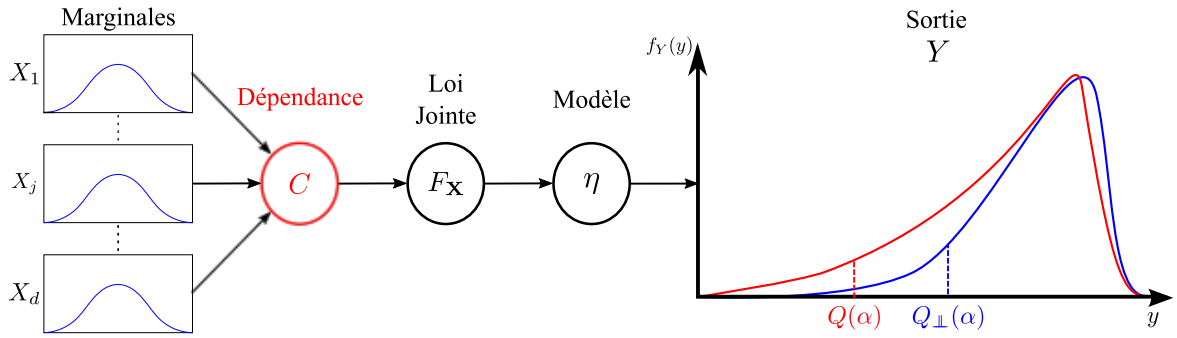


Fig. 1.3 *Propagation de l'incertitude de* $\mathbf{X}$ *par le modèle* η.

définie par

$$\mathbb{P}[Y \leq t] = G(t),$$

où G est la fonction de répartition de Y. Le quantile, quant à lui, représente la valeur de y associée à une probabilité $\alpha \in (0,1)$ et est défini par la fonction inverse généralisée de G :

$$Q(\alpha) = G^{-1}(\alpha) = \inf\{y : G(y) \geq \alpha\}. \tag{1.1}$$

Dans ces travaux, nous choisissons l'utilisation du quantile $Q(\alpha)$ comme quantité d'intérêt pour la quantification du risque. Néanmoins, ces travaux peuvent être appliqués à toute autre mesure $\mathscr{C}(Y)$.

Nous nous plaçons dans le cas où la structure de dépendance, définie par C, est inconnue. Supposer l'indépendance en utilisant seulement l'information issue des marginales peut avoir des répercussions négatives pour la quantification du risque de l'étude de fiabilité comme le montre la figure 1.3 qui illustre la propagation des incertitudes de $\mathbf{X}$ au travers du modèle η. Nous représentons en rouge et en bleu les distributions de Y respectivement lorsqu'une copule C et l'hypothèse d'indépendance sont considérées. Le quantile de Y associé à la copule C et celui associé à l'indépendance des variables d'entrée sont respectivement définis par $Q(\alpha)$ et $Q_{\perp\!\!\!\perp}(\alpha)$. Dans cette situation, la copule C a une influence sur la sortie et les deux distributions de Y diffèrent significativement avec notamment $Q(\alpha) < Q_{\perp\!\!\!\perp}(\alpha)$. Cela illustre que supposer l'indépendance des variables, sans justifications rigoureuses, peut mener vers une sous évaluation du risque.

L'effet de la dépendance a été montré comme influent en fiabilité industrielle. Notamment les travaux de Grigoriu and Turkstra [1979] et Thoft-Christensen and Sørensen [1982] qui montrèrent, pour un cas simple linéaire, que la corrélation entre des variables d'entrée pouvait influencer la probabilité de défaillance d'un système et l'augmenter significativement dans les pires cas. D'autres études [Tang et al., 2013, 2015] ont aussi considéré des structures de dépendance différentes du cas gaussien et montrèrent que l'influence peut être nettement plus importante.

Connaitre l'influence de C sur la sortie du modèle permettrait d'aider à justifier l'hypothèse d'indépendance. Cependant, une pratique plus conservative serait de surévaluer le risque, quitte à faire des hypothèses très pessimistes sur la structure de dépendance. Sachant les distributions marginales connues et fixées, l'objectif est alors de déterminer, parmi un ensemble $\boldsymbol{\mathcal{C}}$ de copules

réalistes, la valeur du quantile $Q(\alpha)$ la plus pessimiste tel que :

$$Q^*(\alpha) = \min_{C \in \mathcal{C}} Q_C(\alpha), \tag{1.2}$$

où Q_C est le quantile de Y associé à une copule C de $\mathbf{X}$. C'est dans ce cadre que se placent nos contributions et vise à proposer une méthodologie plus conservative de propagation des incertitudes et de quantification du risque, lorsque la structure de dépendance est inconnue.

1.4 Organisation du manuscrit et présentation des contributions

Ce travail de thèse est structuré en quatre chapitres. Le chapitre 2 propose une revue de la littérature des méthodes et des outils mathématiques nécessaires aux contributions proposées dans les chapitres suivants. Les deux chapitres suivants sont consacrés à la situation de dépendance connue. Le chapitre 3 est un travail exploratoire autour des mesures d'importance par permutation des forêts aléatoires, de leurs liens avec l'analyse de sensibilité et du traitement de ces mesures dans le cas de données dépendantes. Il vise principalement à proposer de nouvelles mesures adaptées au cas de données dépendantes. Le chapitre 4 porte au contraire sur une méthode spécifiquement développée pour les modèles à entrées dépendantes : les indices de sensibilité basés sur des valeurs de Shapley [Shapley and Shubik, 1954]. Ce chapitre présente une méthode d'estimation de ces indices par krigeage et une prise en compte des différentes erreurs d'estimations liées à cette méthode. Il a fait l'objet d'un article, coécrit avec Kévin Elie-Dit-Cosaque et publié dans la revue ESAIM Proceedings [Benoumechiara and Elie-Dit-Cosaque, 2019]. Le chapitre 5 traite, lui, du cas désigné sous le terme de dépendance inconnue. Il porte sur l'évaluation conservative d'un risque en fiabilité des structures, lorsque la structure de dépendance des variables d'entrée n'est pas spécifiée, et a été l'objet d'un article, coécrit avec Bertrand Michel, Philippe Saint-Pierre et Nicolas Bousquet [Benoumechiara et al., 2018].

1.4.1 Chapitre 2 : État de l'art

Ce chapitre introduit les outils mathématiques nécessaires pour l'ensemble du manuscrit et en propose une revue de la littérature. Sont tout d'abord présentées les méthodes de propagation d'incertitudes ainsi que celles d'analyse de sensibilité. Puis sont introduites la notion de copule et les différentes familles de copules paramétriques permettant de construire, en pratique, des structures de dépendance et des lois jointes multidimensionnelles. Nous explorons ensuite les différentes méthodes d'analyse de sensibilité dans le cas de variables dépendantes. Enfin quelques éléments sont donnés concernant les différentes méthodes permettant de "méta modéliser" les modèles η avec une attention particulière portée sur les forêts aléatoires et les processus gaussiens.

1.4.2 Chapitre 3 : Analyse de sensibilité à partir des forêts aléatoires dans le cas de variables corrélées

Depuis de nombreuses années, le volume des données disponibles augmente de façon spectaculaire dans de nombreuses disciplines. Les techniques d'apprentissage supervisé tentent d'exploiter ces données dans le but de construire des modèles prédictifs permettant de se rapprocher au mieux du phénomène sous-jacent. Parmi ces techniques, l'algorithme des forêts aléatoires, introduit par Breiman [2001], a montré de très bonnes performances en pratique pour des problèmes complexes (relations non linéaires, interactions, grande dimension, etc.). La méthode agrège une collection d'arbres CART [Breiman et al., 1984] construits à partir d'échantillons bootstrap de l'ensemble d'apprentissage et leur agrégation améliore substantiellement les performances des arbres individuels. L'estimation d'un tel modèle $\hat{\eta}$ se fait à partir d'un échantillon d'apprentissage $\mathcal{D}_n = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$ de n observations du couple $(\mathbf{X}, Y)$ provenant de phénomènes physiques réels ou de simulations numériques de codes de calculs complexes et extrêmement coûteux.

Mesurer l'importance des variables d'entrées permet une interprétation du modèle d'apprentissage et une meilleure compréhension du phénomène. Les forêts aléatoires permettent l'évaluation de plusieurs mesures d'importance, dont une en particulier : la mesure d'importance par permutation [Breiman, 2001]. Cette méthode consiste à mesurer l'augmentation moyenne de l'erreur de prédiction en permutant une variable X_j dans un échantillon test ou dans l'échantillon Out-of-Bag (OOB) de la forêt représentant pour chaque arbre l'ensemble des observations qui ne sont pas retenues dans l'échantillon bootstrap de l'arbre. Cette mesure fut formalisée par Zhu et al. [2015] en considérant le modèle η et une fonction de coût quadratique et est définie théoriquement par

$$\mathcal{I}_\eta(X_j) = \mathbb{E}\Big[(Y - \eta(\mathbf{X}_{(j)}))^2\Big] - \mathbb{E}\Big[\,(Y - \eta(\mathbf{X}))^2\,\Big],$$

où $\mathbf{X}_{(j)} = (X_1, \ldots, X'_j, \ldots, X_d)$ est un vecteur aléatoire tel que X'_j est une réplication indépendante de X_j et indépendante des autres variables. Cette mesure, aussi nommée *Mean Decrease Accuracy*, a la propriété très particulière d'être liée théoriquement aux indices de Sobol' totaux [Gregorutti et al., 2015; Wei et al., 2015] :

$$\mathcal{I}_\eta(X_j) = 2\mathrm{Var}[\mathbb{E}[Y|X_j]].$$

Cette relation permet de comprendre qu'une permutation mesure l'effet d'une variable X_j sur la variance de Y. Une de nos contributions consiste à définir une nouvelle mesure d'importance par permutation qui contrairement à $\mathcal{I}_\eta(X_j)$, vise à permuter toutes les variables sauf X_j :

$$\mathcal{J}_\eta(X_j) = \mathbb{E}\Big[(Y - \eta(\mathbf{X}_{(-j)}))^2\Big] - \mathbb{E}\Big[\,(Y - \eta(\mathbf{X}))^2\,\Big],$$

où $\mathbf{X}_{(j)} = (X'_1, \ldots, X_j, \ldots, X'_d)$. Ainsi, nous montrons une relation entre l'indice $\mathcal{J}_\eta(X_j)$ et l'indice de Sobol du premier ordre :

$$\mathcal{J}_\eta(X_j) = 2\mathbb{E}[\mathrm{Var}[Y|X_j]].$$

L'effet d'une permutation de X_j vise à casser le lien entre la variable permutée X'_j et les autres variables. Lorsque les variables sont indépendantes, nous avons $\mathbf{X} \sim \mathbf{X}_{(j)}$ et donc $\eta(\mathbf{X}) \sim \eta(\mathbf{X}_{(j)})$

et ainsi, la permutation de X_j permet de mesurer uniquement son effet sur le modèle η. Or, la permutation de X_j a aussi pour effet de casser le lien avec ses autres composantes, ce qui implique que $\mathrm{Cov}[X'_j, X_k] = 0$ pour tout $k = 1, \ldots, d$ et $k \neq j$. C'est pourquoi, lorsque les variables sont dépendantes, l'effet de la permutation supprime cette dépendance entre la variable permutée X_j et les autres variables. La distribution de $\mathbf{X}_{(j)}$ est alors différente de celle de $\mathbf{X}$. Il en est de même pour la distribution de $\eta(\mathbf{X}_{(j)})$ qui est alors différente de celle de $\eta(\mathbf{X})$. Cette modification de la structure de dépendance des variables d'entrée change l'interprétation de l'indice de permutation obtenu, puisque celui-ci ne mesure plus seulement l'effet de X_j, mais aussi l'effet dû à la modification de la structure de dépendance.

Nous proposons alors de transformer la loi de $\mathbf{X}$ (que nous supposons connue), en une loi iso probabiliste où les variables sont indépendantes dans un espace uniforme. Effectuer la permutation de X_j dans cet espace puis revenir dans l'espace de départ permet de ne pas changer la loi de l'échantillon permuté. Cette approche, inspirée de Mara et al. [2015] en analyse de sensibilité permet d'obtenir des indices de Sobol en retirant les effets de la dépendance. Nous montrons aussi que les mesures d'importance par permutation avec une transformée iso probabiliste sont liées théoriquement à ces indices.

Néanmoins, ces mesures d'importance par permutation sont définies dans le cas théorique où η est considéré. Or en pratique, ce sont les estimateurs η_t des arbres t de la forêt qui sont utilisés. Nous montrons dans ce chapitre que la mesure d'importance par permutation $\mathcal{I}_{\hat{\eta}}(X_j)$ calculée à partir d'une forêt est biaisée par rapport à la mesure considérant le vrai modèle η et que ce biais est défini par

$$\mathcal{I}_{\eta}(X_j) - \mathcal{I}_{\hat{\eta}}(X_j) = \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\left(\hat{\eta}_t(\mathbf{X}_{(j)}) - \eta(\mathbf{X}_{(j)})\right)^2\right] + \frac{2}{T}\sum_{t=1}^{T} \mathbb{E}\left[\left(\hat{\eta}_t(\mathbf{X}_{(j)}) - \eta(\mathbf{X}_{(j)})\right)\left(Y - \hat{\eta}_t(\mathbf{X}_{(j)})\right)\right].$$

Ce biais est aussi valable pour l'indice $\mathcal{J}_{\eta}(X_j)$ et est directement lié à la qualité prédictive de la forêt. À partir de deux cas tests, nous montrons que ce biais peut être très important et peut changer l'interprétation des résultats si l'erreur de la forêt est importante.

1.4.3 Chapitre 4 : Shapley effects avec prise en compte des erreurs de krigeage et de Monte Carlo

L'étape d'analyse de sensibilité dans les études de fiabilité permet de mesurer l'importance de chaque variable sur la sortie du modèle. Les indices de Sobol [Sobol, 1993] sont parmi les indices les plus utilisés en pratique. De nombreuses méthodes permettent d'estimer ces indices en passant par le modèle η ou à partir de modèle de substitution (méta modèles). Cependant, la décomposition de la variance, sur laquelle sont basés les indices de Sobol, nécessite l'indépendance des variables aléatoires. Différentes généralisations des indices dans le cas de variables dépendantes furent proposées [Li and Rabitz, 2010; Chastaing et al., 2012; Mara et al., 2015]. Mais leurs estimations et leurs interprétations restent difficiles en pratique. Récemment, Owen [2014] a montré l'utilité des indices de pouvoir de Shapley [Shapley and Shubik, 1954] en analyse de sensibilité. Il est aussi montré dans Iooss and Prieur [2017] que ces indices ont une grande utilité dans le cas de variables aléatoires dépendantes. En effet, ils permettent une répartition

équitable de l'influence des variables sur le modèle en prenant en compte les effets marginaux, d'interactions et de dépendance. Les indices de Sobol décomposent la variance de Y et l'allouent à chaque sous-ensemble de $\mathbf{X}$ tandis que les indices de Shapley allouent cette variance à chaque variable.

Cependant, la complexité du calcul des indices de Shapley rend leurs estimations difficiles et coûteuses. En effet la formulation initiale demande de calculer toutes les permutations de "joueurs" (représentant les variables) et de calculer leurs contributions dans le jeu. Un algorithme proposé par Castro et al. [2009] et amélioré par Song et al. [2016] permet l'estimation de ces indices avec deux méthodes d'estimation : une méthode par permutations exactes et une méthode par permutations aléatoires. La première considère toutes les permutations de joueurs possibles tandis que la seconde en choisit aléatoirement un nombre fini. La seconde méthode a un intérêt particulier lorsque la dimension du problème augmente, et donc le nombre de permutations possibles. Pour les deux méthodes, l'estimation reste difficilement faisable pour les modèles coûteux en temps de calcul. L'utilisation d'un méta modèle de krigeage en substitution du vrai modèle permet de réduire le coût de l'estimation [Iooss and Prieur, 2017]. Cependant, l'utilisation d'un modèle approché augmente le risque d'erreur, en plus de l'erreur d'estimation de l'algorithme. Notre contribution dans ce chapitre est alors d'adapter l'algorithme de Song et al. [2016] afin d'inclure l'estimation de l'erreur. La procédure se fait par bootstrap et diffère selon la méthode exacte et aléatoire. L'erreur du modèle de krigeage est estimée en faisant un certain nombre de réalisations du processus aléatoire. En nous inspirant de Le Gratiet et al. [2014], nous pouvons séparer l'erreur issue du modèle de krigeage et l'erreur issue de l'estimateur. Nous montrons aussi expérimentalement que les intervalles de confiance convergent vers les bonnes probabilités de couverture.

Les chapitres 3 et 4 sont placés dans une situation de dépendance avérée et connue, situations pour laquelle les méthodes usuelles d'analyse de sensibilité ne sont pas toujours adaptées. Le prochain chapitre se place cette fois dans la situation plus délicate encore pour l'ingénieur fiabiliste, où la dépendance n'est pas connue, c'est-à-dire que la caractérisation du modèle probabiliste pour $\mathbf{X}$ est limitée à la donnée des distributions marginales de $X_1, X_2, \ldots, X_d$ mais qu'aucune information n'est disponible sur la copule C décrivant la structure de dépendance entre ces marges.

1.4.4 Chapitre 5 : Évaluation pénalisante d'un critère de risque lorsque la structure de dépendance est inconnue

Nous souhaitons nous prémunir contre l'éventualité d'une surestimation de la fiabilité du système ou de manière équivalente, d'une sous-estimation du risque que l'on cherche à évaluer. En d'autres termes, dans le cas qui nous intéresse ici où l'évaluation de $Q(\alpha)$ est souhaité, nous ne chercherons pas à estimer directement la vraie valeur de ce quantile - inaccessible - mais un minorant sur l'ensemble des possibilités qui demeurent admissibles, lorsque la connaissance disponible sur les distributions marginales est prise en compte. Cette démarche est conforme à la démarche (déterministe) de pénalisation des entrées, évoquée section 1.1, consistant à effectuer un unique calcul pour une valeur déterminée de chaque entrée incertaine, chacune d'entre elles étant prise de manière à minimiser la sortie y correspondante. Le résultat ainsi obtenu est alors une valeur

conservative de y, c'est-à-dire un minorant de la vraie valeur de y inconnue. De même, lorsque la dépendance est inconnue dans le cadre d'une méthode probabiliste, on cherchera à pénaliser la structure de dépendance c'est-à-dire à calculer un minorant de la quantité d'intérêt par rapport à l'ensemble des dépendances possibles entre les variables d'entrée.

Formellement, l'objectif est d'estimer un minorant de $Q(\alpha)$, lorsque la structure de dépendance de $\mathbf{X}$ est inconnue. On choisit de se placer dans le cadre d'une famille de copules paramétriques, ces différentes familles permettant de définir de façon commode la dépendance entre variables aléatoires, mais aussi afin de limiter le nombre de paramètres sur lesquels portera l'opération de minimisation. Certaines de ces familles (comme la copule de Clayton) sont adaptées à différents types de dépendances, assez courants en fiabilité des structures, avec notamment une forte dépendance de queue pour les événements rares (voir Nelsen [2007]). Nous définissons alors $C_{\boldsymbol{\theta}}$ comme une copule de paramètre $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ faisant le lien entre la distribution $F_{\boldsymbol{\theta}}$ de $\mathbf{X}$ et ses distributions marginales. Ainsi, la minimisation du quantile $Q_{\boldsymbol{\theta}}(\alpha)$, associé à une copule $C_{\boldsymbol{\theta}}$, permet l'estimation conservative de la valeur de y associée à un certain niveau de risque α. Le *quantile minimum* est défini par

$$Q_C^*(\alpha) := \inf_{\boldsymbol{\theta} \in \Theta} Q_{\boldsymbol{\theta}}(\alpha)$$

et, si elle existe, sa structure de dépendance pénalisante est définie par

$$\boldsymbol{\theta}_C^* = \operatorname*{argmin}_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} Q_{\boldsymbol{\theta}}(\alpha).$$

Afin d'estimer la quantité $Q_C^*(\alpha)$, nous proposons d'utiliser une grille $\boldsymbol{\Theta}_N$ de taille N afin d'explorer $\boldsymbol{\Theta}$ et un estimateur empirique du quantile. Cela nous permet d'avoir l'estimateur du quantile minimum

$$\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_N} Q_{\boldsymbol{\theta}}(\alpha)$$

et de définir l'estimateur de la structure de dépendance la plus pénalisante par

$$\hat{\boldsymbol{\theta}}_N = \operatorname*{argmin}_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_N} Q_{\boldsymbol{\theta}}(\alpha).$$

La première contribution de ce chapitre est de montrer la convergence en probabilité de ces estimateurs pour une grille $\boldsymbol{\Theta}_N$ suffisamment fine, et sous certaines hypothèses de régularité du modèle η et de la loi de $\mathbf{X}$.

Néanmoins, il n'est pas simple en pratique d'estimer $\hat{\boldsymbol{\theta}}_N$ lorsque la dimension est supérieure à trois. L'utilisation d'une famille paramétrique impose d'avoir la même forme de dépendance pour chaque paire de variables. Or, il est très probable que dans un même problème, les paires aient une structure de dépendance différente. L'extension la plus usuelle des copules en pratique est l'utilisation des *Vine Copulas* [Bedford and Cooke, 2001]. Cet outil permet de construire des structures de dépendances multivariées à partir de copules bivariées. Cependant, un grand nombre de constructions est possible pour différentes familles et différents agencements des copules bivariées. L'exploration de toutes ces configurations est en pratique infaisable, en particulier pour les codes de calculs très coûteux. C'est pourquoi nous partons de l'hypothèse que seules certaines paires dans un problème de fiabilité ont une influence significative sur le quantile. Nous proposons alors un algorithme permettant d'explorer itérativement l'ensemble des Vines possible et de déterminer les paires les plus influentes en obtenant les paramètres (et les familles de copules)

minimisant le quantile. Les paires considérées comme non influentes sont fixées à l'indépendance. Cela nous permet d'obtenir une structure de dépendance plus simple et facilement interprétable avec composée de seulement quelques paires dépendantes. Ces résultats sont appliqués pour différents cas jouets et le cas d'étude industriel explicité à la section 1.2.

Chapter 2

State of the Art

Abstract. This chapter introduces the mathematical tools necessary for the entire manuscript and provides a review of the literature. We first introduce the methods of sensitivity analysis for independent and dependent random inputs. We then introduce the notion of copula and Vine-copula to model probabilistic dependencies. Last, some elements are given concerning the different methods for "meta-modeling" with a particular attention to random forests and Gaussian processes.

Contents

2.1 Global Sensitivity Analysis with independent inputs 23
2.2 Global Sensitivity analysis with dependent inputs 33
2.3 Modeling Dependencies 38
2.4 Surrogate models 44

2.1 Global Sensitivity Analysis with independent inputs

We consider the input-output system where $\mathbf{X} = \{X_1, \ldots, X_d\}$ is a random vector of d input parameters and $Y = \eta(\mathbf{X})$ is the output random variable of a deterministic model $\eta : \mathbb{R}^d \to \mathbb{R}$ which can be a mathematical function or a computational code. The purpose of sensitivity analysis (SA) is to understand how a change in the input parameters affects the output of the model and to focus on identifying the parameters that most contribute to the model uncertainty [Saltelli et al., 2010].

Three approaches of SA can be distinguished. *Screening* methods [Morris, 1991] qualitatively analyze the importance of an input on the output of a model. These methods are mostly used in high dimensional problems to identify the input parameters that most contribute to the model output variability. *Local sensitivity methods* analyze quantitatively the importance of an input when the input parameters are at a fixed value. These approaches basically compute the local

partial derivatives at a fixed point $\mathbf{x}^*$ in order to rank the importance of a variable at this local point. *Global sensitivity methods* analyze the whole range of variation of the input parameters and can determine precisely the importance of each input but they require a large number of model evaluations. In this thesis, we are interested in global sensitivity methods which are the most general SA methods.

2.1.1 Variance-based methods

Variance-based methods aims at quantifying the input variables that contribute the most in the variability of the model output. In this section, it is assumed that the components of $\mathbf{X}$ are independent and that $\mathbf{X} \sim \mathcal{U}(0,1)^d$, however it is also applicable for any other distribution with independent components.

2.1.1.1 Importance measure

We seek to analyze how the distribution of a model output is changed when one or several inputs are set to a fixed value. The first variance-based sensitivity studies date back to the mid-80's with the *importance measure* of an input variable X_j introduced by Hora and Iman [1986] and defined by

$$I_j = \sqrt{\mathrm{Var}[Y] - \mathbb{E}[\mathrm{Var}[Y|X_j]]}. \tag{2.1}$$

This measure consists in "freezing" an input X_j (e.g. assigning a deterministic value to this variable) and studies the effect induced on the discrepancy of Y. For a more robust measure of importance, the same authors proposed the following measure [Iman and Hora, 1990]:

$$\frac{\mathrm{Var}[\mathbb{E}(\log Y|X_j)]}{\mathrm{Var}[\log Y]}. \tag{2.2}$$

Then, a new index is introduced by McKay [1997] and described as a "correlation ratio". This last quantity allows a simple interpretation of the contribution of X_j to the model discrepancy. It is defined by:

$$\frac{\mathrm{Var}[\mathbb{E}(Y|X_j)]}{\mathrm{Var}[Y]}. \tag{2.3}$$

The upper term quantifies the discrepancy of Y which is due to the variation of X_j. The ratio between this term and the total variance of Y leads to a normalized index giving the share of the variation of Y resulting from the variation of X_j. These measures are generalized using a variance decomposition which is defined in the following section

2.1.1.2 ANOVA decomposition

Most of the variance-based sensitivity methods are based on the *analysis of variance* (ANOVA). Coming from Fisher [1925], this procedure aims at decomposing a model η into a sum of elementary functions, including main and interaction effects of the input variables. Assuming that η is

square integrable, the Hoeffding decomposition [Hoeffding, 1948] decomposes η into 2^d terms. This decomposition states the following scheme:

$$\eta(\mathbf{X}) = \eta_0 + \underbrace{\overbrace{\sum_{j=1}^{d} \eta_j(X_j)}^{\text{order 1}}}_{\text{main effects}} + \underbrace{\overbrace{\sum_{j=1}^{d}\sum_{j<k}^{d} \eta_{jk}(X_j, X_k)}^{\text{order 2}} + \cdots + \overbrace{\eta_{1,2,\ldots,d}(\mathbf{X})}^{\text{order } d}}_{\text{interactions effects}} \tag{2.4}$$

where η_0 is a constant, η_j is a main effect of X_j, $\eta_{j,k}$ is a second-order interaction effect of X_j and X_k, and so on for higher interaction orders. These elementary functions are obtained from

$$\begin{aligned}
\eta_0 &= \mathbb{E}[Y], \\
\eta_j(X_j) &= \mathbb{E}[Y|X_j] - \eta_0, \\
\eta_{j,k}(X_j, X_k) &= \mathbb{E}[Y|X_j, X_k] - \eta_j - \eta_k - \eta_0
\end{aligned}$$

and similarly for higher interaction orders. Such a decomposition seems to offer a clear view on the different effects of the inputs. However, without any additional constraints on the components, there exists an infinite number of decomposition satisfying (2.4). If η is assumed to be square-integrable on its domain, the expansion (2.4) is unique under the following orthogonality constraint [Sobol, 1993]:

$$\int_0^1 \eta_{i_1,\ldots,i_s}(x_{i_1}, \ldots, x_{i_s}) \mathrm{d}x_{i_w} = 0 \text{ if } w \in \{1, \ldots, s\} \tag{2.5}$$

When considering the variance of $\eta(\mathbf{X})$, the representation (2.4) leads to the functional ANAlyse Of VAriance (ANOVA) which consists of expanding the output variance $\mathrm{Var}[Y]$ into $2^d - 1$ partial variance terms of increasing order

$$\mathrm{Var}[Y] = \sum_{j=1}^{d} V_j + \sum_{1 \le j \le k \le d} V_{jk} + \cdots + V_{1,\ldots,d}, \tag{2.6}$$

where $V_j = \mathrm{Var}[\eta_j(X_j)]$ is the first-order partial variance which indicates the part of variance of Y explained by X_j individually, $V_{jk} = \mathrm{Var}[\eta_{j,k}(X_j, X_k)] - V_j - V_k$ is a second-order partial variance measuring the part of variance of Y due to the interaction effect between X_j and X_k, and so on for higher order interaction effects.

2.1.1.3 Sobol' indices

The variance decomposition enables the establishment of variance-based sensitivity indices: the so-called Sobol' indices [Sobol, 1993]. By normalizing the partial variances

$$V_j = \mathrm{Var}[\mathbb{E}[Y|X_j]], \qquad V_{jk} = \mathrm{Var}[\mathbb{E}[Y|X_j, X_k]] - V_j - V_k, \qquad \ldots, \tag{2.7}$$

from (2.6) with $\mathrm{Var}[Y]$, the Sobol' indices are derived by

$$S_j = \frac{V_j}{\mathrm{Var}[Y]}, \qquad S_{jk} = \frac{V_{jk}}{\mathrm{Var}[Y]}, \qquad \ldots \tag{2.8}$$

where S_j is a first-order sensitivity index (also called main effect index), S_{jk} is a second-order sensitivity index and so on for higher interaction orders. Because these indices are normalized by the total variance $\mathrm{Var}[Y]$, they result in the following properties:

$$0 \leq S_j \leq 1, \quad 0 \leq S_{jk} \leq 1, \quad \ldots \tag{2.9}$$

$$\sum_{j=1}^{d} S_j + \sum_{1 \leq j \leq k \leq d} S_{jk} + \cdots + S_{1,\ldots,d} = 1. \tag{2.10}$$

As shown in property (2.9), the Sobol' indices are standardized values between 0 and 1, such as if $S_j = 1$, then X_j individually explains all the variance of Y and inversely. Moreover, property (2.10) shows that their sum is always equal to 1, which means that the sum of all partial variance of input variables explains the variance of Y.

When a model has many interactions, it becomes difficult to interpret the main and interaction effects. Moreover, the number of indices significantly increases with the dimension. Thus, Homma and Saltelli [1996] introduced the Sobol' total effect index which gathers information of the main and interaction effects. For a variable X_j, the total index is defined by

$$ST_j = S_j + \sum_{k=1, k \neq j}^{d} S_{jk} + \cdots + S_{1,\ldots,d} \tag{2.11}$$

$$= 1 - \frac{\mathrm{Var}[\mathbb{E}[Y|\mathbf{X}_{-j}]]}{\mathrm{Var}[Y]} = \frac{\mathbb{E}[\mathrm{Var}[Y|\mathbf{X}_{-j}]]}{\mathrm{Var}[Y]}, \tag{2.12}$$

where $\mathbf{X}_{-j} = \mathbf{X} \backslash X_j$ (the vector $\mathbf{X}$ without X_j). Equation (2.11) defines the total effect of X_j as the sum of its main and interaction effects which corresponds to the excepted variance of Y when all variables but X_j are fixed, as stated in (2.12). The property from (2.9) holds for the total effect, whereas (2.10) does not always hold due the redundancy of interaction effects. Thus, $\sum_{j=1}^{d} ST_j \geq 1$ if there is interaction between variables and on the contrary the model is purely additive if $\sum_{j=1}^{d} ST_j = 1$.

2.1.1.4 Estimating Sobol' indices

Sobol' indices can be estimated using different techniques. To do so, the estimators based on Monte Carlo (MC) sampling are most commonly used. Various estimators exist in the literature and the most popular are introduced in the following.

A "brute force Monte Carlo" approach is a very intuitive way to compute the Sobol' indices, but it is threatens to be costly. Indeed, for an index S_j, one has to estimate as many conditional expectations $\mathbb{E}[Y|X_j]$ to have a correct estimation of the partial variance V_j. This naive technique has a very slow convergence and a more efficient way has been proposed with a faster convergence rate.

We first introduce two independent sampling matrices $\mathbf{A}$ and $\mathbf{B}$ composed of N realization

of the random vector $\mathbf{X}$ as illustrated bellow:

$$\mathbf{A} = \begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,p} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{N,1} & a_{N,2} & \cdots & a_{N,p} \end{pmatrix} \quad \text{et} \quad \mathbf{B} = \begin{pmatrix} b_{1,1} & b_{1,2} & \cdots & b_{1,p} \\ b_{2,1} & b_{2,2} & \cdots & b_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ b_{N,1} & b_{N,2} & \cdots & b_{N,p} \end{pmatrix} . \tag{2.13}$$

Let us now denote by $\mathbf{A}_{\mathbf{B}}^{(j)}$ (resp. $\mathbf{B}_{\mathbf{A}}^{(j)}$) the sampling matrix $\mathbf{A}$ (resp. $\mathbf{B}$) where the j-th column is replaced by the j-th column from $\mathbf{B}$ (resp. $\mathbf{A}$) as shown below:

$$\mathbf{A}_{\mathbf{B}}^{(j)} = \begin{pmatrix} a_{1,1} & \cdots & b_{1,i} & \cdots & a_{1,p} \\ a_{2,1} & \cdots & b_{2,i} & \cdots & a_{2,p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{N,1} & \cdots & b_{N,i} & \cdots & a_{N,p} \end{pmatrix} . \tag{2.14}$$

The evaluation of a a matrix through the model η denotes (for example with $\mathbf{A}$) the set $\{\eta(\mathbf{A})_i \equiv \eta(A_i), i = 1, \ldots, N\}$ of output samples from the evaluation of the N rows where A_i is the i-th row of the matrix $\mathbf{A}$. Then, the combined matrix $\mathbf{A}_{\mathbf{B}}^{(j)}$ (resp. $\mathbf{B}_{\mathbf{A}}^{(j)}$) can be used to compute the partial variances necessary for the computation of the Sobol' indices. The estimator of V_j, introduced by Sobol [1993], is given by

$$\hat{V}_j = \frac{1}{N} \sum_{i=1}^{N} \eta\left(\mathbf{A}\right)_i \eta\left(\mathbf{B}_{\mathbf{A}}^{(j)}\right)_i - \hat{\eta}_0^2, \tag{2.15}$$

where $\hat{\eta}_0$ is the empirical estimator of the mean of Y defined by

$$\hat{\eta}_0 = \frac{1}{N} \sum_{i=1}^{N} \eta\left(\mathbf{A}\right)_k . \tag{2.16}$$

The same can be done to estimate $V_{-j} = \mathrm{Var}[\mathbb{E}[Y|\mathbf{X}_{-j}]]$ for the total Sobol' index and the estimator, introduced by Homma and Saltelli [1996], is given by

$$\hat{V}_{-j} = \frac{1}{N} \sum_{i=1}^{N} \eta\left(\mathbf{A}\right)_i \eta\left(\mathbf{A}_{\mathbf{B}}^{(j)}\right)_i - \hat{\eta}_0^2. \tag{2.17}$$

The proofs of these equations can be found in Saltelli et al. [2010].

The estimator of V_j have been improved by Saltelli [2002] and Tarantola et al. [2007] with the following estimator

$$\hat{V}_j = \frac{1}{N} \sum_{i=1}^{N} \eta\left(\mathbf{A}\right)_i \left(\eta\left(\mathbf{B}_{\mathbf{A}}^{(j)}\right)_i - \eta(\mathbf{B})_i\right). \tag{2.18}$$

Equivalently, Sobol' [2007] has numerically improved the estimator of V_{-j} and by considering the alternative estimator

$$\hat{V}_{-j} = \hat{V} - \frac{1}{N} \sum_{i=1}^{N} \eta\left(\mathbf{A}\right)_i \left(\eta\left(\mathbf{A}\right)_i - \eta\left(\mathbf{A}_{\mathbf{B}}^{(j)}\right)_i\right), \tag{2.19}$$

where $\hat{V}$ is the estimated variance of Y defined by

$$\hat{V} = \frac{1}{N}\sum_{i=1}^{N}\eta(\mathbf{A})_i - \hat{\eta}_0^2. \tag{2.20}$$

Alternative estimators for V_j and V_{-j} have also been offered by Jansen et al. [1994] and Jansen [1999]. The so-called "Jansen's formulae" are respectively

$$\hat{V}_j = \hat{V} - \frac{1}{2N}\sum_{i=1}^{N}\left(\eta\left(\mathbf{B}\right)_i - \eta\left(\mathbf{A}_{\mathbf{B}}^{(j)}\right)_i\right)^2 \tag{2.21}$$

and

$$\hat{V}_{-j} = \frac{1}{2N}\sum_{i=1}^{N}\left(\eta\left(\mathbf{A}\right)_i - \eta\left(\mathbf{A}_{\mathbf{B}}^{(j)}\right)_i\right)^2 \tag{2.22}$$

A summary of the most known estimators of the Sobol' indices is given in Table 2.1.

V_j for S_j	Reference
(a) $\frac{1}{N}\sum_{i=1}^{N}\eta\left(\mathbf{A}\right)_i\eta\left(\mathbf{B}_{\mathbf{A}}^{(j)}\right)_i - \hat{\eta}_0^2$	Sobol [1993]
(b) $\frac{1}{N}\sum_{i=1}^{N}\eta\left(\mathbf{B}\right)_i\left(\eta\left(\mathbf{A}_{\mathbf{B}}^{(j)}\right)_i - \eta\left(\mathbf{A}\right)_i\right)$	Saltelli [2002]
(c) $\hat{V} - \frac{1}{2N}\sum_{i=1}^{N}\left(\eta\left(\mathbf{B}\right)_i - \eta\left(\mathbf{A}_{\mathbf{B}}^{(j)}\right)_i\right)^2$	Jansen et al. [1994]
$V_{\sim j}$ for ST_j	
(d) $\frac{1}{N}\sum_{i=1}^{N}\eta\left(\mathbf{A}\right)_i\eta\left(\mathbf{A}_{\mathbf{B}}^{(j)}\right)_i - \hat{\eta}_0^2$	Homma and Saltelli [1996]
(e) $\hat{V} - \frac{1}{N}\sum_{i=1}^{N}\eta\left(\mathbf{A}\right)_i\left(\eta\left(\mathbf{A}\right)_i - \eta\left(\mathbf{A}_{\mathbf{B}}^{(j)}\right)_i\right)$	Sobol' [2007]
(f) $\frac{1}{2N}\sum_{i=1}^{N}\left(\eta\left(\mathbf{A}\right)_i - \eta\left(\mathbf{A}_{\mathbf{B}}^{(j)}\right)_i\right)^2$	Jansen [1999]

Table 2.1 *Different estimators of the partial variances from Monte Carlo sampling.*

The MC sampling procedure has the great property to provide asymptotic behaviors. Confidence intervals can be obtained from these asymptotic results [Janon et al., 2014] or can be estimated from bootstrap sampling [Archer et al., 1997]. Unfortunately, the MC sampling-based methods are relatively costly and a large number of model evaluations is necessary to obtain precise results (convergence rate approximately in $\sqrt{N}$). The use of quasi-Monte Carlo sequences instead of MC samples can reduce the estimation cost by exploring with a better discrepancy the variation space of $\mathbf{X}$ [Saltelli et al., 2008]. However, the asymptotic properties are lost and no confidence intervals can be obtained from the quasi-MC. Another estimation technique considers a multi-dimensional Fourier transform with the FAST method [Cukier et al., 1978; Saltelli et al., 1999], which has then been improved in Tissot and Prieur [2012]. However, according to Tissot and Prieur [2012] the FAST method can be biased, unstable and costly in high dimensional problems. Alternatives to reduce the sampling size considered surrogate models exist, such as kriging [Oakley and O'Hagan, 2004; Marrel et al., 2009; Le Gratiet et al., 2014] and polynomial

chaos expansion [Sudret, 2007]. The latter method relies into expanding the model response onto a suitable polynomial basis. A relationship between the functional ANOVA decomposition and the polynomial representation makes easier the computation of the Sobol' indices. However, this polynomial representation also requires the independence of the random variables, although a new polynomial decomposition for dependent random variables has been recently proposed in Rahman [2018].

Variance-based methods are largely used in Global Sensitivity Analysis. However, other studies have been proposed to consider different quantity of interests, such as quantiles, probabilities and other moments. The Goal Oriented Sensitivity Analysis (GOSA) introduced by Fort et al. [2016] aims at measuring the relation between a variable X_j with a quantity of interest of Y through different contrast functions. Another method generalized the study by considering the entire distribution of Y and is detailed in the following section.

2.1.2 A distribution-based method

2.1.2.1 Definition

Introduced by Borgonovo [2007], the delta indices quantify the change in the distribution of Y if a variable X_j is fixed. Let f_Y and $f_{Y|X_j}$ be respectively the probability density functions (pdf) of Y and $Y|X_j$. The change in the output distribution when the input parameter X_j is fixed to a given value x_j^* can be measured by the *shift* computed as

$$s(x_j^*) = \int_{\mathcal{D}_Y} |f_Y(y) - f_{Y|X_j=x_j^*}(y)| \mathrm{d}y \quad , \tag{2.23}$$

which represents the area between the two distributions f_Y and $f_{Y|X_j}$. Figure 2.1 illustrates this quantity and shows the shift $s(x_j^*)$ in the blue area between the densities f_Y (red line) and $f_{Y|X_j=x_j^*}$ (blue line). A low value of $s(x_j^*)$ indicates a low influence of X_j of Y when fixed to x_j^*. On the contrary, if the shift is large, then fixing X_j to x_j^* has a high influence on Y.

As shown in Equation (2.23), the definition of the shift is dependent to the value of x_j^*. In order to have a global index of the influence of X_j, the excepted value of $s(x_j^*)$ is taken over the domain of X_j, which can be written by

$$\begin{aligned} \mathbb{E}_{X_j}[s(X_j)] &= \int_{\mathcal{D}_{X_j}} f_{X_j}(x_j) s(x_j) \mathrm{d}x_j \\ &= \int_{\mathcal{D}_{X_j}} f_{X_i}(x_j) \left[\int_{\mathcal{D}_Y} |f_Y(y) - f_{Y|X_j}(y)| \mathrm{d}y \right] \mathrm{d}x_j \quad . \end{aligned} \tag{2.24}$$

From (2.24), Borgonovo [2007] proposed the following delta index

$$\delta_i = \frac{1}{2} \mathbb{E}_{X_j}[X(x_j)] \quad . \tag{2.25}$$

As for the Sobol' index, the delta index can measure the importance of groups of variables in order to measure interactions. For a subset $\mathbf{u} = (X_j, X_k)$, the delta index of (X_j, X_k) is given

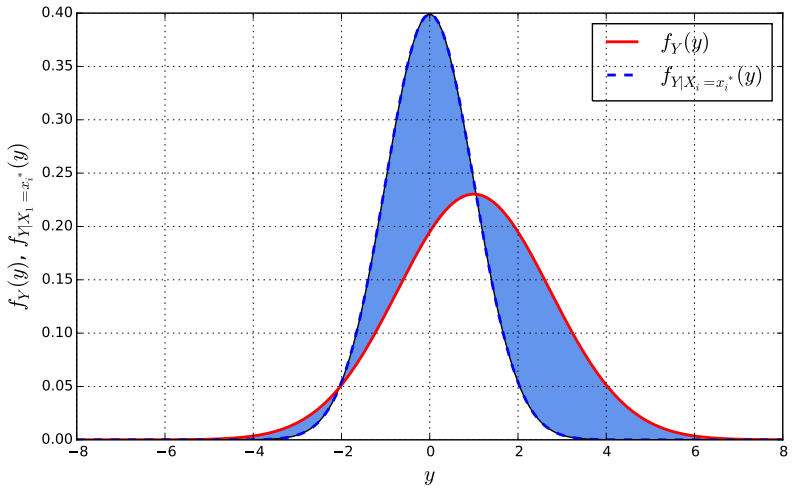


Figure 2.1 *Illustration of the shift $s(x_i^*)$ for a simple additive problem with input Gaussian distributions.*

by

$$\delta_{jk} = \frac{1}{2}\mathbb{E}_{X_jX_k}[s(X_j, X_k)], \tag{2.26}$$

where

$$s(X_j, X_k) = \int_{\mathcal{D}_Y} |f_Y(y) - f_{Y|X_jX_k}(y)|\mathrm{d}y \tag{2.27}$$

is the shift when $X_j = x_j$ and $X_k = x_k$.

The computation of this index can be costly due to the estimation of the whole distributions. Thus, a variant of the estimation has been proposed in Borgonovo et al. [2011] by using the cumulative distribution functions (cdf). Let $\mathcal{D}_Y$ be decomposed in two sub-domains $\mathcal{D}_Y^+$ and $\mathcal{D}_Y^-$ such that

$$\begin{aligned} \mathcal{D}_Y^+ &= \left\{y : f_Y(y) > f_{Y|X_i}(y)\right\} \quad , \\ \mathcal{D}_Y^- &= \left\{y : f_Y(y) \leq f_{Y|X_i}(y)\right\} \quad . \end{aligned} \tag{2.28}$$

Thus, as shown in Borgonovo et al. [2011], the shift can be computed by

$$\begin{aligned} s(x_j^*) &= 2\left[F_Y(\mathcal{D}_Y^+) - F_{Y|X_j=x_j^*}(\mathcal{D}_Y^+)\right] \\ &= 2\left[F_{Y|X_j=x_j^*}(\mathcal{D}_Y^-) - F_Y(\mathcal{D}_Y^-)\right] \quad . \end{aligned} \tag{2.29}$$

Figure 2.2 illustrates the sub-domains $\mathcal{D}_Y^+$ and $\mathcal{D}_Y^-$ for the computation of $s(x_j^*)$. The two black dots represent the values y_a and y_b for which $f_Y(y) = f_{Y|X_j}(y)$. The sub-domains are defined by $\mathcal{D}_Y^+ =]-\infty, y_a] \cup [y_b, +\infty[$ and $\mathcal{D}_Y^- =]y_a, y_b[$. Using Equation (2.29), the shift is computed for this illustration as

$$s(x_j^*) = 2\left[F_Y(y_a) + F_{Y|X_j=x_j^*}(y_b) - F_Y(y_b) - F_{Y|X_j=x_j^*}(y_a)\right] \quad . \tag{2.30}$$

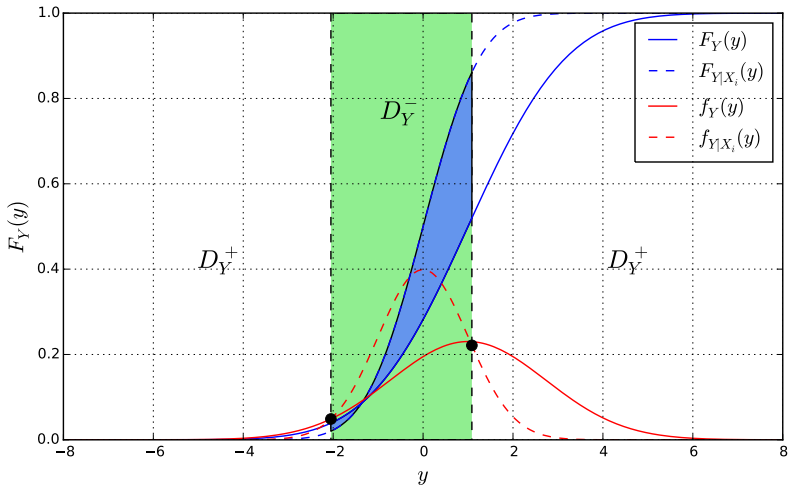


Figure 2.2 *Illustration of $\mathcal{D}_Y^+$ and $\mathcal{D}_Y^-$*

Generally, this variant makes the computation of the shift simpler with only four evaluations of the CDF. However, an optimization problem must be solved to determine $\mathcal{D}_Y^+$ and $\mathcal{D}_Y^-$.

As for the Sobol' indices, the delta indices are normalized, such as $0 \leq \delta_j \leq 1$, which implies that if $\delta_j = 0$, then X_j has no influence on Y. However, contrary to the Sobol' indices, no mathematical interpretation can give significance to $\sum_{j=1}^{d} \delta_j$. Note that the delta index is a special case of another class of sensitivity indices introduced by Da Veiga [2015] based on the distance correlation measure [Székely and Rizzo, 2013] and Hilbert-Schmidt independence criterion [Gretton et al., 2005].

2.1.2.2 Estimation of the delta indices

Because the distribution of Y is unknown in practice, one must estimate the distribution of Y and $Y|X_j$ to compute the delta index of X_j. Borgonovo [2007] proposed to approximate these distributions using a maximum likelihood estimator followed with a Kolmogorov-Smirnov test for validation. However, this approach is parametric and makes some assumptions on the output distribution. Thus, the use of a kernel density estimation (KDE) [Parzen, 1962] is a relevant non-parametric alternative.

The KDE method allows to create a continuous estimate of the distribution of Y given a sample data $\mathcal{Y} = \{y^{(1)}, \ldots, y^{(N)}\}$ of N observations. A kernel is created for each element of $\mathcal{Y}$ and the density estimation is computed by

$$\hat{f}_Y(y) = \frac{1}{N h_K} \sum_{i=1}^{N} K\left(\frac{y - y^{(i)}}{h_K}\right), \tag{2.31}$$

where K is a kernel function with a window h_K governing the degree of smoothing of the estimate. One example of window is the the Silverman window used for Gaussian Kernels which is compute by minimizing the integrated root mean square error of the estimated density [Silverman, 1986].

For a given point x_j^*, the shift $s(x_j^*)$ can be estimated using (2.23) or (2.30) and by replacing the true densities of Y and $Y|X_j$ with their KDE estimate. For example, using Equation (2.30), the estimated shift is given by:

$$\hat{s}(x_j^*) = 2\left[\hat{F}_Y(y_a) + \hat{F}_{Y|X_j=x_j^*}(y_b) - \hat{F}_Y(y_b) - \hat{F}_{Y|X_j=x_j^*}(y_a)\right],$$

where $\hat{F}_Y$ are KDE estimation $\hat{F}_{Y|X_j=x_j^*}$ of the output densities using (2.31). Finally, the delta index of X_j is computed from the mean of the estimated shift over a number N_q of fixed values $\{x_j^{(l)}\}_{l=1,\dots,N_q}$:

$$\hat{\delta}_j = \frac{1}{2N_q}\sum_{l=1}^{N_q}\hat{s}(x_j^{(l)}). \tag{2.32}$$

Figure 2.3 illustrates the estimated densities used to measure $\hat{\delta}_j$. The full blue line represents the estimated density of f_Y, while the dotted lines are the estimated densities of $f_{Y|X_j=x_j^*}$ for multiple x_j. A Gaussian kernel is used in the estimation of these densities.

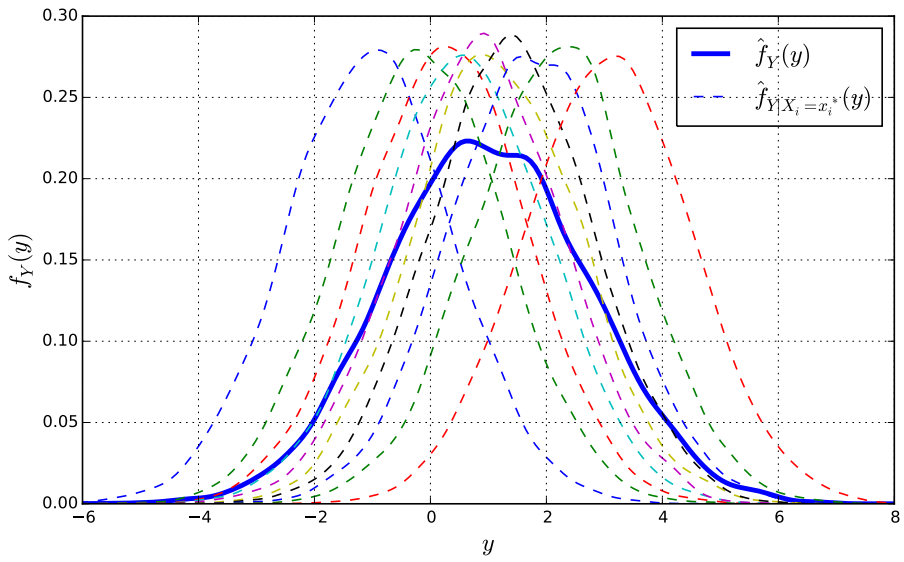


Figure 2.3 *Example of a Gaussian density estimations for f_Y and $f_{Y|X_j=x_j^*}$.*

The KDE based estimation can still remain too much costly for some studies. New techniques proposed to combine the principle of fractional moment-based maximum entropy with the use of Nataf transformation to improve the computational efficiency. However, this technique rests on various technical assumptions such as independence between inputs [Zhang et al., 2014]. Derennes et al. [2019] later proposed. To overcome this constraint, Wei et al. [2013]introduces a single-loop Monte Carlo simulation scheme which needs only one set of samples for computing all the indices. Derennes et al. [2019] shows via simulation that this method may be inaccurate and proposes a new estimation scheme which greatly improves accuracy of the single-loop and combines importance sampling and kernel estimation procedures. Nevertheless, the last method

may still lead to inacurate estimations, mostly due to the kernel approximation of the joint density of the output and the considered input. Thus, another estimation scheme valid for dependent model inputs is proposed by Derennes et al. [2018]. This method is build on the copula representation of indices and uses maximum entropy methods to estimate this copula.

The delta index is a very general index that measures the importance of a variable X_j on the whole distribution of Y. The definition of the shift given in (2.23) makes no assumptions on the model nor on the dependence structure of the input random variables. However, the assumption on the independence of the random variables can be necessary for the density estimations of Y. Moreover, contrary to the Sobol' index, the definition of the delta index does not a have share of contributions between the input variables, but is more like a bounded distance between the unconditional and conditional output distribution. Moreover, the interpretation of these indices in the case of dependent inputs is not clear because we ignore how the conditional output distribution is influenced by the dependencies.

2.2 Global Sensitivity analysis with dependent inputs

Most of the SA methods were initially defined for independent components. However, some methods give erroneous (or not calculable) results in the presence of stochastic dependencies between the input random variables. As stated in Section 2.1.1.2, the functional ANOVA decomposition only holds if the random variables are independent. In the case of dependent inputs, the uniqueness of this decomposition becomes unverified. In this section, we introduce two existing alternatives to obtain interpretable sensitivity measures in the presence of dependencies.

2.2.1 Influence of dependencies for the Sobol' indices

To illustrate the influence of dependencies in the interpretation of the Sobol' indices, we consider a simple linear model defined in the form of $Y = 3X_1 + 2X_2 + X_3$. Because the model is additive, the first order and total Sobol' indices are equal for this configuration. Let $\mathbf{X} \sim \mathcal{N}(0, \Sigma)$, where Σ is a linear correlation matrix defined by:

$$\Sigma = \begin{pmatrix} 1 & 0.8 & 0.4 \\ 0.8 & 1 & 0.7 \\ 0.4 & 0.7 & 1 \end{pmatrix}. \tag{2.33}$$

We compute the Sobol' indices of this problem for the three random variables in the independent and dependent case (with the correlation Σ). Figure 2.4 shows the results of these indices and we observe a variation of the Sobol' indices between the independent and dependent cases. The importance of X_1 decreases while it is the opposite for X_2 and X_3. The correlation between the variables influence the interpretation of the indices. Indeed, it is difficult to understand if the contribution of a variable X_j is due to its marginal importance or due to its correlation with another variable.

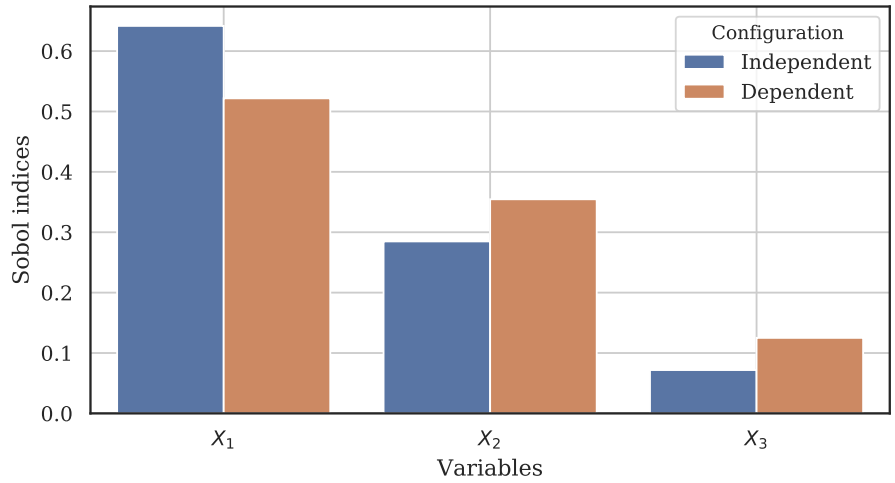


Figure 2.4 *Estimation of the Sobol' indices in the independent and dependent case for a linear Gaussian model in* $d = 3$.

Ignoring the dependence structure and using unsuited SA methods can lead to misinterpretations and eventually bad decisions. Therefore, we propose a short review of new techniques that have been proposed to deal with dependencies in SA.

Different approaches tried to counter this issue and generalize the Sobol' indices for dependent inputs [Chastaing et al., 2012; Kucherenko et al., 2012; Caniou, 2012]. One in particular, developed in Mara et al. [2015], relies on an iso-probabilistic transformation to transform $\mathbf{X} \sim p_{\mathbf{X}}$ into a random vector $\mathbf{U} \sim \mathcal{U}^n(0,1)$ with independent and uniformly distributed components. This approach led to the definition of new indices which are introduced in the next section.

2.2.2 Full and independent Sobol' indices

Such a transformation can be done using the Rosenblatt Transformation (RT) [Rosenblatt, 1952] as defined in the following definition.

Definition 1 (Rosenblatt Transformation)**.** *Let* $\mathbf{X} = (X_1, \ldots, X_k, \ldots, X_d)$ *be a random vector with joint distribution* $F_{\mathbf{X}}$ *defined by its marginal cumulative distribution functions* $F_1, \ldots, F_d$*. The Rosenblatt transformation* T *of* $\mathbf{X}$ *reads:*

$$\mathbf{U} = T(\mathbf{X})$$

where T is defined by

$$T : \mathbb{R}^d \to \mathbb{R}^d$$
$$\mathbf{X} \mapsto \mathbf{U} = \begin{pmatrix} F_1(X_1) \\ \vdots \\ F_{k|1,\dots,k-1}(X_k|X_1,\dots,X_{k-1}) \\ \vdots \\ F_{d|1,\dots,d-1}(X_d|X_1,\dots,X_{d-1}) \end{pmatrix}$$

where $F_{k|1,\dots,k-1}$ is the conditional cumulative distribution function of X_k conditioned by $X_1,\dots,X_{k-1}$.

However, multiple transformations can be done due to the $d!$ possible permutations of the components of $\mathbf{X}$. For example with $d = 3$, the RT of (X_3, X_2, X_1) would be different than the RT of (X_1, X_2, X_3). Nevertheless, in the following procedure only the d circular permutations with the associated RT are considered in order to keep the original ordering of the variables. We denote as $\mathbf{U}^j = (U_1^j, \dots, U_d^j)$ the random vector obtained from the RT of the set $(X_j, X_{j+1}, \dots, X_d, X_1, \dots, X_{j-1})$ such as

$$(X_j, X_{j+1}, \dots, X_d, X_1, \dots, X_{j-1}) \sim p_{\mathbf{X}} \xrightarrow{T} (U_1^j, \dots, U_d^j) \sim \mathcal{U}^d(0,1). \tag{2.34}$$

This RT corresponds to a particular j-th ordering and changing it would correspond to another RT.

Such a mapping is bijective and we can consider a function g_j such as $Y = \eta(\mathbf{X}) = g_j(\mathbf{U}^j)$. Because the elements of $\mathbf{U}^j$ are independent, the ANOVA decomposition is unique and variance-based sensitivity indices can be computed. Thus, we can write

$$g_j(\mathbf{U}^j) = g_\emptyset + \sum_{k=1}^{d} g_k(U_k^j) + \sum_{k=1}^{d}\sum_{k<l}^{d} g_{k,l}(U_k^j, U_l^j) + \cdots + g_{1,\dots,d}(U_1^j, \dots, U_d^j) \tag{2.35}$$

where $g_\emptyset = \mathbb{E}[g_j(\mathbf{U}^j)]$. Because the summands in (2.35) are orthogonal, the variance based decomposition can be derived, such that

$$\mathrm{Var}[Y] = \mathrm{Var}[g_j(\mathbf{U}^j)] = \sum_{k=1}^{d} V_k + \sum_{k=1}^{d}\sum_{k<l}^{d} V_{k,l} + \cdots + V_{1,\dots,d} \tag{2.36}$$

where $V_k = \mathrm{Var}[\mathbb{E}(g_j(\mathbf{U}^j)|U_k^j)]$, $V_{k,l} = \mathrm{Var}[\mathbb{E}(g_j(\mathbf{U}^j)|U_k, U_l)] - V_k - V_l$ and so on for higher orders. Just like for the independent case, the first-order Sobol' indice is defined by dividing V_k by the total variance, such that

$$S_k^j = \frac{\mathrm{Var}[\mathbb{E}[g_j(\mathbf{U}^j)|U_k^j]]}{\mathrm{Var}[g_j(\mathbf{U}^j)]}. \tag{2.37}$$

The total Sobol' indice can also be written as

$$ST_k^j = \frac{\mathbb{E}[\mathrm{Var}[g_j(\mathbf{U}^j)|\mathbf{U}_{-k}^j]]}{\mathrm{Var}[g_j(\mathbf{U}^j)]}, \tag{2.38}$$

where $\mathbf{U}^j_{-k}$ is the vector $\mathbf{U}^j$ not containing U^j_k. Each transformed vector $\mathbf{U}^j$ from a particular ordered RT, now have d new indices. To interpret the results, we may establish a link between the sensitivity indices of $\mathbf{U}^j$ and those of X_j. We remind that (2.37) and (2.38) are derived from the RT of the ordered set $(X_j, X_{j+1}, \ldots, X_d, X_1, \ldots, X_{j-1})$ which can be seen as the following bijective mapping

$$\Big[(X_j), (X_{j+1}|X_j), \ldots, (X_1|X_j, X_{j+1}, \ldots, X_d), \ldots, (X_{j-1}|\mathbf{X}_{-(j-1)})\Big] \longleftrightarrow (U^j_1, U^j_2, \ldots, U^j_d),$$

where $U^j_1 = F_j(X_j)$, $U^j_2 = F_{j+1|j}(X_{j+1}|X_j)$ and so on for other variables. From here, Mara et al. [2015] proposed to consider only the variables U^j_1 and U^j_d due to their interesting properties. The variable U^j_1 is representative of the behavior of X_j taking into account the dependence with other variables. On the opposite, the variable U^j_d represents the effects of X_{j-1} that is not due to its dependence with other variables. As a consequence, Mara et al. [2015] introduced the following indices:

- the *full* Sobol' indices which describe the influence of a variable including its dependence with other variables

$$S^{full}_j = \frac{\mathrm{Var}[\mathbb{E}[g_j(\mathbf{U}^j)|U^j_1]]}{\mathrm{Var}[g_j(\mathbf{U}^j)]} = \frac{\mathrm{Var}[\mathbb{E}[\eta(\mathbf{X})|X_j]]}{\mathrm{Var}[\eta(\mathbf{X})]} \tag{2.39}$$

$$ST^{full}_j = \frac{\mathbb{E}[\mathrm{Var}[g_j(\mathbf{U}^j)|\mathbf{U}^j_{-1}]]}{\mathrm{Var}[g_j(\mathbf{U}^j)]} = \frac{\mathbb{E}[\mathrm{Var}[\eta(\mathbf{X})|(\mathbf{X}_{-j}|X_j)]]}{\mathrm{Var}[\eta(\mathbf{X})]} \tag{2.40}$$

  where $\mathbf{X}_{-j}|X_j$ represent all components except X_j not taking into account the dependence with the variable X_j.
- the *independent* Sobol' indices which describe the influence of variables without its dependence with other variables.

$$S^{ind}_j = \frac{\mathrm{Var}[\mathbb{E}[g_{j+1}(\mathbf{U}^{j+1})|U^{j+1}_d]]}{\mathrm{Var}[g_{j+1}(\mathbf{U}^{j+1})]} = \frac{\mathrm{Var}[\mathbb{E}[\eta(\mathbf{X})|(X_j|\mathbf{X}_{-j})]]}{\mathrm{Var}[\eta(\mathbf{X})]} \tag{2.41}$$

$$ST^{ind}_j = \frac{\mathbb{E}[\mathrm{Var}[g_{j+1}(\mathbf{U}^{j+1})|\mathbf{U}^{j+1}_{-d}]]}{\mathrm{Var}[g_{j+1}(\mathbf{U}^{j+1})]} = \frac{\mathbb{E}[\mathrm{Var}[\eta(\mathbf{X})|\mathbf{X}_{-j}]]}{\mathrm{Var}[\eta(\mathbf{X})]} \tag{2.42}$$

  where $X_j|\mathbf{X}_{-j}$ represent the component X_j not taking account the dependence with other variables and with the convention that $\mathbf{U}^{d+1} = \mathbf{U}^1$ and $g_{d+1} = g_1$.

Thanks to the RT, we can also define the sensitivity indices of $(X_i|X_u), i = 1, \ldots, d$ and $u \subset \mathcal{D}\backslash\{i\}, u \neq \emptyset$ via U^i_u which represent the effect of X_i without its mutual dependent contribution with X_u. These indices can be estimated with a Monte Carlo algorithm and the procedure is very similar to the case of the Sobol' indices as described in Section 2.1.1.4.

The estimation of (S^{full}_j, ST^{full}_j, S^{ind}_{j-1}, ST^{ind}_{j-1}) can be done with four samples using the "pick and freeze" strategy (see Saltelli et al. [2010]). The procedure is divided in two steps:

- Generate and prepare the samples:
  - generate two independent sampling matrices $\mathbf{A}$ and $\mathbf{B}$ of size $N \times d$ with $\mathcal{U}(0,1)^d$ rows,

  - creates $\mathbf{B}_{\mathbf{A}}^{(1)}$ (resp. $\mathbf{B}_{\mathbf{A}}^{(d)}$) in the following way: keep all columns from $\mathbf{B}$ except the 1-th (resp. d-th) column which is taken from $\mathbf{A}$.
- Compute the indices with a given estimator:

$$\widehat{S}_j^{full} = \frac{\frac{1}{N}\sum_{i=1}^{N} g_j(A)_j g_j(\mathbf{B}_{\mathbf{A}}^{(1)})_i - g_{j_0}^2}{\widehat{V}} \tag{2.43}$$

$$\widehat{ST}_j^{full} = 1 - \frac{\frac{1}{N}\sum_{i=1}^{N} g_j(B)_i g_j(\mathbf{B}_{\mathbf{A}}^{(1)})_i - g_{j_0}^2}{\widehat{V}} \tag{2.44}$$

$$\widehat{S}_{j-1}^{ind} = \frac{\frac{1}{N}\sum_{i=1}^{N} g_j(A)_i g_j(\mathbf{B}_{\mathbf{A}}^{(d)})_i - g_{j_0}^2}{\widehat{V}} \tag{2.45}$$

$$\widehat{ST}_{j-1}^{ind} = 1 - \frac{\frac{1}{N}\sum_{i=1}^{N} g_j(B)_i g_j(\mathbf{B}_{\mathbf{A}}^{(d)})_i - g_{j_0}^2}{\widehat{V}} \tag{2.46}$$

where g_{j_0} is the estimate of the mean and $\widehat{V} = \frac{1}{N}\sum_{i=1}^{N}(g_j(A)_i)^2 - g_{j_0}^2$

This procedure considers the estimator from Janon et al. [2014] and the overall cost is $4dN$ with N being the number of samples. However, another estimator can be used to estimate the indices. See Saltelli et al. [2010] for a review of various estimators of sensitivity indices.

2.2.3 Shapley effects

We are now changing the paradigm and we are taking advantage of game theory to understand how the reward is evenly distributed among the different players. The purpose of Sobol' indices is to decompose $\mathrm{Var}[Y]$ and allocate it to *each subset* $\mathcal{J} \subseteq \mathcal{D} = \{1, 2, \ldots, d\}$ whereas the Shapley effects decompose $\mathrm{Var}[Y]$ and allocate it to *each input* X_j. This difference allows us to consider any variable regardless of their dependence with other inputs.

One of the main issues in cooperative games theory is to define a relevant way to allocate the earnings between players. A fair share of earnings of a d players coalition has been proposed in Shapley [1953]. Formally, in Song et al. [2016] a d-player game with the set of players $\mathcal{D} = \{1, 2, \ldots, d\}$ is defined as a real-valued function that maps a subset of $\mathcal{D}$ to its corresponding cost, i.e., $c : 2^{\mathcal{D}} \mapsto \mathbb{R}$ with $c(\emptyset) = 0$. Hence, $c(\mathcal{J})$ represents the cost that arises when the players in the subset $\mathcal{J}$ of $\mathcal{D}$ participate in the game. The Shapley value of player j with respect to $c(\cdot)$ is defined as

$$v^j = \sum_{\mathcal{J} \subseteq \mathcal{D} \setminus \{j\}} \frac{(d - |\mathcal{J}| - 1)!|\mathcal{J}|!}{d!} \left(c\left(\mathcal{J} \cup \{j\}\right) - c\left(\mathcal{J}\right)\right) \tag{2.47}$$

where $|\mathcal{J}|$ indicates the size of $\mathcal{J}$. In other words, v^j is the incremental cost of including player j in set $\mathcal{J}$ averaged over all sets $\mathcal{J} \subseteq \mathcal{D}$ $\{j\}$.

This formula can be transposed to the field of global sensitivity analysis [Owen, 2014] if we consider the set of inputs of $\eta(\cdot)$ as the set of players $\mathcal{D}$. We then need to define a cost function $c(\cdot)$ such that for $\mathcal{J} \subseteq \mathcal{D}$, $c(\mathcal{J})$ measures the part of variance of Y caused by the uncertainty of the inputs in $\mathcal{J}$. To this aim, we want the cost function to verify $c(\emptyset) = 0$ and $c(\mathcal{D}) = \mathrm{Var}(Y)$. Two

functions can be used: $\tilde{c}(\mathcal{J}) = \mathrm{Var}\left[\mathbb{E}\left[Y|\mathbf{X}_{\mathcal{J}}\right]\right]/\mathrm{Var}(Y)$ and $c(\mathcal{J}) = \mathbb{E}\left[\mathrm{Var}\left[Y|\mathbf{X}_{-\mathcal{J}}\right]\right]/\mathrm{Var}(Y)$ which satisfy the two above conditions. According to Theorem 1 of Song et al. [2016], the Shapley values calculated using both cost functions $\tilde{c}(\mathcal{J})$ and $c(\mathcal{J})$ are equivalent. However, for some reasons described in Song et al. [2016], the cost function c instead of $\tilde{c}$ is privileged to define the Shapley effects. A valuable property of the Shapley effects defined in this way is that they are non-negative and they add up to one. Each one can therefore be interpreted as a measure of the part of the variance of Y related to the j-th input of η.

An issue with the Shapley value is its computational complexity, as all possible subsets of players need to be considered. Castro et al. [2009] proposed an estimation method based on an alternative definition of the Shapley values, which expresses the values in terms of all possible permutations of players. Let us denote by $\Pi(\mathcal{D})$ the set of all possible permutations with player set $\mathcal{D}$. Given a permutation $\pi \in \Pi(\mathcal{D})$, define the set $P_j(\pi)$ as the players that precede player j in π. Thus, the Shapley value can be rewritten in the following way :

$$v^j = \frac{1}{d!} \sum_{\pi \in \Pi(\mathcal{D})} \left[c\left(P_j(\pi) \cup \{j\}\right) - c\left(P_j(\pi)\right)\right] \tag{2.48}$$

From this formula, Castro et al. [2009] proposed to estimate v^j with $\widehat{v}^j$ by drawing randomly m permutations in $\Pi(\mathcal{D})$ and thus we have :

$$\widehat{v}^j = \frac{1}{m} \sum_{l=1}^{m} \Delta_j c(\pi_l) \tag{2.49}$$

where $\Delta_j c(\pi_l) = c\left(P_j(\pi) \cup \{j\}\right) - c\left(P_j(\pi)\right)$ and $c(\cdot)$ is the cost function.

An improvement of Castro's algorithm is proposed in Song et al. [2016] by including the Monte Carlo estimation $\widehat{c}$ of the cost function $c(\mathcal{J}) = \mathbb{E}\left[\mathrm{Var}\left[Y|\mathbf{X}_{-\mathcal{J}}\right]\right]/\mathrm{Var}(Y)$ to estimate the Shapley effects. The estimator writes:

$$\widehat{Sh}^j = \frac{1}{m} \sum_{l=1}^{m} \left[\widehat{c}\left(P_j(\pi_l) \cup \{j\}\right) - \widehat{c}\left(P_j(\pi_l)\right)\right] \tag{2.50}$$

where m refers to the number of permutations. Two estimation algorithms are defined, the main features of which are spelled out below:

- the *exact permutation method* if d is small, one does all possible permutations between the inputs (i.e. $m = d!$);
- the *random permutation method* which consists in randomly sampling m permutations of the inputs in $\Pi(\mathcal{D})$.

For each iteration of this loop on the permutations, a conditional variance expectation is computed. The cost C of these algorithms is the following $C = N_v + m(d-1)N_o N_i$ where N_v is the sample size for the variance computation of Y, N_o is the outer loop size for the expectation, N_i is the inner loop size for the conditional variance of Y and m is the number of permutations according to the selected method. Note that, in addition to the Shapley values, the full first-order Sobol' indices and the independent total Sobol' indices can also be obtained from these algorithms.

Based on theoretical results, Song et al. [2016] recommend setting the parameters at the following values to obtain an accurate approximation of Shapley effects that is computationally affordable:

- The *exact permutation method*: N_o as large as possible and $N_i = 3$;
- The *random permutation method*: $N_o = 1$, $N_i = 3$ and m as large as possible.

The choice of N_v is independent from these values and Iooss and Prieur [2017] illustrated the convergence of two numerical algorithms in the estimation of the Shapley effects.

2.3 Modeling Dependencies

Copula functions are great tools to model dependence structure of probability distributions [Sklar, 1959]. Multiple shapes of dependencies can be considered with multiple parametric families. In high dimension, other methods like *vine copulas* [Joe, 1994], can describe multidimensional dependencies by combining multiple pair copulas. A graphical tool in the form of tree describes the high dimensional dependence structure by connecting the bivariate copulas. This section introduces the notion of copula and the construction of vine copulas.

2.3.1 The Copula Theory

The word *copula*, that means "a link, tie, bond," was first employed by Sklar [1959] in the theorem which now bears his name. It is explained by Sklar [1996] as "*a grammatical term for a word or expression that links a subject and predicate,*" which is appropriate for a function that links a multidimensional distribution to its one-dimensional margins. Formally, a copula describes the dependence structure between a group of random variables. It is therefore a practical and powerful tool to describe dependencies. Many parametric copula families are available and are based on different dependence structures and parameters. Most of these families have bi-dimensional dependencies, but some can be extended to higher dimensions.

A copula is a multidimensional CDF linking the margins of $\mathbf{X} = (X_1, \ldots, X_d)$ to its joint distribution. With $F_j(x_j) = \mathbb{P}[X_j \leq x_j]$ for $j = 1, \ldots, d$, the continuous CDF of X_j, we can denote the random vector

$$(U_1, \ldots, U_d) = (F_1(X_1), \ldots, F_d(X_d))$$

with uniformly distributed margins. The copula, denoted by C, is the joint CDF of $(U_1, \ldots, U_d)$, such as for $(u_1, \ldots, u_d) \in [0,1]^d$,

$$C(u_1, \ldots, u_d) = \mathbb{P}[U_1 \leq u_1, \ldots, U_d \leq u_d].$$

Definition 2. *A d-dimensional copula is a function $C : [0,1]^d \to [0,1]$ such that [Nelsen, 2007],*

- $C(u_1, \ldots, u_d) = 0$ *if* $u_j = 0$ *for at least one* $j = 1 \ldots d$,

- *the margins of* C *are uniformly distributed, i.e.,* $C(1,\dots,u_j,\dots,1) = u_j$*, for each* $j = 1,\dots,d$*,*
- C *is* d*-non-decreasing, i.e., the volume of each hyper-rectangle* $B = \prod_{j=1}^{d}[a_j, b_j] \subseteq [0,1]^d$ *is non-negative :*

$$\sum sgn(u_1,\dots,u_d)C(u_1,\dots,u_d) \geq 0$$

where the sum is taken among the vertices of B *and*

$$sgn(u_1,\dots,u_d) = \begin{cases} 1 & \text{if } u_k = a_k \text{ for any even number } k = 1,\dots,d\ , \\ -1 & \text{if } u_k = a_k \text{ for any odd number } k = 1,\dots,d\ , \end{cases}$$

with the example of a bivariate copula $C : [0,1]^2 \to [0,1]$*, for all* $0 \leq u_1 \leq u_2 \leq 1$ *and* $0 \leq v_1 \leq v_2 \leq 1$ *we have* $C(u_2,v_2) - C(u_2,v_1) - C(u_1,v_2) + C(u_1,v_1) \geq 0$*.*

The first property says that the copula C is zero on the lower bounds. The second involves the uniformity of the margins. The thirs states that the copula is bounded, and the fourth shows that C is d-increasing so long as the probability measure in any hyper-rectangle B embedded in the unit square is positive.

An interesting property of a d-variate copula C is that it satisfies the Lipschitz condition, such as for all $\mathbf{u}, \mathbf{v} \in [0,1]^d$ we have

$$|C(\mathbf{u}) - C(\mathbf{v})| \leq \sum_{j=1}^{d} |u_j - v_j|.$$

Consequently copulas are uniformly continuous. Moreover, if all marginal CDFs are continuous, then it exists a unique copula C associated to F. The converse is also true. The following theorem, called *Sklar theorem* from Sklar [1959], describes this fundamental result in copula modeling.

Theorem 1. *Let* $F_{\mathbf{X}}$ *be a* d*-dimensional CDF with continuous margins* $F_1,\dots,F_d$*. Then, there exists a unique copula such as*

$$F(x_1,\dots,x_d) = C\left(F_1(x_1),\dots,F_d(x_d)\right). \tag{2.51}$$

Conversely, if C *is a copula and* $F_1,\dots,F_d$ *are CDFs, then the function* F *described by Equation (2.51) is a* d*-variate CDF with margins* $F_1,\dots,F_d$*. Furthermore,* C *is unique if* $F_1,\dots,F_d$ *are continuous.*

One can also rewrite (2.51) for $(u_1,\dots,u_d) \in [0,1]^d$ as

$$C(u_1,\dots,u_d) = F(F_1^{-1}(u_1),\dots,F_d^{-1}(u_d)).$$

Equation (2.51) shows that the joint CDF F can be constructed with the margins F_j for $j = 1,\dots,d$ and a copula C. Therefore, the copula is a perfect description of the dependence structure.

Definition 3. *A d-dimensional copula C is said to be absolutely continuous if $C(u_1, \ldots, u_d)$ has a density $c(u_1, \ldots, u_d)$ and for all $(u_1, \ldots, u_d) \in [0,1]^d$,*

$$C(u_1, \ldots, u_d) \equiv \int_0^{u_1} \ldots \int_0^{u_d} c(s_1, \ldots, s_d) \mathrm{d}s_1, \ldots, \mathrm{d}s_d.$$

If copulas are absolutely continuous (regarding the Lebesgue measure), Theorem 1 also states that the joint PDF f is decomposed in the product of the margins PDF and the density of the copula c, such that

$$f(x_1, \ldots, x_d) = c(F_1(x_1), \ldots, F_d(x_d)) \prod_{j=1}^{d} f_j(x_j).$$

Definition 4. *The copula density c, if it exists, is defined by*

$$c(u_1, \ldots, u_d) = \frac{\partial^d}{\partial u_1 \ldots \partial u_d} C(u_1, \ldots, u_d).$$

Any copula C has natural upper and lower bounds: the so-called Fréchet-Hoeffding copula bounds. These bounds, describes in Definition 5, are functions that bounds the copula C as stated in Theorem 2. The name states in recognition of the independent work from Hoeffding and Fréchet [Hoeffding, 1940; Frechet, 1951] during Second World War.

Definition 5. *The Fréchet-Hoeffding upper bound is given by the function*

$$M(u_1, \ldots, u_d) = \min_{j=1,\ldots,d} u_j$$

and the Fréchet-Hoeffding lower bound is given by the function

$$W(u_1, \ldots, u_d) = \max \left\{ \sum_{j=1}^{d} u_j + 1 - d, 0 \right\}.$$

The upper bound M is sharp such that it is always a copula and corresponds to comonotone random variables. However, the lower bound W is point-wise sharp and is a copula only for $d = 2$, which corresponds to countermonotonic random variables.

Theorem 2 (Fréchet–Hoeffding Theorem)**.** *For every copula C and for any vector $\mathbf{u} \in [0,1]^d$, C satisfies:*

$$W(\mathbf{u}) \leq C(\mathbf{u}) \leq M(\mathbf{u}).$$

The dependence between two random variables can be characterized by dependence measures such as the Spearman's ρ or the Kendall's τ and can directly obtained from the copula of a joint distribution [Nelsen, 2007]. These two dependence measures between two random variables X_1 and X_2 are respectively defined by

$$\rho_S(X_1, X_2) = 12 \int \int_{[0,1]^2} C(u, v) \mathrm{d}u\mathrm{v} - 3$$

and

$$\tau(X_1, X_2) = 4 \int \int_{[0,1]^2} C(u, v) \mathrm{d}C(u, v) - 1.$$

2.3.2 Vine copulas

The theorem of Sklar states that a copula C is unique and admits a density c if the margins $F_1, \dots, F_d$ are continuous. Using the chain rule, the joint density f associated to the random vector $\mathbf{X} = (X_1, \dots, X_d)$ can be written as

$$f(x_1, \dots, x_d) = c(F_1(x_1), \dots, F_d(x_d)) \times \prod_{j=1}^{d} f_j(x_j),$$

where $f_j(x_j)$, $j = 1, \dots, d$ are the marginal densities of $\mathbf{X}$. By recursive conditioning, the joint density can also be written as

$$f(x_1, \dots, x_d) = f_d(x_d) \times f(x_{d-1}|x_d) \times f(x_{d-2}|x_{d-1}, x_d) \times \cdots \times f(x_1|x_2, \dots, x_d), \quad (2.52)$$

which is unique up to a re-labelling of the variables. For instance, in the bivariate case

$$f(x_1, x_2) = c_{1,2}(F_1(x_1), F_2(x_2)) \times f_1(x_1) \times f_2(x_2),$$

where $c_{1,2}$ is the pair-copula density. For a conditional density, one can derived that

$$f(x_1|x_2) = c_{1,2}(F_1(x_1), F_2(x_2)) \times f_1(x_1).$$

This can be used for instance to decompose the factor $f(x_{n-1}|x_n)$ in equation (2.52) into pair-copula and a marginal density. For three random variables,

$$f(x_1|x_2, x_3) = \frac{f(x_1, x_2|x_3)}{f(x_2|x_3)} = c_{1,2|3}(F(x_1|x_3), F(x_2|x_3)) \times f(x_1|x_3), \quad (2.53)$$

where the second equality comes from Sklar's theorem. Note that another decomposition is

$$f(x_1|x_2, x_3) = c_{1,3|2}(F(x_1|x_2), F(x_3|x_2)) \times f(x_1|x_2) \quad (2.54)$$

where $c_{1,3|2}$ is different from the pair-copula in (2.53). Decomposing for instance $f(x_1|x_2)$ in (2.54) leads to a pair-copula decomposition. It follows that each term in (2.52) can be decomposed into a product of pair copula and marginal density using the general formula

$$f(x_k|\mathbf{v}) = c_{k,j|-j}(F(x_k|\mathbf{v}_{-j}), F(v_j|\mathbf{v}_{-j})) \times f(x_k|\mathbf{v}_{-j}), \quad (2.55)$$

where $\mathbf{v}$ is a d-dimensional vector, v_j is one arbitrary chosen component of $\mathbf{v}$ and $\mathbf{v}_{-j}$ is $\mathbf{v}$ excluding the component j. By combining the two previous results (2.52) and (2.55) we may derive a decomposition of $f(x_1, \dots, x_d)$ that only consists of marginal distributions and $\frac{d(d-1)}{2}$ bivariate copulas. Such a decomposition, called the pair-copula construction (PCC), has been introduced by Joe [1996]. This allows to model more flexible dependence structures using only bivariate copulas.

A re-labeling of the variables can lead to a large number of different PCCs. Depending on the order of conditional variables, there are multiple ways to create a PCC. The number of possible decompositions increases significantly with the dimension d. To help organize them, a graphical model called regular vine (R-vine) is introduced by Bedford and Cooke [2001, 2002] and detailed in

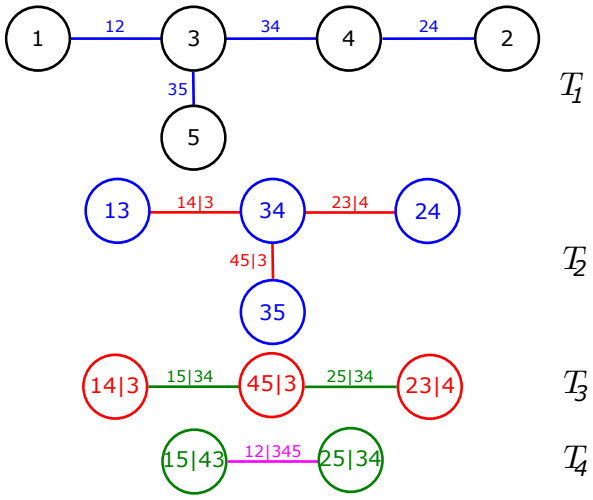


Figure 2.5 *R-vine structure of a $d = 5$ problem.*

Kurowicka and Cooke [2006]. Each of these sequences of tree give a specific way of decomposing the density. A R-vine describe a d-dimensional PCC and is a sequence of linked trees where the nodes and edges correspond to the $d(d-1)/2$ pair-copulas. According to Definition 8 of Bedford and Cooke [2001], a R-vine consists of $d-1$ trees $T_1, \dots, T_{d-1}$ with several constraints. Each tree T_i is composed of $d-i+1$ nodes which are linked by $d-i$ edges for $i = 1, \dots, d-1$. A node in a tree T_i must be an edge in the tree T_{i-1}, for $i = 2, \dots, d-1$. Two nodes in a tree T_i can be joined if their respective edges in tree T_{i-1} share a common node, for $i = 2, \dots, d-1$. The pair copula in the first tree characterize pairwise unconditional dependencies, while the pair copula in higher order trees model the conditional dependency between two variables given a set of variables. The number of conditioning variables grows with the tree order. Note that a PCC where all trees have a path like structure define the D-vine subclass while the star like structures correspond to C-vine subclass [Bedford and Cooke, 2001].

The following example explains the density construction of a R-vine in dimension 5. A PCC can be derived from Equation (2.52) and (2.55), for instance

$$
\begin{aligned}
f(x_1,x_2,x_3,x_4,x_5) = {} & f_1(x_1)f_2(x_2)f_3(x_3)f_4(x_4)f_5(x_5) \text{(margins)} \\
\text{(unconditional pairs)} \times {} & c_{12}(F_1(x_1),F_2(x_2)) \cdot c_{35}(F_3(x_3),F_5(x_5)) \cdot c_{34}(F_3(x_3),F_4(x_4)) \cdot c_{24}(F_2(x_2),F_4(x_4)) \\
\text{(conditional pair)} \times {} & c_{14|3}(F_{1|3}(x_1|x_3),F_{4|3}(x_4|x_3)) \cdot c_{23|4}(F_{2|4}(x_2|x_4),F_{3|4}(x_3|x_4)) \cdot c_{45|3}(F_{4|3}(x_4|x_3),F_{5|3}(x_5|x_3)) \\
\text{(conditional pair)} \times {} & c_{15|34}(F_{1|34}(x_1|x_3,x_4),F_{5|34}(x_5|x_3,x_4)) \cdot c_{25|34}(F_{2|23}(x_3|x_3,x_3),F_{5|23}(x_5|x_3,x_4)) \\
\text{(conditional pair)} \times {} & c_{12|345}(F_{1|345}(x_1|x_3,x_4,x_5),F_{2|345}(x_2|x_3,x_4,x_5)) \qquad (2.56)
\end{aligned}
$$

This joint density is associated to a specific R-vine, which is illustrated in Figure 2.5.

For R-vines in general, there was no efficient way of storing the indices of the pair-copulas required in the joint density. Morales Nápoles [2010]; Dissmann et al. [2013] proposed a way to store the construction in a matrix. This approach uses the specification of a lower triangular

matrix A where the entries belong to $1, \dots, d$.

Such matrix representation of R-vines allows to directly derived the tree structure (or alternatively the PCC of the R-vine distribution). Each row from 2 to d provides a tree where each considered pair is identified by a diagonal entry and by the corresponding column entry of the row under consideration. The conditioning sets are given by the column entries below this row. The row number 2 defines the lowest tree whereas row number d defines the top of the tree structure (unconditional pairs). Note that such matrix representation also allows to store the copula parameter and type of each bivariate copula. Dissmann et al. [2013] show that the R-vine density associated to the R-vine matrix $A = \{a_{i,j}\}_{i,j=1,\dots,d}$ is given by the following equation

$$f(x_1, \dots, x_d) = \underbrace{\prod_{k=1}^{d} f_k(x_k)}_{\text{marginal densities}} \times \underbrace{\prod_{j=d-1}^{1} \prod_{i=d}^{j+1} c_{a_{j,j},a_{i,j}|a_{i+1,j},\dots,a_{n,j}}}_{\text{pair copula densities}}$$

where pair-copulas arguments are $F(x_{a_{j,j}}|x_{a_{i+1,j}}, \dots, x_{a_{n,j}})$ and $F(x_{a_{i,j}}|x_{a_{i+1,j}}, \dots, x_{a_{n,j}})$.

The following example illustrates the lecture of the R-vine matrix associated to the PCC given in (2.56). The following R-vine matrix A is a way to store the information in (2.56) :

$$A = \begin{pmatrix} 1 & & & & \\ 2 & 2 & & & \\ 5 & 5 & 4 & & \\ 4 & 3 & 5 & 3 & \\ 3 & 4 & 3 & 5 & 5 \end{pmatrix}$$

The element of the diagonal represents the first variable of the pair and another element of the same column is the second variable. The elements that are below the second variable in the matrix are the conditional variables. The array is read from the top left to the bottom right of the matrix. The first element of the matrix is the pair $X_1, X_2|X_3, X_4, X_5$ which is the last pair of the R-vine. Then we climb the trees by taking the second pair of the first column $(X_1, X_5|X_3, X_4)$ and the first pair of the second column $(X_2, X_5|X_3, X_4)$. We do this recursively until the R-vine is complete.

2.4 Surrogate models

Reliability analysis or sensitivity analysis on numerical models may require a high number of model evaluations in order to obtain a precise measure of the estimated quantity. However, when the model is too costly and cannot be run in a reasonable timeframe, one can substitute the original model with a mathematical approximation built from a set of sample data. This approximated model is referred as a surrogate model (or metamodel). In this section, we make an overview of two important techniques used in the following chapters of this thesis: Random forests and Gaussian processes. In the case of random forest, we also introduce the permutation variable importance measure and its connection with Sobol' indices.

2.4.1 Random forests

2.4.1.1 Definition

Let $\mathcal{D}_n = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_d, Y_d)\}$ be a training set of n replications of $(\mathbf{X}, Y)$, where $\mathbf{X}_i = (X_{i,1}, \ldots, X_{i,d})$. Random forests, introduced by Breiman [2001], are an ensemble method that combines Classification And Regression Trees (CART) [Breiman et al., 1984]. The prediction accuracy of a random forest is highly increased compared to individual CART trees. Indeed, the algorithm generates bootstrap samples from the original sample $\mathcal{D}_n$ and fits an unpruned CART tree on each bootstrap sample. The randomness from the bootstrap sampling gives slightly different learning samples for each tree which adjust the instability from a single tree prediction. Random forests can be used in classification or regression problems. However, in this document, we focus on the use of random forests for regression tasks.

A CART tree is generated by recursively splitting its training sample into new homogeneous subspaces (called *nodes*). Starting from the *root* node, a *cut* is processed among one of the d variables to split the current node into two new partitions, also called child nodes. The cut is chosen such as the loss of variance is maximized between the parent node and the child nodes. When no more cuts can decrease the variance of a node, it is called a *leaf* and has minimum variance. A tree is considered maximum when no more nodes can be split. In a classical CART algorithm, a *pruning* procedure can follow the procedure by unifying leaves in order to improve the bias-variance trade-off. However, in random forest algorithm, this step is not applied and only unpruned trees are considered.

The random forest algorithm constructs a predictor from a collection of M randomized regression trees. The predicted value of the m-th tree at a new query point $\mathbf{x}$ is denoted by $\hat{\eta}(\mathbf{x}; \Theta_m, \mathcal{D}_n)$, where $\Theta_1, \ldots, \Theta_M$ are independent random variables, distributed the same as a generic random variable Θ and independent of $\mathcal{D}_n$. The variable Θ is used to randomize the trees by resampling the training set $\mathcal{D}_n$ prior to the growing of individual trees. The trees are then combined to form the (finite) forest estimate

$$\hat{\eta}_M(\mathbf{x}; \Theta_1, \ldots, \Theta_M, \mathcal{D}_n) = \frac{1}{M} \sum_{m=1}^{M} \hat{\eta}(\mathbf{x}; \Theta_m, \mathcal{D}_n). \tag{2.57}$$

In a modeling point of view, it makes sense to make M tend towards infinity, and consider the (infinite) forest estimate

$$\hat{\eta}_\infty(\mathbf{x}; \mathcal{D}_n) = \mathbb{E}_\Theta \left[\hat{\eta}(\mathbf{x}; \Theta, \mathcal{D}_n) \right], \tag{2.58}$$

where $\mathbb{E}_\Theta$ denotes the expectation with respect to the random parameter Θ, conditionally to $\mathcal{D}_n$. The random forest algorithm constructs a number T of trees grown from different bootstrap samples of $\mathcal{D}_n$. The training sample $\mathcal{D}_n(\Theta_m)$, of a tree m is then a sub-sample of $\mathcal{D}_n$. This randomness makes the trees in a random forest grow differently, which diversify the results and increases the prediction accuracy. Another hyper-parameter of the random forest algorithm, called *mtry*, randomly selects a subset of candidate variables among $X_1, \ldots, X_d$ and selects the cutting variable among this subset. This feature adds additional randomness and can diversify the variables selected for the cutting in each tree.

Random forest became more and more popular with broad applications to machine learning and data mining. However, it is not clearly elucidated from a mathematical point of view. Consistency results where shown for different variants [Biau et al., 2008; Scornet et al., 2015; Ishwaran and Kogalur, 2010; Denil et al., 2014] as for the estimation of confidence intervals [Wager et al., 2014; Mentch and Hooker, 2016].

The m-th bootstrap of $\mathcal{D}_n$, denoted by $\mathcal{D}_n(\Theta_m)$, is used to train the tree m. The out of sample data of $\mathcal{D}_n(\Theta_m)$, called *out-of-bag* (OOB) sample [Breiman, 1996], is denoted by $\bar{\mathcal{D}}_n(\Theta_m) = \mathcal{D}_n \backslash \mathcal{D}_n(\Theta_m)$. Since OOB sample are observations not used to train the m-th forest, its main utility is to be used as a test sample to estimate the prediction error of a tree. The OOB error is defined by

$$\widehat{OOB}_m = \frac{1}{|\bar{\mathcal{D}}_n(\Theta_m)|} \sum_{i:(\mathbf{X}_i, Y_i) \in \bar{\mathcal{D}}_n(\Theta_m)} \ell\Big(Y_i, \hat{\eta}(\mathbf{X}_i; \Theta_m, \mathcal{D}_n)\Big) \tag{2.59}$$

where ℓ is a loss function and $|\bar{\mathcal{D}}_n(\Theta_m)|$ is the sample size of the OOB sample $\bar{\mathcal{D}}_n(\Theta_m)$. Usually in standard regression, the mean squared error $\ell(y, q) = (y - q)^2$ is considered, making (2.59) written as

$$\widehat{OOB}_m = \frac{1}{|\bar{\mathcal{D}}_n(\Theta_m)|} \sum_{i:(\mathbf{X}_i, Y_i) \in \bar{\mathcal{D}}_n(\Theta_m)} \Big(Y_i - \hat{\eta}(\mathbf{X}_i; \Theta_m, \mathcal{D}_n)\Big)^2. \tag{2.60}$$

The interest of this method, compared to cross-validation techniques, is that no test sample of new observations is necessary to evaluate the model accuracy. However, Genuer et al. [2010] highlighted that while the OOB error can be slightly optimistic, there is still a fair representation of the model accuracy and therefore it can be very interesting for model comparison.

2.4.1.2 Permutation variable importance

The cutting rule of decision trees makes simpler their visualization and interpretation. However, this is lost in the case of random forests due to the randomness and the large number of trees. In order to understand the behavior of random forest, variable importance (VI) measures were introduced to describe the link between the input variables and the target. One of the most commonly used VI measures in random forest are the permutation based VI (PVI) [Breiman, 2001]. The PVI value of a variable X_j is described by the increase of OOB error when the values of X_j are randomly permuted in the OOB sample. The effect of a permutation is to break the link between X_j and Y. The more a permutation of X_j increases the error and the more the variable is important, and inversely. Let $\widehat{OOB}^j_m$ be the error over $\bar{\mathcal{D}}^j_n(\Theta_m)$ which is the OOB sample of the m-th tree where X_j is randomly permuted. The estimated PVI value of a random forest model is defined as the mean increase of OOB error over all trees:

$$\hat{\mathcal{I}}_{\hat{\eta}_M}(X_j) = \frac{1}{M} \sum_{m=1}^{M} \Big[\widehat{OOB}^j_m - \widehat{OOB}_m\Big]. \tag{2.61}$$

Different studies formalized this empirical measure [Ishwaran et al., 2007; Zhu et al., 2015] and stated that (2.61) is an estimator of

$$\mathcal{I}_\eta(X_j) = \mathbb{E}\Big[(Y - \eta(\mathbf{X}_{(j)}))^2\Big] - \mathbb{E}\Big[\,(Y - \eta(\mathbf{X}))^2\,\Big]. \tag{2.62}$$

where $\mathbf{X}_{(j)} = (X_1, \ldots, X'_j, \ldots, X_d)$ is a random vector such as X'_j is an independent copy of X_j and independent to Y and the other variables. The first term of (2.62) measures the increase of error due to a permutation of X_j whereas the second withdraws the noise in the case of stochastic model: $Y = \eta(\mathbf{X}) + \epsilon$, where ϵ is a random noise. However, in our context, we suppose the model is deterministic, such as $Y = \eta(\mathbf{X})$, making the second term null. We can then simplify (2.62) as

$$\mathcal{I}_\eta(X_j) = \mathbb{E}\Big[(\eta(\mathbf{X}) - \eta(\mathbf{X}_{(j)}))^2\Big]. \tag{2.63}$$

Little is known regarding the inner working of these indices. A detailed theoretical development of a simplified version of PVI values has been done in Ishwaran et al. [2007], while Zhu et al. [2015] showed convergence of the estimator of the PVI. Given the difficulties to obtain theoretical results with these indices, Ishwaran and Lu [2018] proposed to approximate the distribution of the PVI values through some form of resampling.

Recent work made a theoretical connection between the PVI values and the total Sobol' indices in sensitivity analysis [Gregorutti et al., 2015; Wei et al., 2015] which helps to understand the behavior of the PVI values.

2.4.1.3 Relationship with the total Sobol' indices

The relation between the total Sobol' indices and the PVI values was established in Gregorutti et al. [2015]; Wei et al. [2015] and is given by the following proposition.

Proposition 1. *Suppose $\mathbf{X}$ is a random-vector of independent components and $Y = \eta(\mathbf{X})$. Then for any component X_j we have*

$$ST_j = \frac{\mathcal{I}_\eta(X_j)}{2\mathrm{Var}[Y]}. \tag{2.64}$$

This relation helps to understand the effect of a permutation on X_j and is an important link between sensitivity analysis of model output and feature importance in data science. The total Sobol' index can be estimated using the PVI estimation. However, no relation has been established for the first-order Sobol' index. The Chapter 3 introduces a modified version of the PVI which is related to the first-order Sobol' index.

2.4.2 Gaussian Processes

Kriging, also referred to as Gaussian process (GP), was first introduced in the field of geostatistics in Krige [1951] and formalized by Matheron [1962]. The method consider the values of a costly numerical model η in which the interpolated values are modeled by a Gaussian process. More precisely, it is based on the assumption that the function $\eta(\mathbf{x})$ is the realization of a Gaussian random process. The Gaussian hypothesis then provides an explicit formula for the law of the process conditionally to the value taken by η on a design of experiments $\mathbf{D}$. We refer to Santner et al. [2003] for a more detailed review of the topic.

The function $\eta(\mathbf{x})$ is assumed to be a realization of a Gaussian random process H. Let μ be the mean of the random field H and Z be a zero-mean stationary GP. The GP can be defined by

$$H(\mathbf{x}) = \mu(\mathbf{x}) + Z(\mathbf{x}). \tag{2.65}$$

The mean μ is usually defined as a linear combination of deterministic functions $\{f_i\}_{i=1,\dots,p}$ such that $\mu(\mathbf{x}) = \mathbf{f}(\mathbf{x})^T\boldsymbol{\beta}$. The first term of (2.65) is the global trend of the model output which can be constant, linear or polynomial. The second term corresponds to a local perturbation of the model output. This assumption can be limited because of the choice of decoupling large-scale (the mean μ) and small-scale (the supposed stationary random field Z) effects. Thus, the nature of μ is assumed with a constant or linear trend in practice and Z is assumed to be a second-order stationary field such that Z has a constant finite variance $\sigma^2(\mathbf{x}) = \sigma^2$ and its covariance kernel (or autocorrelation function) $R(\mathbf{x}, \tilde{\mathbf{x}})$ is only function of the shift $\mathbf{x} - \tilde{\mathbf{x}}$.

Let $\mathcal{D}_n = \{(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_d, Y_d)\}$ be a training set of n replications of $(\mathbf{X}, Y)$ using the true model η. We also introduce $\boldsymbol{\mathcal{X}} = \{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}\}$ as the input samples and $\mathcal{Y} = \{y^{(k)} = \eta(\mathbf{x}^{(k)}\}_{k=1,\dots,n}$ as the output samples. The GP approximation builds a distribution of H using the n observations of $\mathcal{D}_n$. Determining the conditional distribution is not straightforward. It can be obtained as a constrained optimization problem. The following matrix notation is used in the following:

$$\begin{aligned}
\mathbf{r}(\mathbf{x}) &= \{R(\mathbf{x}, \mathbf{x}^{(1)}), \dots, R(\mathbf{x}, \mathbf{x}^{(n)})\} \\
\mathbf{F} &= \Big(f_i(x^{(k)})\Big)_{1\leq i\leq p, 1\leq k\leq n} \\
\mathbf{R} &= \Big(R(\mathbf{x}^{(k)}, \mathbf{x}^{(l)})\Big)_{1\leq k\leq n, 1\leq l\leq n},
\end{aligned}$$

where $\mathbf{F}$ and $\mathbf{R}$ are referred to as the experiments and autocorrelation matrices. Thus, the conditional model response at an given point $\mathbf{x}$ is normally distribution with a law $\mathcal{N}(\mu_H(\mathbf{x}), \sigma_H(\mathbf{x}))$ such that:

$$\mu_H(\mathbf{x}) = \mathbf{f}^T(\mathbf{x})\boldsymbol{\beta} + \mathbf{r}^T(\mathbf{x})\mathbf{R}^{-1}(\mathcal{Y} - \mathbf{F}\boldsymbol{\beta}) \tag{2.66}$$

$$\sigma_H^2 = \sigma^2 - \mathbf{r}^T(\mathbf{x})\mathbf{R}\mathbf{r}(\mathbf{x}). \tag{2.67}$$

These equations involve properties of H (i.e., mean, autocovariance) that are unknown in practice. The autocorrelation function has to be selected and is usually defined in the form of tensorized stationary functions defined by:

$$R(\mathbf{x}, \tilde{\mathbf{x}}) = \prod_{i=1}^{n} R_i(\mathbf{x}_i - \tilde{\mathbf{x}}).$$

The correlation between two realizations only depends on their distance to one another. Several types of autocorrelation functions are used in practice:

- linear:

$$R(\mathbf{x} - \tilde{\mathbf{x}}, \mathbf{l}) = \prod_{i=1}^{n} \max\left(0, 1 - \frac{\mathbf{x}_i - \tilde{\mathbf{x}}_i}{l_i}\right)$$

- exponential:

$$R(\mathbf{x} - \tilde{\mathbf{x}}, \mathbf{l}) = \exp\left(-\sum_{i=1}^{n} \frac{|\mathbf{x}_i - \tilde{\mathbf{x}}_i|}{l_i}\right)$$

- Gaussian

$$R(\mathbf{x} - \tilde{\mathbf{x}}, \mathbf{l}) = \exp\left(-\sum_{i=1}^{n}\left(\frac{\mathbf{x}_i - \tilde{\mathbf{x}}_i}{l_i}\right)^2\right)$$

where $\mathbf{l} = \{l_i\}_{i=1,\dots,n}$ are the scale parameters.

Therefore, a Gaussian process can be fitted by estimating its parameters: the mean (i.e., regression coefficients defined by $\boldsymbol{\beta}$), the variance σ^2 and the autocorrelation functions parameters $\mathbf{l}$. A *maximum likelihood estimation* is used for the estimation. The estimation of the regression parameters $\boldsymbol{\beta}$ is a function of $\mathbf{l}$ such that:

$$\hat{\boldsymbol{\beta}} = (\mathbf{F}^T \mathbf{R}_{\mathbf{l}}^{-1} \mathbf{F})^{-1} \mathbf{F}^T \mathbf{R}_{\mathbf{l}}^{-1} \mathcal{Y}$$

where the subscript $\mathbf{l}$ denotes the dependency of $\mathbf{R}$ with $\mathbf{l}$. The estimation of the variance σ^2 is also a function of l and is given by

$$\hat{\sigma}_H^2 = \frac{1}{n}(\mathcal{Y} - \mathbf{F}\boldsymbol{\beta}^T)^T \mathbf{R}_{\mathbf{l}}{}^{-1} (\mathcal{Y} - \mathbf{F}\boldsymbol{\beta}^T)$$

The estimation of the autocorrelation parameters $\mathbf{l}$ is obtained by solving the optimization problem:

$$\hat{\mathbf{l}} = \underset{\mathbf{l}}{\operatorname{argmin}} (\det \mathbf{R}_{\mathbf{l}})^{1/n} \hat{\sigma}_H^2$$

The mean is the last quantity to be estimated. Its calculation is based on the other computed parameters $\hat{\boldsymbol{\beta}}$, $\hat{\sigma}_H^2$ and $\hat{\mathbf{l}}$, such that:

$$\hat{\mu}_H(\mathbf{x}) = \hat{\eta}(\mathbf{x}) = \mathbf{f}^T(\mathbf{x})\hat{\boldsymbol{\beta}} + \mathbf{r}_{\hat{\beta}}(\mathbf{x})^T \mathbf{R}_{\hat{\mathbf{l}}}{}^{-1} (\mathcal{Y} - \mathbf{F}\hat{\boldsymbol{\beta}}).$$

Finally, the obtained metamodel $\hat{\eta}(\mathbf{x})$ interpolates the output observations in $\boldsymbol{\mathcal{X}}$. Moreover, from the Gaussian assumption, one can estimate the variance of the model (also called Kriging variance), defined by

$$\hat{\sigma}_H^2(\mathbf{x}) = \hat{\sigma}_H^2 - \mathbf{r}_{\hat{\beta}}(\mathbf{x})^T \mathbf{R}_{\hat{\mathbf{l}}} \mathbf{r}_{\hat{\beta}}(\mathbf{x}).$$

This estimated variance is very useful for deriving confidence intervals on the predictions of $\hat{\eta}(\mathbf{x})$.

Chapter 3

Sensitivity Analysis and Random Forest

Abstract. Variance-based methods have been shown to be very popular in sensitivity analysis of model output. These techniques describe the sensitivity pattern of a model with its random inputs using, whereby the variance of a model output can be decomposed into terms which can be attributed to the inputs or their interactions. The first-order Sobol' index captures the individual effect of an input on the output variance, while the total Sobol' effect gathers the marginal and interaction effects. The permutation variable importance (PVI) based on random forest is an effective and popular technique in machine learning for measuring variable importance. Previous studies have shown a relationship between the PVI values and the total Sobol' indices. In this chapter, we propose new PVI values which are related to the first-order Sobol' indices. We also introduce a new permutation procedure to deal with dependent variables, using a Rosenblatt transformation. We then show that this new procedure leads to PVI values which are connected to the full and independent Sobol' indices. A numerical study shows that the permutation procedure can be a good alternative to the Monte Carlo sampling to estimate the Sobol' indices. Furthermore, using a random forest model, we identified a systematic bias which is strongly related to the accuracy of the estimated model. We observed that this bias is also significantly increased when variables are more dependent. We also numerically show with toy examples that the new permutation procedure using the Rosenblatt transformation is a suitable alternative for measuring the influence of dependent variables, as it is less subject to the identified bias.

Contents

3.1 Introduction 50
3.2 New permutation variable importance values and their relations with the Sobol' indices 52
3.3 Sobol indices estimation and comparison 54
3.4 Using Random Forest 64
3.5 Conclusion 72
3.6 Appendix 73

3.1 Introduction

In many disciplines, such as risk analysis, numerical models are greatly used to approximate the behavior of physical phenomenon. They allow engineers to get rid of expensive or unfeasible real experiments and have a further understanding of the natural system. With the increase of computing power, models have a more accurate representation of reality. However, the accuracy increases the complexity, making the model interpretation more difficult. A model is described by input parameters, sometimes in a large number, with more or less influence on the prediction. Oftentimes, the model inputs are subject to uncertainties due to the lack of knowledge of their exact values, which means that the resulting output can be regarded as random (e.g., the wind speed at the site of a forest [Salvador et al., 2001]). In this case, a good practice is to use Uncertainty Analysis (UA) to quantify the overall uncertainty in the model response. However, some of the input variables strongly affect the model output while others have a small effect.

In addition to UA studies, Sensitivity Analysis (SA) tries to quantify these effects by evaluating how much each input is contributes to the model output. Two types of SA methods exists: local and global. Local SA measures the contribution of the inputs around a particular value. However, as mentioned by Saltelli et al. [2008], local methods are difficult to justify when dealing with random inputs since we cannot know the particular value of the inputs. On the other hand, Global SA (GSA) measures how the uncertainty in the model output can be apportioned to the uncertainty in the inputs over their whole variation range [Saltelli et al., 2000]. Several GSA methods exist and target different quantities of the model output distribution. Variance-based methods decompose the output variance and allocate it to the variance of each input, leading to the Sobol' indices. A first-order effect [Sobol, 1993] measures the expected variance reduction of the output when an input is fixed, whereas the total effect [Homma and Saltelli, 1996] measures the expected remaining variance of the output when all other input values are fixed. The Goal-Oriented SA[Fort et al., 2016] is a more generalized version of the first-order effect. It measures the importance of an input depending on the selected quantity of interest of the output distribution (e.g., quantile [Browne et al., 2017; Maume-Deschamps and Niang, 2018]). Other techniques try to compare the whole output distribution with its conditional counterpart when one of the input variables is fixed [Borgonovo, 2007; Da Veiga, 2015]. However, among these techniques, the Sobol' indices are one of the most popularly used, and have gained the most attention as they quantify the individual and total contributions of each input. Many estimation

techniques and algorithms have been proposed to compute the Sobol' indices (see for example Tarantola et al. [2006]; Saltelli et al. [2010]).

In data science, the volume of available data is expanding dramatically in many disciplines and their exploitation is a topical challenge, such as data linked to observations of real-life phenomena. Machine Learning techniques consider these data and try to build predictive models that approximate these phenomena. They are very similar to what computer codes were designed for. The prime difference stems from their design: a computer code is mainly developed based on knowledge and application theories(e.g., Navier–Stokes equations, Black–Scholes model) while machine learning models are built entirely by inferring the observed data. Similarly to SA, it is important to understand why and how a model made a certain prediction. Thus, many machine learning methods provide Variable Importance (VI) to measure the influence of each variable on the model. Two of the most popular methods that provide VI measures are linear regression models [Thomas et al., 1998; Johnson and LeBreton, 2004] and tree ensemble methods such as gradient boosting machines [Friedman, 2001] and Random Forests [Breiman, 2001]. VI measures from linear regression and random forest models have been compared in Grömping [2009]. They concluded that VI from linear regression models are only suitable to linear phenomenon unlike random forest models, which are non-parametric and non-linear models. A random forest is developed for classification or regression and is not only used as a predictor but also for its VI measures. Three primary ways of computing VI values exist for random forest: split count [Chen and Guestrin, 2016], gain (also called mean decrease in impurity) [Breiman et al., 1984] and permutation (also called mean decrease in accuracy) [Breiman, 2001]. The gain approach has been shown to overestimate the importance of the categorical variables with more candidate values [Strobl et al., 2007]. We refer to Louppe et al. [2013] for more detail on the two first techniques. More recently, Lundberg et al. [2018] proposed new VI measures based on the Shapley effects [Shapley and Shubik, 1954] from game theory which aim to appropriately share the importance of a variable on the prediction.

In this chapter, we restrict our attention to the permutation variable importance (PVI). Some studies empirically showed that the PVI capture individual and interaction effects of the input variables [Lunetta et al., 2004; Strobl et al., 2009; Winham et al., 2012]. Wei et al. [2015] and Gregorutti et al. [2015] provided a theoretical evidence between the PVI values and the Sobol' total indices. These results provided a first connection between SA and VI in data science. However, not many practical tests were established to study the properties of the estimated PVI values, regarding their connection with the Sobol' total indices. Moreover, no relation has been shown with the first-order Sobol' indices. Convergence results of the estimated PVI values have also been shown in Zhu et al. [2015], but Ishwaran and Lu [2018] showed difficulties to obtain consistent confidence intervals for the estimated values. In this chapter, we aim to demonstrate that the PVI values estimated using a random forest model are biased. This bias is directly related to the accuracy of the random forest model and increases with the correlation between input variables.

A necessary condition in the computation of the Sobol' indices is the independence of the input random variables. When input variables are dependent, the variance decomposition cannot be unique. In this case, the influence of a variable can hardly be distinguished from its marginal effect or due to its dependence with another variable. Several studies have been established

to deal with dependencies in SA [Da Veiga, 2015; Chastaing et al., 2012; Kucherenko et al., 2012]. In particular, Mara et al. [2015] used an iso-probabilistic transformation to evaluate the Sobol' indices in an uniform and independent space. In the case of PVI values, they were shown to be biased when dealing with correlated observations [Strobl et al., 2008]. Indeed, when the variables are dependent, the effect of a permutation modifies the correlation between the input variables. To counter this effect, Strobl et al. [2008] proposed a permutation scheme that conditionally permutes the variables, so that the correlations between the input variables remain unchanged. However, this method permutes the data within certain groups and can hardly be applicable for small sample-sizes. This is why in this chapter we propose to use an iso-probabilistic transformation on the input sample to perform the permutation schema in an independent space. Next we apply the inverse transformation to obtain a permuted sample with the same distribution as the initial sample.

This chapter is organized as follows. Section 3.2 introduces the new PVI values for the first-order effects and a new permutation procedure proposed to deal with dependent inputs. Section 3.3 presents numerical experiences to compare the estimation of the Sobol' indices using the permutation procedure and the Monte Carlo sampling. Section 3.4 considers an estimated model from a random forest and reveals a systematic bias in the estimation of the PVI values. We then study the influence of this bias in the estimation of the PVI values with and without dependencies and we also apply the new permutation procedure for dependent variables. Section 3.5 concludes on these contributions and proposes future work.

3.2 New permutation variable importance values and their relations with the Sobol' indices

The Sobol indices and PVI values both aim at measuring the influence of a variable X_j on the model η. However, they both target the output variance as a quantity of interest. In the following section we introduce a new permutation schema which values are related to the first-order Sobol' indices. We also present different schema that consider the Rosenblatt Transformation and are related to the full and independent Sobol' indices.

3.2.1 Relationship with the first-order Sobol' indices

The relationship between the permutation PVI value and the total Sobol' index (see Section 2.4.1.3) shows that a permutation of X_j is equivalent to the study of the partial variance of Y when all variables are fixed except X_j. One could expect that the opposite is also possible: permuting all variables except X_j would be the equivalent to studying the partial variance of Y when only X_j is fixed. Such a relation would corresponds to the definition of the first-order Sobol' indices.

Let $\mathbf{X} = (X_1, \ldots, X_j, \ldots, X_d)$ be a random-vector of independent components, we define $\mathbf{X}_{(-j)} = (X'_1, \ldots, X_j, \ldots, X'_d)$ such as for $k = 1, \ldots, d$ and $k \neq j$, X'_k is an independent copy

of X_k. We now introduce the first-order PVI value (fPVI) of X_j defined by

$$\mathcal{J}_\eta(X_j) = \mathbb{E}\left[\left(\eta(\mathbf{X}) - \eta(\mathbf{X}_{(-j)})\right)^2\right]. \tag{3.1}$$

This value is very close to the original PVI value of X_j. The fPVI value studies the effect on the model output by permuting all variables except X_j. Moreover, Proposition 2 states a relation with the first-order Sobol' index. The proof of Proposition 2 is given in Section 3.6.1 in Appendix.

Proposition 2. *Suppose* $\mathbf{X}$ *is a random-vector of independent components and* $Y = \eta(\mathbf{X})$*, then for any component* X_j*:*

$$S_j = 1 - \frac{\mathcal{J}_\eta(X_j)}{2\mathrm{Var}[Y]} \tag{3.2}$$

This relation shows that the fPVI value is related to the first-order Sobol' index. However, its tendency is the opposite of the first Sobol' index. The lower the fPVI value is, the higher the first-order Sobol' index is. Thus, contrary to the PVI value, the higher the fPVI value of X_j is, the less X_j has a marginal effect on the output model variance.

In variance-based SA, the first-order and total Sobol' indices are two important indices to be computed together in a study. Both indices give important informations about the influence of each variable X_j. Thus, introducing the fPVI value follows this analysis by measuring the marginal effect of a variable and may help to have a better interpretation of the study. However, the relation between the permutation based values and the Sobol' indices (Proposition 1 and 2) are established for independent components. Therefore, we propose in the next section to modify the permutation procedure leading to new PVI values which are related to the full and independent Sobol' indices to be used for dependent variables.

3.2.2 For dependent random variables

The procedure from Mara et al. [2015] to measure the Sobol' indices in the case of dependent inputs (see 2.2.2) used the Rosenblatt transformation (RT) to transform $\mathbf{X}$ into an uniformly and independent random vector. Thus, we propose to apply the RT transformation to dependent data in order to compute PVI values.

We remind that $\mathbf{U}^j$ is the resulting random vector from the RT of $(X_j, \ldots, X_1, \ldots, X_d)$ after reordering the set $(X_1, \ldots, X_d)$ as shown in (2.34). Using the transformed sample $\mathbf{U}^j$ we can consider permuting its elements to compute the PVI values. We then introduce the following

new permutation VI in the case of dependent variables:

$$\mathcal{I}_\eta^{full}(X_j) = \mathbb{E}\left[\left(g_j(\mathbf{U}^j) - g_j(\mathbf{U}^j_{(1)})\right)^2\right] \tag{3.3}$$

$$\mathcal{J}_\eta^{full}(X_j) = \mathbb{E}\left[\left(g_j(\mathbf{U}^j) - g_j(\mathbf{U}^j_{(-1)})\right)^2\right] \tag{3.4}$$

$$\mathcal{I}_\eta^{ind}(X_j) = \mathbb{E}\left[\left(g_{j+1}(\mathbf{U}^j) - g_j(\mathbf{U}^{j+1}_{(d)})\right)^2\right] \tag{3.5}$$

$$\mathcal{J}_\eta^{ind}(X_j) = \mathbb{E}\left[\left(g_{j+1}(\mathbf{U}^j) - g_j(\mathbf{U}^{j+1}_{(-d)})\right)^2\right], \tag{3.6}$$

with the convention that $\mathbf{U}^{d+1} = \mathbf{U}^1$ and $g_{d+1} = g_1$. The random vector $\mathbf{U}^j_{(1)} = (U_1'^j, \dots, U_d^j)$ corresponds to permuting the first element of $\mathbf{U}^j$ while $\mathbf{U}^j_{(-1)} = (U_1^j, U_2'^j, \dots, U_d'^j)$ corresponds to permuting all elements of $\mathbf{U}^j$ except the first. This is the equivalent for $\mathbf{U}^{j+1}_{(d)}$ and $\mathbf{U}^{j+1}_{(-d)}$.

Corollary 1. *Using Proposition 1 and 2, we can state a relation between the PVI values from Equation (3.3)-(3.6) and the already stated full and independent Sobol' indices introduced in Section 2.2.2 of Chapter 2:*

$$\begin{aligned}
\mathcal{I}_\eta^{full}(X_j) &= 2\mathbb{E}[\mathrm{Var}[g_j(\mathbf{U}^j)|U^j_{-1}]] = 2ST_j^{full} \times \mathrm{Var}[g_j(\mathbf{U}^j)] \\
\mathcal{J}_\eta^{full}(X_j) &= 2\mathrm{Var}[\mathbb{E}[g_j(\mathbf{U}^j)|U^j_1]] = 2(1 - S_j^{full}) \times \mathrm{Var}[g_j(\mathbf{U}^j)] \\
\mathcal{I}_\eta^{ind}(X_j) &= 2\mathbb{E}[\mathrm{Var}[g_{j+1}(\mathbf{U}^{j+1})|U^{j+1}_{-d}]] = 2ST_j^{ind} \times \mathrm{Var}[g_{j+1}(\mathbf{U}^{j+1})] \\
\mathcal{J}_\eta^{ind}(X_j) &= 2\mathrm{Var}[\mathbb{E}[g_{j+1}(\mathbf{U}^{j+1})|U^{j+1}_d]] = 2(1 - S_j^{ind}) \times \mathrm{Var}[g_{j+1}(\mathbf{U}^{j+1})].
\end{aligned}$$

And because $Y = \eta(\mathbf{X}) = g_j(\mathbf{U}^j) = g_{j+1}(\mathbf{U}^{j+1})$*, we have*

$$\mathcal{I}_\eta^{full}(X_j) = 2ST_j^{full} \times \mathrm{Var}[Y] \tag{3.7}$$

$$\mathcal{J}_\eta^{full}(X_j) = 2(1 - S_j^{full}) \times \mathrm{Var}[Y] \tag{3.8}$$

$$\mathcal{I}_\eta^{ind}(X_j) = 2ST_j^{ind} \times \mathrm{Var}[Y] \tag{3.9}$$

$$\mathcal{J}_\eta^{ind}(X_j) = 2(1 - S_j^{ind}) \times \mathrm{Var}[Y]. \tag{3.10}$$

The full and independent PVI values are similar to the full and independent Sobol' indices and as such, have a different interpretation than the PVI values without a transformation. The use of a RT when the variables are dependent enables the permutation to be done without changing the distribution of $\mathbf{X}$. For a variable X_j, the full indices shall give the importance of X_j on the output variance with the effect of its dependencies while this is the opposite case for the independent ones.

3.3 Sobol indices estimation: a comparison between the Monte Carlo sampling and the permutation procedure

The previous relations make possible the estimation of the Sobol' indices using the PVI values instead of the classical Monte Carlo sampling (see Sobol [2001]; Saltelli et al. [2010]). Algorithm

1 and 2 summarize respectively how to estimate the Sobol' indices using a permutation scheme for the independent and dependent case. In this section, we propose to compare the estimation accuracy of the Sobol' indices using the Monte Carlo sampling and the permutation procedure. In this comparison, the estimator from Janon et al. [2014] has been used for the Monte Carlo sampling.

3.3.1 Numerical examples

The comparison is established on three examples with different behavior: the Ishigami model, the G-function and an additive Gaussian model. We now detail these examples and their theoretical sensitivity indices.

3.3.1.1 Ishigami model

Introduced in Ishigami and Homma [1990], the Ishigami function is typically used as a benchmarking function for uncertainty and sensitivity analysis. It is interesting because it exhibits a strong non-linearity and has interactions between variables. We consider the random vector $\mathbf{X} = (X_1, X_2, X_3)$, such as $X_j \sim \mathcal{U}(-\pi, \pi)$. The model function can be written as

$$\eta(\mathbf{X}) = \sin(X_1) + a\sin^2(X_2) + bX_3^4\sin(X_1), \tag{3.11}$$

where a and b are constants. The theoretical results of the output variance and partial variances are known (see for example Section 4 of Baudin and Martinez [2014]) and are expressed as

$$\begin{aligned}
&\mathrm{Var}[Y] = \frac{1}{2} + \frac{a^2}{8} + \frac{b^2\pi^8}{18} + \frac{b\pi^4}{5} \\
&V_1 = \frac{1}{2}\left(1 + b\frac{\pi^4}{5}\right)^2, \quad V_2 = \frac{a^2}{8}, \quad V_3 = 0, \\
&V_{-1} = V_1 + b^2\pi^8\frac{8}{225}, \quad V_{-2} = V_2, \quad V_{-3} = b^2\pi^8\frac{8}{225}.
\end{aligned}$$

Many studies considered $a = 7$ and $b = 0.1$ to compute the Sobol' indices [Marrel et al., 2009; Crestaux et al., 2009]. Thus, we have

$$\begin{aligned}
&\mathrm{Var}[Y] = 13.845 \\
&V_1 = 4.346, \quad V_2 = 6.125, \quad V_3 = 0, \\
&V_{-1} = 7.720, \quad V_{-2} = 6.125, \quad V_{-3} = 3.374.
\end{aligned}$$

If we divide these partial variances by the total variance $\mathrm{Var}[Y]$, we obtain the theoretical first-order and total Sobol' indices, which are given by

$$\begin{aligned}
&S_1 = 0.314, \quad S_2 = 0.442, \quad S_3 = 0, \\
&ST_1 = 0.558, \quad ST_2 = 0.442, \quad ST_3 = 0.244.
\end{aligned}$$

However, there are no analytical solutions for the full and independent Sobol' indices. This example is interesting because the model is strongly non linear, X_2 has no interaction with other variables ($V_2 = V_{-2}$) and X_3 is only influential due to its interaction with X_1 ($V_3 = 0$ and $V_{-3} > 0$).

3.3.1.2 G-function

The G-function (or also called Sobol's G-function), introduced in Archer et al. [1997] is an interesting use-case in many sensitivity analysis studies because it can be used to generate test cases over a wide spectrum of difficulties. Let $\mathbf{X} \sim \mathcal{U}(0,1)^d$ be the input random vector and we define the function as

$$\eta(\mathbf{X}) = \prod_{j=1}^{d} g_j, \text{ with } g_j = \frac{|4X_j - 2| + a_j}{1 + a_i}, \tag{3.12}$$

where $a_j \in \mathbb{R}^+$ for all $j = 1, \ldots, d$. The function is driven by the dimensionality d and each coefficient a_j. The theoretical partial variances and output variance are known and are given for each X_j by

$$\begin{aligned} \mathrm{Var}[\mathbb{E}[Y|X_j]] &= \frac{1/3}{(1+a_j)^2} \\ \mathbb{E}[\mathrm{Var}[Y|X_{-j}]] &= V_j \prod_{j \neq k} (1 + V_k) \\ \mathrm{Var}[Y] &= \prod_{j=1}^{d} (1 + V_j) - 1. \end{aligned}$$

The first-order and total Sobol' indices can be obtained by dividing the partial variances with $\mathrm{Var}[Y]$. Low values of a_j implies an important first-order index and high interactions effects. The main interest of this example is to exhibit large interaction effects which are interesting for benchmarking.

3.3.1.3 Additive model with Gaussian framework

Another simple and very common example case is the additive Gaussian framework. The probabilistic model considered $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} = [\mathbb{E}[X_1], \ldots, \mathbb{E}[X_d]]^T$ is the mean vector and $\boldsymbol{\Sigma} = \mathrm{Cov}[[X_j, X_k]; 1 \leq j, k \leq d]$ is the covariance matrix. The additive model can be written as

$$\eta(\mathbf{X}) = \boldsymbol{\beta}^T \mathbf{X} = \sum_{j=1}^{d} \beta_j X_j, \tag{3.13}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)$ is a weight vector of constants. The theoretical results of the output variance and partial variances can easily be obtained since the model is additive:

$$\mathrm{Var}[Y] = \boldsymbol{\beta}^T \Sigma \boldsymbol{\beta}. \tag{3.14}$$

Since the model has no interactions, the Sobol' indices of a random variable X_j at independence are

$$S_j = ST_j = \frac{\beta_j^2 \sigma_j^2}{\mathrm{Var}[Y]}. \tag{3.15}$$

The full and independent indices can be calculated. The full indices add the effect of dependencies in the calculation:

$$S_j^{full} = ST_j^{full} = \frac{\left(\beta_j \sigma_j + \sum_{k \neq j} \beta_k \sigma_k \rho_{j,k}\right)^2}{\mathrm{Var}[Y]}$$

where σ_k is the standard deviation of X_k and $\rho_{j,k} = \mathrm{Cov}[X_j, X_k]/(\sigma_j \sigma_k)$ is the linear correlation between X_j and X_k. The independent indices subtract the effect of indices such as:

$$S_j^{ind} = ST_j^{ind} = \frac{\boldsymbol{\beta}^T \left(\Sigma - \Sigma_{\cdot,-j} \Sigma_{-j}^{-1} \Sigma_{\cdot,-j}^T\right) \boldsymbol{\beta}}{\mathrm{Var}[Y]}$$

where $\Sigma_{\cdot,-j}$ corresponds to Σ without the j-th column and Σ_{-j}^{-1} is the inverse covariance matrix of all variables except X_j.

This example is interesting because we can increase the problem dimension, change the weights β_j of all random variables X_j and have analytical results of the full and independent Sobol' indices.

3.3.2 Numerical experiments

For the three numerical examples, we compare the estimation error rates in function of the number of model evaluations for the classical Monte Carlo (MC) sampling and the permutation procedure. The experiments are done with an increasing number of model evaluations and we also estimate the confidence interval from a bootstrap sampling for both methods.

3.3.2.1 Computational costs

In the independent case, the estimation of the first and total indices for the MC procedure requires $d + 2$ model evaluations which is due to the evaluation of two matrices A and B and d other matrices created from the crossing of A with B (see Saltelli et al. [2010]). For the permutation schema, $2d+1$ model evaluations are necessary due to the $2d$ permutations of the original sample and one from the evaluation of the original sample.

In the dependent case, the schema changes slightly and the Rosenblatt transformation procedure increases the number of evaluations for the MC case, which brings the number of evaluations to $4d$. Two matrices A and B are generated but both are transformed differently for each variable X_j, which require $2d$ evaluations. Then, d evaluations are necessary for the crossing of A and B for the computation of the independent indices and d others for a different crossing of A and B for the full indices (see Mara et al. [2015] for more details). The permutation procedure

needs $4d + 1$ model evaluations. One for the original sample and $4d$ for the four permutations (and transformations) for each variable X_j. The evaluation cost is summarized in the following tabular:

	Independent	Dependent
Classical MC	$d + 2$	$4d$
Permutation	$2d + 1$	$4d + 1$

At independence, the permutation procedure is approximately two times costlier than the classical MC while it is almost equivalent for the dependent case. The following experiments compare both approaches for the same computational budget, and compares the decrease of the estimation error with the number of model evaluations. These experiments are done for an increasing number of model evaluations, from 10^2 to 10^5. Confidence intervals are also estimated using 1000 bootstrap samples for the MC and permutation methods. In order to correctly estimate the excepted error for a given computational budget, each experiment is done 500 times with new samples. Moreover, we also estimate the probability of coverage (POC) which measures, for each experiment, the probability that the confidence interval at 80% of each estimated indice covers the true indice value.

3.3.2.2 Ishigami at independence

We first consider the typical Ishigami example with independent variables. The experiment is largely known in sensitivity analysis literature because it exhibits interactions between variables and has theoretical results of the first and total indices. Figure 3.1 shows, for the first and total Sobol indices, the absolute error of the estimated indices (with full lines) and the POC (with dotted lines) in function of the number of model evaluations (in log scale). The scale of the absolute error is represented at left and the scale of the POC is represented at right. The blue and the red lines respectively correspond to the results for the Monte Carlo schema and the Permutation method. The dark line represents the true probability of the confidence interval fixed at 80 % for these experiments.

We observe that the POC for the first and total indices converge rapidly and closely to the true probability (80%) for all variables. The absolute error of all indices also converges to zero for the first and total indices, and the MC and permutation schemes. However, the rates differ for both methods. For the first indices, the MC estimation has a lower error compared to the permutation technique and this is the opposite for the Total indices.

3.3.2.3 G-function at independence

We now consider the G-function for $d = 10$ and coefficients of increasing weights $a_j = j - 1$ for $j = 1, \ldots, d$. Figure 3.2 represents the same quantities as Figure 3.1.

The absolute error for both the first and total indices seems to decrease with the same exponential rate as the Ishigami example. The error of the permutation procedure is greater

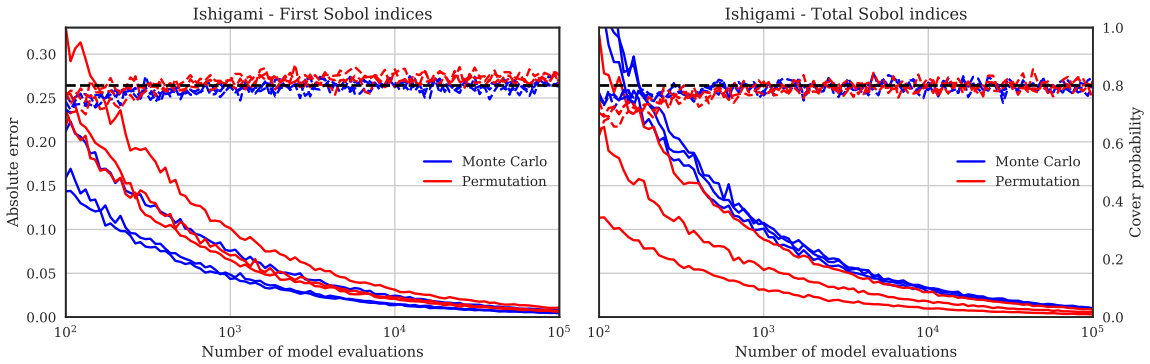


Figure 3.1 *Convergence of the indice estimations and their confidence intervals for the MC and permutation schemes for the Ishigami example.*

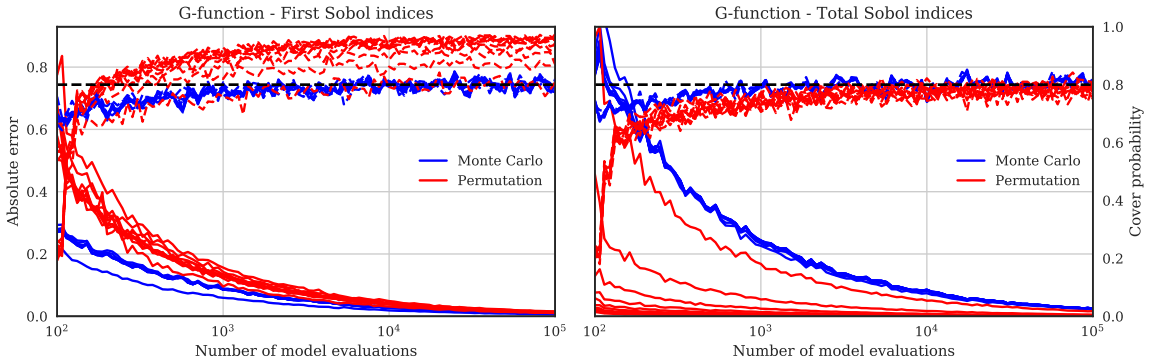


Figure 3.2 *Convergence of the indices estimations and their confidence intervals for the MC and permutation schemes for the G-function example ($d = 10$).*

than the MC sampling for the estimation of the first indices, while it is the opposite for the total indices. The POC converges towards the correct probability for the MC sampling, but not for the permutation procedure, which is slightly overestimated for the first indices.

3.3.2.4 Additive Gaussian at independence

We first consider the independent additive Gaussian framework for $d = 10$ with increasing weights for the variables ($\beta_j = j$ for $j = 1, \ldots, d$). We also consider distribution of unit variance and zero mean ($\mu_j = 0$, $\sigma_j = 1$ for $j = 1 \ldots, d$). Since the model is additive and has no interactions, the first and total indices are equivalent. The true values are given by Equation 3.15 and where $S_j \propto \beta_j^2$. The Figure 3.3 represents the same quantities as Figure 3.1.

The POC of the MC schema converges rapidly to the true probability, however it is not the case for the permutation schema, especially for the first Sobol indices. For the permutation

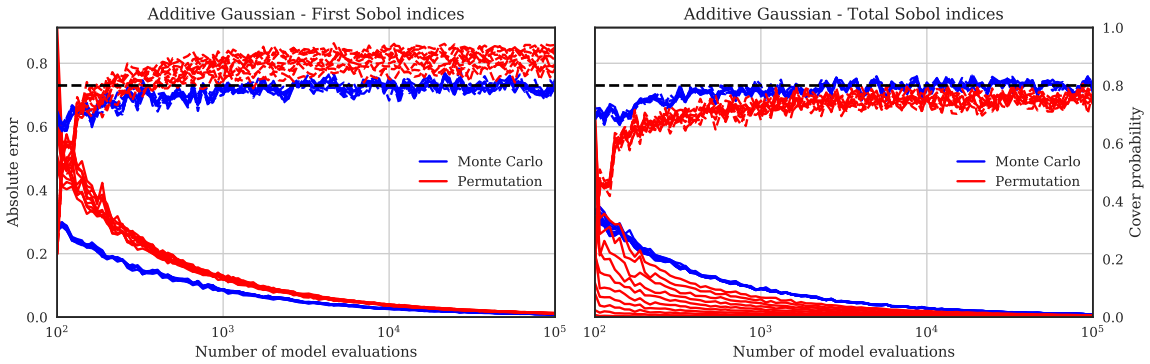


Figure 3.3 *Convergence of the indice estimations and their confidence intervals for the MC and permutation schemes for the additive Gaussian example ($d = 10$).*

schema, the confidence intervals are overestimated for the first indices and underestimated for the total.

We observe that the absolute errors of the first and total indices are equivalent for the MC schema and for all indices. However, this is not the case for the permutation technique where the error decreases much slower for the first indices than for the total ones. Moreover, the estimation error of the total indices with the permutation technique is larger for large indices, and inversely.

We now consider the same example, but with an increasing dimension in order to compare the influence of d on the absolute error and the POC for a given sample size of $n = 10^4$. Figure 3.4 shows at the top the absolute error for both the MC sampling and permutation procedure for the first and total indices. The bottom represents the POC, where each point represents the POC of the indice of a variable X_j.

For the first indices, the absolute error increases slowly with the dimension for both methods. This seems equivalent for the absolute error of the MC sampling for the total indice, but not for the permutation procedure, which decreases slowly with the dimension. The POC of the MC sampling is close very close to the true probability for both the first and total indices. Whereas the permutation procedure still overestimates the POC for the first indices and underestimates for the total indices.

3.3.2.5 Additive Gaussian for dependent inputs

In this example we consider a dependent additive Gaussian framework for $d = 3$ and equal weights for all variables ($\beta_j = 1$ for $j = 1, \ldots, d$). We also consider distribution of unit variance and zero mean ($\mu_j = 0$, $\sigma_j = 1$ for $j = 1 \ldots, d$). This three-dimensional problem has three pairs of variables with linear correlations described by $\boldsymbol{\rho} = (\rho_{12}, \rho_{13}, \rho_{23}) = (0, 0.2, 0.9)$. The true indices are given in the table 3.1. We observe that $S_1^{full} < S_3^{full} < S_3^{full}$, because the full indices consider the effect of dependencies. This explains why X_3 has a larger indice value due

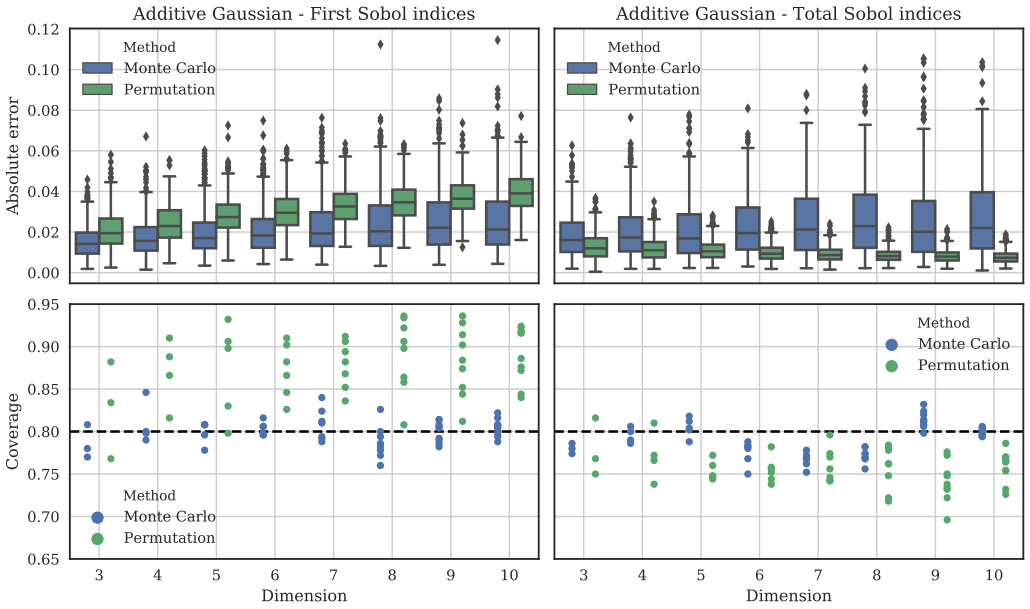


Figure 3.4 *Convergence of the indice estimations and their confidence intervals for the MC and permutation schemes for the additive Gaussian example ($d = 10$).*

to its large correlations with X_1 and X_2. On the other hand, X_1 has a larger independent Sobol indice, because the effect of dependencies are subtracted. Figure 3.5 shows the same quantities as above for the full and total respectively in the first and second row.

	X_1	X_2	X_3
Full	0.277	0.694	0.848
Independent	0.152	0.030	0.029

Table 3.1 *Theoretical values of the full and independent Sobol indices for the additive Gaussian model with $\boldsymbol{\rho} = (0, 0.2, 0.9)$*

The POC for the MC schema seems to correctly converge to the true probability for all variables and all types of indices. However, it is not the case for the permutation schema. As for the results for the independent case, there is a bias in the convergence of the estimated confidence intervals for the permutation case. The Monte Carlo procedure seems to have a better accuracy in the estimation of the indices than the permutation method, except for the total independence indices, where the permutation method is significantly better.

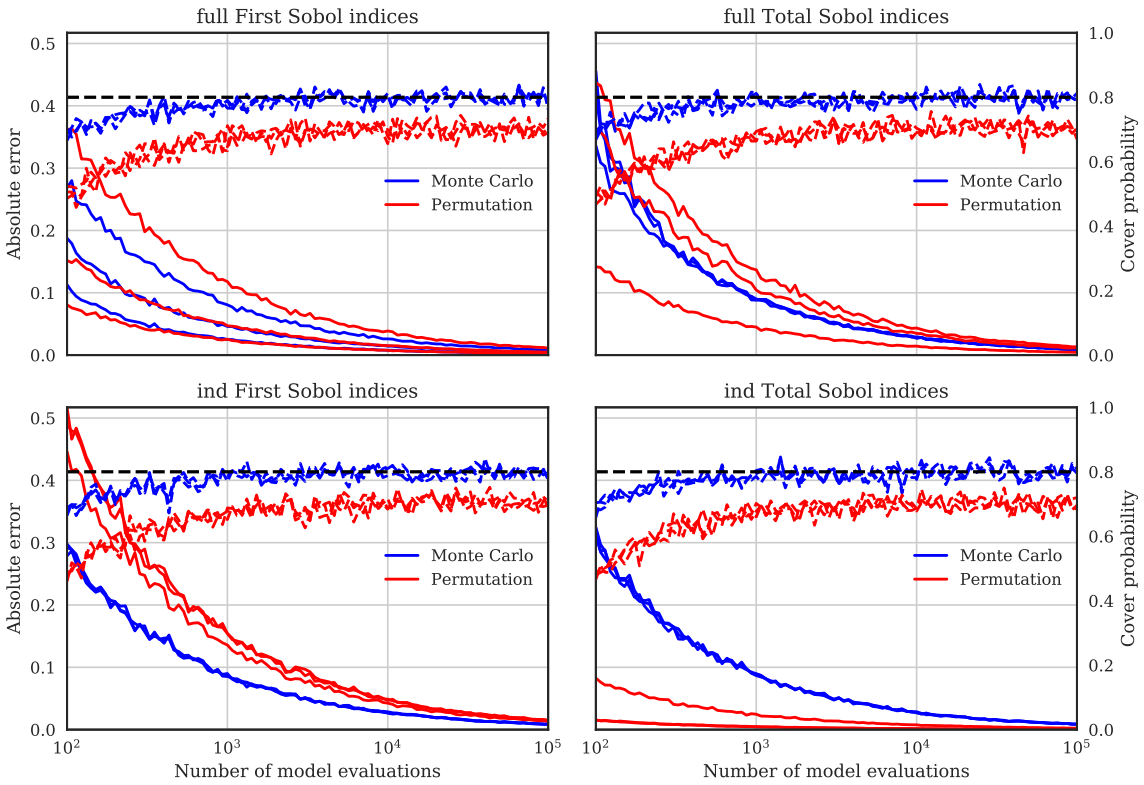


Figure 3.5 *Convergence of the indices estimations and their confidence intervals for the MC and permutation schemes for the additive Gaussian example for* $d = 3$ *and* $\boldsymbol{\rho} = [0., 0.2, 0.9]$.

3.3.2.6 Conclusion

These experiments show the difference in the estimation of the Sobol' indices using the classical MC sampling and the permutation technique on the true model η. The MC sampling has been shown in various studies to be the classical method for the estimation of the Sobol' indices with great statistical properties and confidence intervals (with bootstrap or asymptotic). However, little is known about the permutation based indices. In our experiments, we observe that the confidence intervals with bootstrap are not always consistent and tend to be slightly overestimated or underestimated. The estimation error seems larger for the first Sobol' indices compared to the MC schema, and inversely for the total indices. The estimation of the Sobol' indice with the permutation schema is a good alternative to the classical MC procedure, especially for the computation of the total indices. However, supplementary studies should be established to understand the bias in bootstrap estimations of the confidence interval. Ishwaran and Lu [2018] took interest in the consistency of the bootstrap confidence intervals of PVI values when using random-forest and observed important difficulties in their estimations.

3.3.3 Effect of the permutation on dependent variables

The aim of a permutation is to break the link between Y and the permuted variable X_j. This makes the permuted variable X'_j independent to Y, but has the side effect of also breaking the link with the other input variables. It can be easily checked that a permuted variable X'_j is uncorrelated with the other components of $\mathbf{X}$ (see Section 3.6.2 in Appendix). Thus, if X_j is correlated to other variables, its permutation shall make the distribution of the permuted sample $\mathbf{X}_{(j)}$ different from $\mathbf{X}$. Comparing the responses $\eta(\mathbf{X})$ and $\eta(\mathbf{X}_{(j)})$ does not really make sense since $\mathbf{X}$ and $\mathbf{X}_{(j)}$ have different distributions and probably do not have the same variances. Therefore, a good comparison should be with a permuted sample with the same distribution as $\mathbf{X}$. The interest in using a Rosenblatt Transformation (RT) is one can obtain a permuted sample without changing the probability distribution of $\mathbf{X}$ (see Algorithm 2 in Appendix for details on the procedure).

To illustrate the influence of the permutation on dependent variables, we consider the Ishigami example (see Section 3.3.1.1. We apply a correlation of $\boldsymbol{\rho} = (\rho_{12}, \rho_{13}, \rho_{23}) = (0, 0.9, 0)$. Figure 3.6 shows the effect of permuting X_1. The first row represents scatter plots of the input samples where the left figure is for the original sample, the middle figure is for the permuted sample and the right figure is for the permuted sample using a RT. We denote as X'_1 the permutation of X_1 and $X_1'^{RT}$ the permutation of X_1 using a RT. The second row represents the estimated model output distributions for each input samples of the first row.

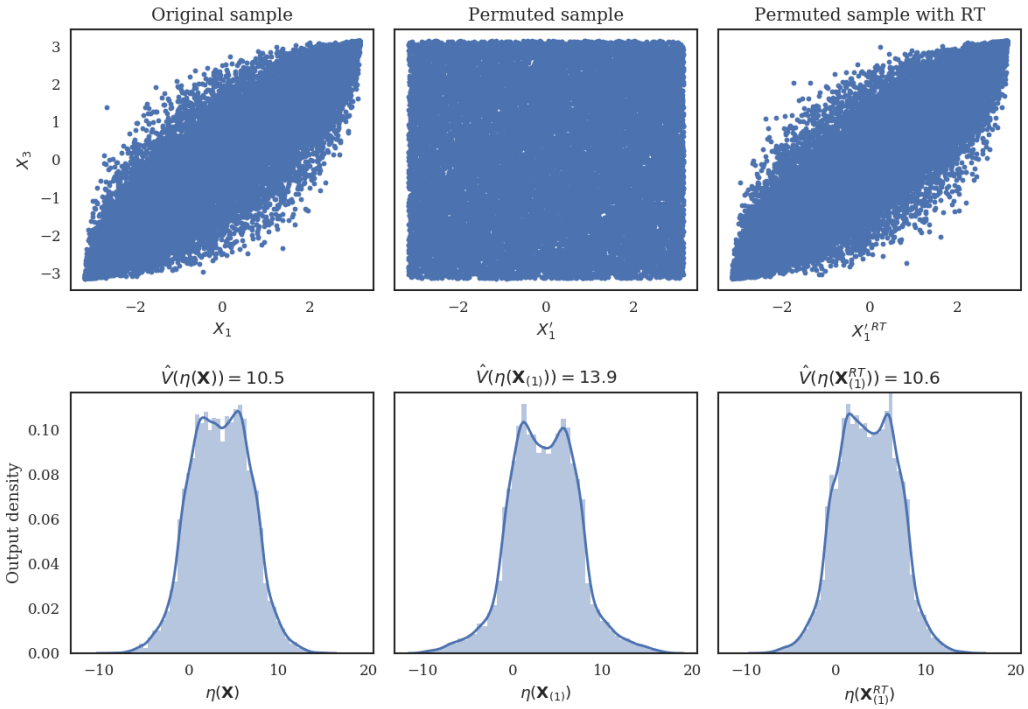


Figure 3.6 *Effect of permuting X_1 on the Ishigami problem with a correlation $\rho_{13} = 0.9$.*

In the first row, we observe that the permuted sample no longer has correlation between X_1' and X_3, while the permuted sample using a RT still holds the initial correlation. This major change is observed in the second row which shows different distributions between the response of the original sample and the permuted one. The variance of $\eta(\mathbf{X}_{(1)})$ corresponds to the model when $\rho_{13} = 0$. Whereas the distribution of $\eta(\mathbf{X}_{(1)}^{RT})$ is the same as the original one.

Having a different distribution of $\eta(\mathbf{X}_{(1)})$ with $\eta(\mathbf{X})$ can lead to a mis-interpretation of the PVI values. As defined in (2.63), when comparing $\eta(\mathbf{X}_{(1)})$ with $\eta(\mathbf{X})$, the main increase of the PVI value is due to the difference on their respective variances. On the other hand, when using the RT, the comparison is more admissible, and the influence of the permutation is not absorbed by the difference on the output variances. The new full and independent PVI values, introduced in Section 3.2.2 are obtained from the Rosenblatt Transformation, and they give a different interpretation of the influence of variables in the presence of dependencies.

To show the different interpretations of these PVI values when the variables are dependent, we compare using the Ishigami example the results with the PVI values from (2.63), the full PVI values from (3.3) and the independent PVI values from (3.5) in function of the correlation ρ_{13}. The results are shown in Figure 3.7. We observe that at independence ($\rho_{13} = 0$), the results of each variable are equivalent for the all the PVI values. Since there is only a correlation between X_1 and X_3, and X_2 does not have interactions with other variables, it is logical that the importance of X_2 remains constant with ρ_{13} for all PVI values. We also observe that the correlation has more impact on the variable importances for the full and independent variables. Moreover, when reaching perfect dependencies ($\rho_{13} = 1, -1$), the importance of X_1 and X_3 is equal when reaching perfect dependencies for the full and independent PVI values because $X_1 = X_3$. The indice full PVI value $\mathcal{I}_\eta^{full}(X_3)$ has a non-monotonic behavior with the correlation, with a slight increase for $-0.5 < \rho_{13} < 0.5$ before decreasing. This behavior is probably due to the non-linearity and the interaction between X_1 and X_3 in the Ishigami function.

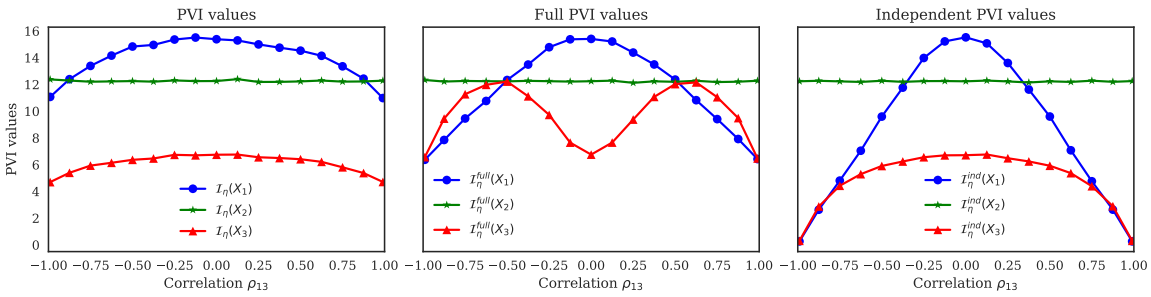


Figure 3.7 *Comparison between the PVI values, the full PVI values and the independent PVI values in function of the correlation ρ_{13} for the Ishigami example.*

3.4 Using Random Forest

In previous Sections, we considered the true model η in the computation of the PVI and fPVI values and showed good experimental convergence of the estimations However, η is not available in practice. As stated in Section 2.4.1.2, the PVI values are usually calculated from the predictors trees in a random forest model. In this section we consider the estimated model from a random forest and we show that the estimations of the PVI values can significantly differ. The following results only consider the PVI values, but the procedure is equivalent for the fPVI values.

3.4.1 Bias identification in PVI values

For a given sample $\mathcal{D}_n$, we first consider the estimator of an infinite random forest $\hat{\eta}_\infty$ as defined in (2.58) to compute the PVI values such as

$$\mathcal{I}_{\hat{\eta}_\infty}(X_j) = \mathbb{E}_\Theta\Big[\mathbb{E}\Big[\Big(Y - \hat{\eta}(\mathbf{X}_{(j)}; \Theta, \mathcal{D}_n)\Big)^2\Big]\Big] - \mathbb{E}_\Theta\Big[\mathbb{E}\Big[(Y - \hat{\eta}(\mathbf{X}; \Theta, \mathcal{D}_n))^2\Big]\Big]. \tag{3.16}$$

In the following, we now write $\hat{\eta}(\cdot; \Theta) = \hat{\eta}(\cdot; \Theta, \mathcal{D}_n)$ for clarity purposes. The first term of (3.16) can be developed as follow:

$$\begin{aligned}
&\mathbb{E}_\Theta\Big[\mathbb{E}\Big[\Big(Y - \hat{\eta}(\mathbf{X}_{(j)}; \Theta)\Big)^2\Big]\Big] = \mathbb{E}_\Theta\Big[\mathbb{E}\Big[\Big(Y - \eta(\mathbf{X}_{(j)}) - (\hat{\eta}(\mathbf{X}_{(j)}; \Theta) - \eta(\mathbf{X}_{(j)}))\Big)^2\Big]\Big] \\
&= \mathbb{E}_\Theta\Big[\mathbb{E}\Big[\Big(Y - \eta(\mathbf{X}_{(j)})\Big)^2\Big] + \mathbb{E}\Big[\Big(\eta(\mathbf{X}_{(j)}) - \hat{\eta}(\mathbf{X}_{(j)}; \Theta)\Big)^2\Big] - 2\mathbb{E}\Big[\Big(Y - \eta(\mathbf{X}_{(j)})\Big)\Big(\hat{\eta}(\mathbf{X}_{(j)}; \Theta) - \eta(\mathbf{X}_{(j)})\Big)\Big]\Big] \\
&= \mathcal{I}_\eta(X_j) + \mathbb{E}_\Theta\Big[\mathbb{E}\Big[\Big(\eta(\mathbf{X}_{(j)}) - \hat{\eta}(\mathbf{X}_{(j)}; \Theta)\Big)^2\Big]\Big] + 2\mathbb{E}_\Theta\Big[\mathbb{E}\Big[\Big(\eta(\mathbf{X}_{(j)}) - \hat{\eta}(\mathbf{X}_{(j)}; \Theta)\Big)\Big(Y - \eta(\mathbf{X}_{(j)})\Big)\Big]\Big].
\end{aligned} \tag{3.17}$$

Which make the PVI values using the true model equivalent out to:

$$\begin{aligned}
\mathcal{I}_\eta(X_j) = \mathbb{E}_\Theta\Big[\mathbb{E}\Big[\Big(Y - \hat{\eta}(\mathbf{X}_{(j)}; \Theta)\Big)^2\Big]\Big] - \mathbb{E}_\Theta\Big[\mathbb{E}\Big[\Big(\eta(\mathbf{X}_{(j)}) - \hat{\eta}(\mathbf{X}_{(j)}; \Theta)\Big)^2\Big]\Big] \\
- 2\mathbb{E}_\Theta\Big[\mathbb{E}\Big[\Big(\eta(\mathbf{X}_{(j)}) - \hat{\eta}(\mathbf{X}_{(j)}; \Theta)\Big)\Big(Y - \eta(\mathbf{X}_{(j)})\Big)\Big]\Big].
\end{aligned} \tag{3.18}$$

This interesting results shows that the use of an estimated model adds two supplementary terms. Thus, by merging (3.17) with (3.16), we find that

$$\begin{aligned}
\mathcal{I}_{\hat{\eta}_\infty}(X_j) = \mathcal{I}_\eta(X_j) + \mathbb{E}_\Theta\Big[\mathbb{E}\Big[\big(\eta(\mathbf{X}_{(j)}) - \hat{\eta}(\mathbf{X}_{(j)}; \Theta)\big)^2\Big]\Big] - \mathbb{E}_\Theta\Big[\mathbb{E}\Big[\big(Y - \hat{\eta}(\mathbf{X}; \Theta)\big)^2\Big]\Big] \\
+ 2\mathbb{E}_\Theta\Big[\mathbb{E}\Big[\big(\eta(\mathbf{X}_{(j)}) - \hat{\eta}(\mathbf{X}_{(j)}; \Theta)\big)\big(Y - \eta(\mathbf{X}_{(j)}; \Theta)\big)\Big]\Big].
\end{aligned} \tag{3.19}$$

The supplementary terms are all related to the accuracy of the random-forest. The first and the second additional terms correspond to the mean square errors of the predictors $\hat{\eta}$ of all Θ respectively on the permuted sample $\mathbf{X}_{(j)}$ and the original sample $\mathbf{X}$. The third one is an interaction effect between the accuracy of all predictors Θ on $\mathbf{X}_{(j)}$ and the deviation between

$Y - \eta(\mathbf{X}_{(j)}; \Theta)$, which is linked to the importance of X_j. Note that $\mathbb{E}_\Theta\Big[\mathbb{E}\Big[\big(Y - \hat{\eta}(\mathbf{X}; \Theta)\big)^2\Big]\Big]$ directly comes from (3.16) and can be computed.

The first and second additional terms were identified by Zhu et al. [2015] in the additional material. However, the third term was ignored as it has a non-negligible importance. These terms are generally non null and they create a bias on the PVI values evaluations. Moreover, this bias cannot be evaluated without the true numerical model η, making it very difficult to estimate. However, one can eventually estimate a part of the terms using the variance estimation from Wager et al. [2014].

In the case of a finite random forest estimator $\hat{\eta}_M$ of M trees, Equation (3.19) can be written as

$$\mathcal{I}_{\hat{\eta}_M}(X_j) = \mathcal{I}_\eta(X_j) + \frac{1}{M}\sum_{m=1}^{M}\Big[\mathbb{E}\Big[\big(\eta(\mathbf{X}_{(j)}) - \hat{\eta}(\mathbf{X}_{(j)}; \Theta_m)\big)^2\Big] + \mathbb{E}\Big[\big(Y - \hat{\eta}(\mathbf{X}; \Theta_m)\big)^2\Big] - 2\mathbb{E}\Big[\big(\eta(\mathbf{X}_{(j)}) - \hat{\eta}(\mathbf{X}_{(j)}; \Theta_m)\big)\big(Y - \eta(\mathbf{X}_{(j)}; \Theta_m)\big)\Big]\Big]. \tag{3.20}$$

We now study the estimation accuracy of the PVI values when evaluated with a random forest model using some of the examples introduced in Section 3.3.1. In the following the random forest hyper-parameters are not tuned and default values are considered, such as $mtry = d$, a number of trees of $M = 500$ with no maximum number of leaves. We also evaluate the PVI values using $p = 20$ number of permutations in order to properly catch the uncertainty from the permutations.

3.4.1.1 Bias in the estimation

Let consider the Ishigami example introduced in Section 3.3.1.1 and fit a random forest model $\hat{\eta}_M$ of M trees on a train sample $\mathcal{D}_n$. To illustrate the existence of a systematic bias, we estimate the fPVI and PVI values using 500 random forests built from unique train samples $\mathcal{D}_n$ and compare the estimation with the theoretical values.

Figure 3.8 represents the box-plot of the estimated fPVI and PVI values for all variables with their true values in black dot points. We observe that the mean of the boxplots are not centered on the theoretical values. This deviation can be significant for some variables and more importantly, it can alter the ranking such as, for the fPVI values, $\hat{\mathcal{J}}_{\hat{\eta}}(X_1) < \hat{\mathcal{J}}_{\hat{\eta}}(X_2)$ instead of $\mathcal{J}_\eta(X_1) > \mathcal{J}_\eta(X_2)$. The ranking is not changed for the PVI values, but we observe that $\hat{\mathcal{I}}_{\hat{\eta}}(X_2)$ is largely underestimated.

This experiment confirms the existence of the bias identified in the previous section and shows that it can alter the interpretation of the values, which can be very important in some studies.

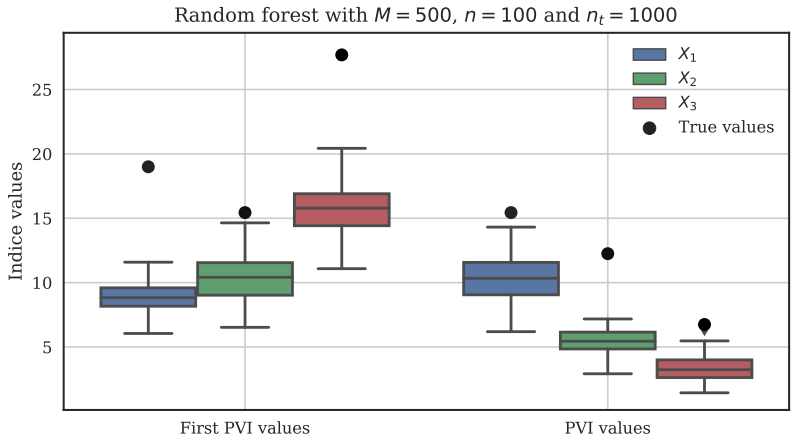


Figure 3.8 *Estimation of the PVI and fPVI values when using a random forest model $\hat{\eta}_M$ with $M = 500$ and a trained sample of $n = 100$ and a test-sample size $n_t = 1000$.*

3.4.1.2 Convergence with the test sample size

We now consider the same example as in the previous section. As stated in (2.61), the PVI values can be estimated using the OOB error of a tree or an independent test sample. In the following, we consider an independent test sample for the estimation. The test sample size is increased and we compute for each variable X_j the residual error of the estimated PVI value, defined by $r_{\hat{\eta}_M}(X_j) = \hat{\mathcal{I}}_{\hat{\eta}_M}(X_j) - \mathcal{I}_\eta(X_j)$. Figure 3.9 shows the residual error in function of the test sample size for the PVI (bottom) and fPVI (top) values. It also compares the estimation error when using the true function η and using a random forest model $\hat{\eta}_M$ with $M = 500$ trees and a trained sample size of $n = 100$. These experiments are done overs 500 replications and the mean of $r_{\hat{\eta}_M}(X_j)$ over all replicas is shown in plain line, while the dotted lines represent the 95 % quantiles of the replications.

When the true model is considered, we observe that the estimation is unbiased with a residual error close to zero for both the fPVI and PVI values. The quantiles of the errors decrease with the sample size. When using the random forest model $\hat{\eta}_M$, we observe that $r_{\hat{\eta}}$ is not centered on zero for any of the indices. This deviation corresponds to bias introduced in (3.20). This bias is directly linked to the accuracy of the random forest. Thus, it cannot be reduced with the increase of the sample size. We observe a ranking of the error terms, such as for the PVI values $r_{\hat{\eta}}(X_3) < r_{\hat{\eta}}(X_1) < r_{\hat{\eta}}(X_2)$. This can be explained by the term $(Y - \hat{\eta}(\mathbf{X}_{(j)}; \Theta_m))$ in Equation (3.20) which is proportional to the importance of X_j on the Y. However, it is the opposite for the fPVI values, since $(Y - \hat{\eta}(\mathbf{X}_{(-j)}; \Theta_m))$ is lower for high influential variables (because the low fPVI value corresponds to an influential variable and inversely).

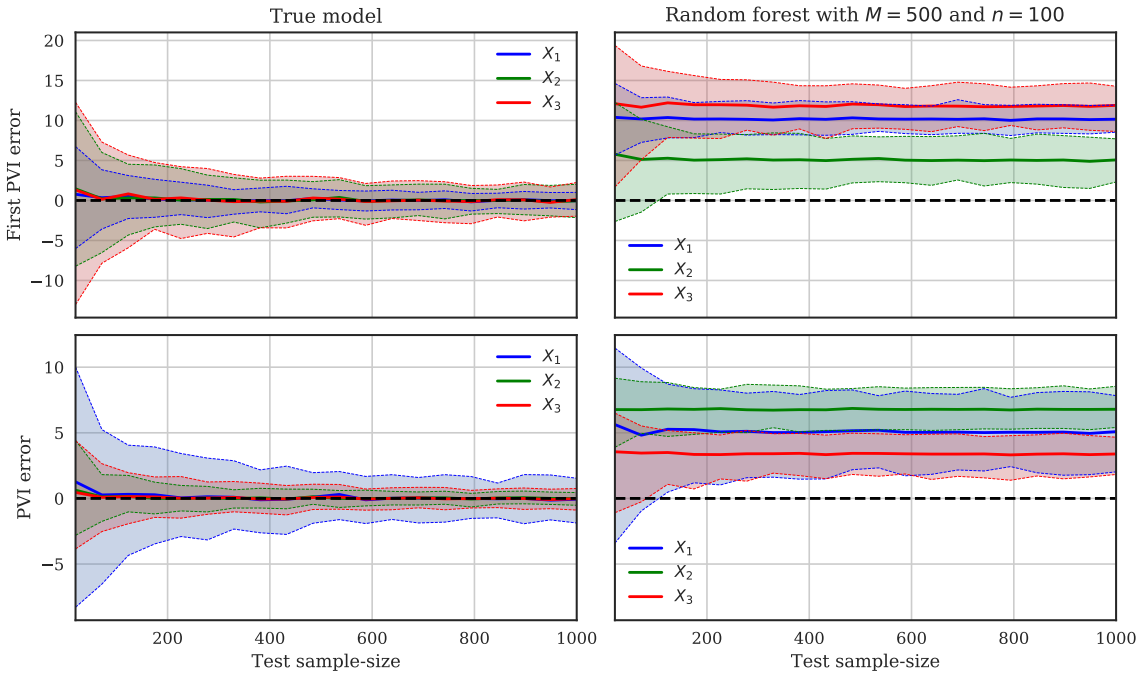


Figure 3.9 *Estimation error in function of the test sample-size for the PVI and fPVI values when using the true model η and a random forest model $\hat{\eta}_M$ with $M = 500$ and a trained sample of $n = 100$.*

3.4.1.3 Convergence with the train sample-size

Still using the Ishigami example, we now aim to numerically show how the accuracy of the random forest influences the bias of the PVI values. To measure this accuracy, we compute the Nash–Sutcliffe model efficiency coefficient (also called Q^2) for multiple train sample $\mathcal{D}_n$. The Q^2 score is defined by

$$Q^2(\hat{\eta}, \mathcal{D}_n) = 1 - \frac{\sum_{(\mathbf{x},y)\in\mathcal{D}_n}(\hat{\eta}(\mathbf{x}) - y)^2}{\sum_{(\mathbf{x},y)\in\mathcal{D}_n}(\hat{\eta}(\mathbf{x}) - \bar{y})^2}, \quad \text{where} \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i,$$

and is equivalent to the coefficient of regression for the case of regression procedures. The greater the Q^2 and the better the model accuracy. Figure 3.10 represents the Q^2 of random forest models in function of the train sample-size n, established over 100 unique train samples $\mathcal{D}_n$ for each sample size n. We observe that the accuracy slowly increases with the sample-size and almost reaches 1 for $n = 10^4$ observations. However, the accuracy could be increased if the hyper-parameters of the random forest are tuned.

Using the same random forest models, we estimate the PVI and fPVI values using the OOB samples of each tree. We then compute the residual error of each estimated value. The Figure 3.11 shows in plain lines the estimation error of the PVI and fPVI values in function of the train sample-size with the 95 % quantiles from the 100 replications. The error decreases with the sample size for both the PVI and fPVI and for all variables. This confirms the results from

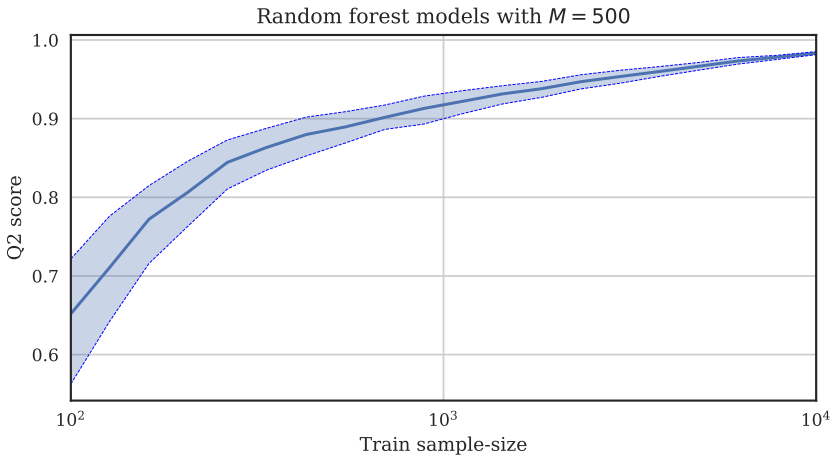


Figure 3.10 *Q^2 accuracy of random forest models for the Ishigami model in function of the training sample-size. 100 samples $\mathcal{D}_n$ are randomly sampled for each sample-size n.*

Equation (3.20) which showed the relation between the bias and the accuracy of the random forest model.

3.4.2 The effect of dependencies

3.4.2.1 Interpretation using an estimated model

The effect of dependencies on the interpretation of PVI values has been studied in Strobl et al. [2008] and it was observed that dependencies make the interpretation of the PVI values difficult by differentiating the marginal effects and the conditional effects. The bias observed by Strobl et al. [2008] can be explained by Equation (3.20). As explained above, a bad accuracy of a random forest model influence can significantly bias the estimation of the PVI values. However, when there are high dependencies among the variables, the original procedure of permutation increases the bias.

This can be illustrated with Figure 3.12 which considers an additive Gaussian example with $d = 2$ and a correlation of $\rho_{12} = -0.9$ between X_1 and X_2. The figure shows a train sample $\mathcal{D}_n$ (small black dots) and the permuted sample $\mathbf{X}_{(1)}$ (big dots). The gradient of colors within the dots of $\mathbf{X}_{(1)}$ represents their mean squared errors averaged over the OOB samples of each tree. As stated in Section 3.3.3, the permutation of X_1 breaks the link between X_1 and X_2, making X_1' uncorrelated to X_2, which explains why X_1' and X_2 are uncorrelated. This artifact, forces the estimated model to evaluate points that are outside the space of $\mathcal{D}_n$. This area is unknown to the random forest and the model accuracy can be extremely low. As shown in the figure, the points on the top right and bottom left (outside the space of $\mathcal{D}_n$) have larger errors (up to ≈ 25). The estimation of the PVI is even more biased because these estimation errors increase the bias from the terms of (3.19). Thus, the more a variable X_j is dependent to its other variables, the

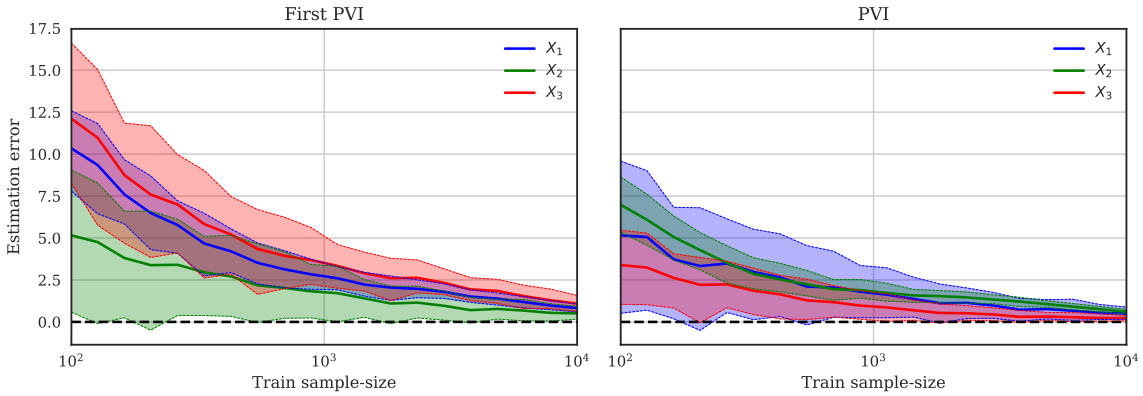


Figure 3.11 *Variation of the estimation error with the sample-size for the fPVI (left) and PVI (right) values estimated from the OOB samples over $M = 500$ trees, $p = 20$ permutations and 100 different training samples of the Ishigami model.*

more its PVI value is biased.

3.4.2.2 Strong relation between the bias and the correlations

Dependencies can have a significant impact on the bias of the PVI values. The use of a Rosenblatt transformation can counterpart this difficulty and make the permuted samples follow the same distribution as the original sample. To illustrate the influence of dependencies, we consider the same experiment as in Section 3.3.3, but instead of using the true model η to estimate the PVI values, we consider a random forest model built from a large train sample size of $n = 10000$. Figure 3.13 represents at the top the PVI, full PVI and independent PVI values in function of the correlation ρ_{13}. The bottom figures shows the estimation bias computed from Equation by (3.20). If we compare the estimated values from the top figures with the results of Figure 3.7, we observe than the estimated full and independent PVI values using a random forest are very close with the estimation using the true model. However, it is not the case for the PVI values (computed with the original procedure), which have very different values for extreme correlations. We also observe in the figure below that the bias for the PVI values increases significantly for extreme correlations. Note that the sum of the value and the bias would lead to the same results as in Figure 3.7. The bias is still non null for the full and independent PVI values due to the accuracy of the random forest.

This experiment confirms that the procedures from the full and independent PVI values have more coherent results when using a random forest model and that the bias is not increased for large correlations.

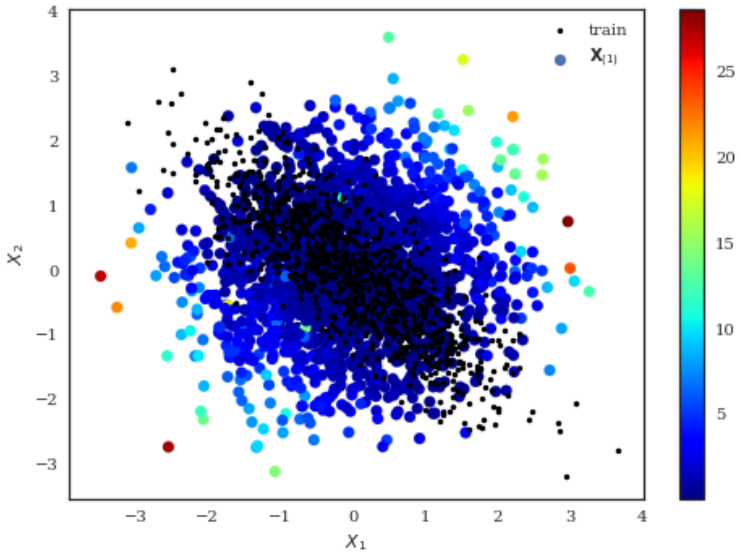


Figure 3.12 *Prediction error of a random forest on the permuted samples of correlated variables.*

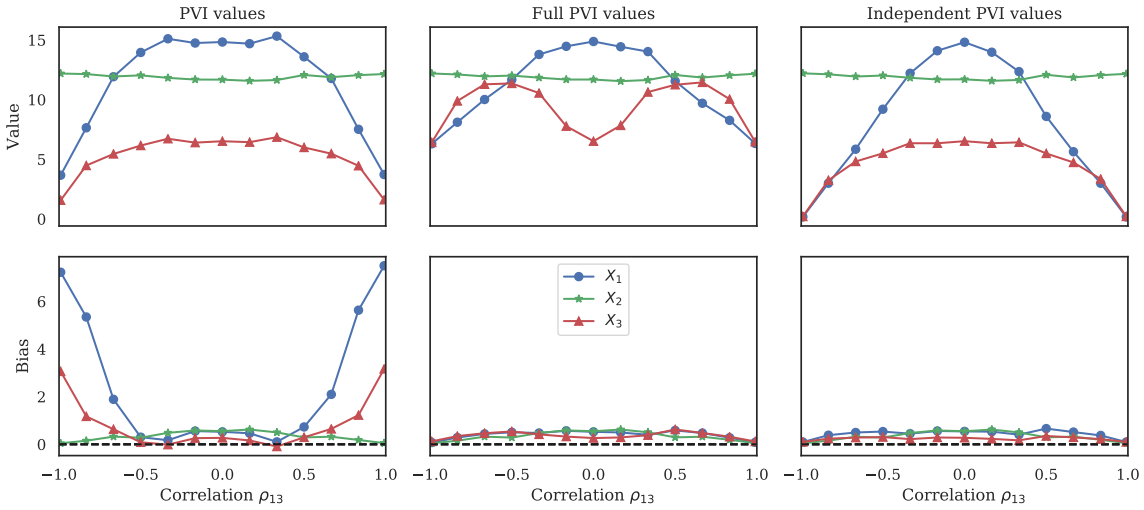


Figure 3.13 *Estimation of the PVI values, full PVI values and independent PVI values in function of the correlation between X_1 and X_3 for the Ishigami example using a finite random forest $\hat{\eta}_M$ with $M = 100$ and a fixed training sample size of size $n = 10000$.*

3.4.2.3 Application of the new PVI values

In order to validate the utility of the RT procedure, we consider the same example as in Strobl et al. [2008] but without the random noise. This example is a $d = 12$ additive Gaussian model with large correlations between some variables. The covariance structure Σ is chosen such that all variables have unit variance $\sigma_{jj} = 1$ and only the first four components are block-correlated with $\sigma_{jk} = 0.9$ for $j \neq k \leq 4$, while the rest are independent $\sigma_{jk} = 0$. From the twelve variables, only six are influential with a non-null weight β_j and are given in Table 3.2.

Variables	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	...	X_{12}
Coefficients	5	5	2	0	-5	-5	-2	0	...	0

Table 3.2 *Coefficient values $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{12})$ of the Gaussian additive model.*

Since the model is additive and has no interactions, we do not consider the fPVI values because they are equivalent to the PVI values for the full and independent Sobol' indices. The boxplots of the estimations (over the 500 replications) are shown in Figure 3.14 for all values. The full and independent PVI values obtain very different results than the classical PVI values except for the non-influential and independent variables X_8 to X_{12} which are measured as non-influential for all values. As explained in the example from Figure 3.13, the classical PVI value is subject to a large bias towards strongly dependent variables, which alters the interpretations of the PVI values. This explains why the correlated variables X_1 and X_2 have a more important influence than the independent variables X_5 and X_6 with same weights. The same is observed for X_3 which has more influence than X_7 even thought they have the same absolute weights. Moreover, it also increases the bias toward X_4, which is normally not directly influent ($\beta_4 = 0$). On the other hand, the full and independent PVI values are less subject to the bias from the accuracy of the random forest model, yet their interpretations differ. The full PVI values integrate the influence of a variable along with its correlation with others, which explains why X_1 and X_2 have larger values (they have large β and are strongly correlated). Variables X_3 and X_4 have large influence, but it is mainly due to their correlations. Variables X_5 and X_6 are still measured as influential, while X_7 is very close to zero. The main interest in this example is for the independent PVI values which show interesting results. The procedure detects an important influence of X_5 and X_6 due to their large weights, but also because they are independent to others. Variables X_1 and X_2 are still influential, but their values are significantly less important. This is equivalent for X_7 which has larger values than X_3. Moreover, the independent PVI value is the only method to also give a null influence to X_4.

3.5 Conclusion

This chapter aims at showing the relations between the field Sensitivity Analysis of Model Output and Variable Importance in machine learning. We have illustrated some of the various difficulties in interpreting the influence of variables in a model, especially when the variables are dependent.

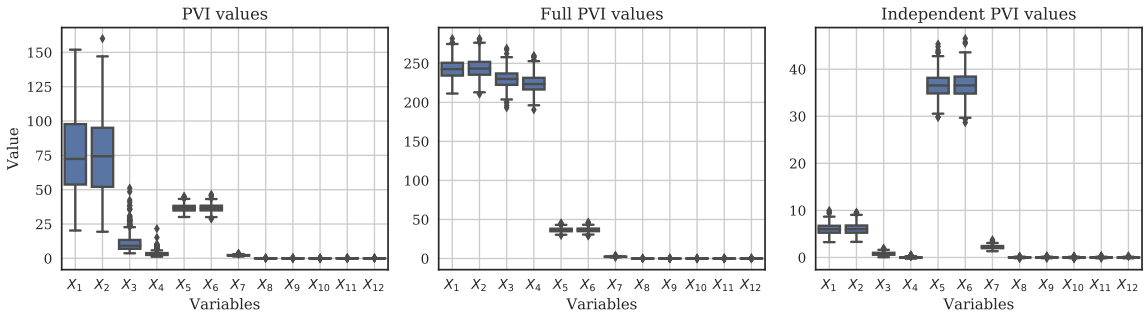


Figure 3.14 *Estimation of the PVI, full PVI and independent PVI on the $d = 12$ additive Gaussian model with four block-correlated variables and six influential variables.*

In Sensitivity Analysis, the Sobol' indices are largely used when it comes to target the variance of the model output and are strongly related to the Permutation Variable Importance. When the variables are dependent, one can consider the use of the full and independent Sobol' indices using the Rosenblatt Transformation to compute the Sobol' indices on an independent space of the input variables.

Our contributions are to enhance the relations between the Sobol' indices and the PVI values by proposing new permutation based values (the fPVI values) and to show that they measure the effect of the first-order Sobol' indices. Moreover, when the variables are dependent, the use of a Rosenblatt Transformation can also be applied for the PVI values. Using the true mathematical model η, we showed that estimation of the Sobol' indices using the permutation procedure can be a good alternative to the Monte Carlo sampling. An notable contribution is to have identified a systematic bias in the estimation of the PVI values when considering an estimated model (such as a random forest). This bias is strongly related to the accuracy of the estimation model and we observe that the original permutation procedure increases this bias significantly when some variables are strongly correlated. The use of the new full and independent PVI values can be a great alternative to deal with dependent variables as they are not impacted by the increase of bias with the dependencies. Moreover, they have a different interpretation of the influence of a variable when there are dependencies, and the independent PVI values are strongly useful to detect non-influential variables even with a strong correlation.

Future work can be done in the analysis of the systematic bias observed in Section 3.4.1. For example, propose to estimate the bias quantities through some bootstraping (for example in Wager et al. [2014]). We are convinced that the difficulties observed by Ishwaran and Lu [2018] in the evaluation of the confidence intervals for the PVI values can be linked to the observed bias.

The use of a Rosenblatt Transformation is a natural idea to generate new independent samples from strongly dependent ones. However, this transformation requires knowing the whole distribution of $\mathbf{X}$ which is not always available in practice. In that case, one can infer the distribution of $\mathbf{X}$ through different techniques, or consider another transformation such as the procedure from Iman and Conover [1982] which was also proposed in Mara et al. [2015], but this can only be

applied when the dependence structure is defined by a rank correlation matrix.

3.6 Appendix

3.6.1 Proof of Proposition 2

We can develop (3.1) such as

$$\begin{aligned}\mathcal{J}_\eta(X_j) &= \mathbb{E}\left[\Big(\eta(\mathbf{X}) - \eta(\mathbf{X}_{(-j)})\Big)^2\right] \\ &= \mathbb{E}\Big[\Big(\eta(\mathbf{X}) - \mathbb{E}[Y]\Big)^2\Big] + \mathbb{E}\Big[\Big(\eta(\mathbf{X}_{(-j)}) - \mathbb{E}[Y]\Big)^2\Big] - 2\mathbb{E}\Big[\Big(\eta(\mathbf{X}) - \mathbb{E}[Y]\Big)\Big(\eta(\mathbf{X}_{(-j)}) - \mathbb{E}[Y]\Big)\Big]\end{aligned}$$

Because the components of $\mathbf{X}$ are independent, we can say that $\mathbb{E}\Big[\Big(\eta(\mathbf{X}_{(-j)}) - \mathbb{E}[Y]\Big)^2\Big] = \mathrm{Var}[Y]$. Thus, we have

$$\mathcal{J}_\eta(X_j) = 2\mathrm{Var}[Y] - 2\mathbb{E}\Big[\Big(\eta(\mathbf{X}) - \mathbb{E}[Y]\Big)\Big(\eta(\mathbf{X}_{(-j)}) - \mathbb{E}[Y]\Big)\Big] \tag{3.21}$$

The second term of (3.21) can be reduced by writing it in form of integrals:

$$\begin{aligned}&\mathbb{E}\Big[\Big(\eta(\mathbf{X}) - \mathbb{E}[Y]\Big)\Big(\eta(\mathbf{X}_{(-j)}) - \mathbb{E}[Y]\Big)\Big] \\ &= \int_{\mathbb{R}}\left[\int_{\mathbb{R}^{d-1}}(\eta(\mathbf{x}) - \mathbb{E}[Y])\prod_{i=1,i\neq j}^{d} f_{X_i}(x_i)\mathrm{d}x_i\right]\left[\int_{\mathbb{R}^{d-1}}(\eta(\mathbf{x}_{(-j)}) - \mathbb{E}[Y])\prod_{k=1,k\neq j}^{d} f_{X_k}(x'_k)\mathrm{d}x'_k\right] f_{X_j}(x_j)\mathrm{d}x_j \\ &= \int_{\mathbb{R}}\left[\int_{\mathbb{R}^{d-1}}(\eta(\mathbf{x}) - \mathbb{E}[Y])\prod_{k=1,k\neq j}^{d} f_{X_k}(x_k)\mathrm{d}x_k\right]^2 f_{X_j}(x_j)\mathrm{d}x_j \\ &= \int_{\mathbb{R}}(\mathbb{E}[Y|X_j] - \mathbb{E}[Y])^2 f_{X_j}(x_j)\mathrm{d}x_j \\ &= \mathrm{Var}[\mathbb{E}[Y|X_j]].\end{aligned} \tag{3.22}$$

We now merge (3.22) with (3.21):

$$\mathcal{J}_\eta(X_j) = 2\mathrm{Var}[Y] - 2\mathrm{Var}[\mathbb{E}[Y|\mathbf{X}_j]] \tag{3.23}$$

$$= 2\mathbb{E}[\mathrm{Var}[Y|\mathbf{X}_j]]. \tag{3.24}$$

The fPVI value of X_j is equivalent to measuring the excepted part of variance on Y due to fixing X_j. Using the definition of the first-order Sobol' index in (2.8), we have the following relation:

$$\mathcal{J}_\eta(X_j) = 2(1 - S_i)\mathrm{Var}[Y]. \tag{3.25}$$

3.6.2 Uncorrelation of permuted variables

We consider a sample of random variables $(X_1, Y_1), \dots, (X_n, Y_n)$ with a joint law (X, Y). We also introduce $(\tilde{X}_1, \dots, \tilde{X}_n)$ the permuted variables of $(X_1, \dots, X_n)$ such as $\tilde{X}_1$ is defined by:

$$\tilde{X}_1 = X_i 1_{Z=i}, \tag{3.26}$$

where Z is an uniform and independent variable on $[1, \dots, n]$.

We aim to study the covariance between $\tilde{X}_1$ and Y_1. Thus we can write :

$$\mathrm{Cov}[\tilde{X}_1, Y_1] = \mathbb{E}[\tilde{X}_1 Y_1] - \mathbb{E}[\tilde{X}_1]\mathbb{E}[Y_1].$$

Because all X_i have the same distribution, we known that $\mathbb{E}[\tilde{X}_1] = \mathbb{E}[X_1]$. Using the law of total expectation and replacing $\tilde{X}_1$ with the expression of (3.26), we have:

$$\begin{aligned}
\mathrm{Cov}[\tilde{X}_1, Y_1] &= \mathbb{E}[X_i 1_{Z=i} Y_1] - \mathbb{E}[X_1]\mathbb{E}[Y_1] \\
&= \mathbb{E}\left[\mathbb{E}[X_i 1_{Z=i} Y_1] | Z\right] - \mathbb{E}[X_1]\mathbb{E}[Y_1] \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[X_i Y_1] - \mathbb{E}[X_1]\mathbb{E}[Y_1] \\
&= \frac{1}{n}\sum_{i=2}^{n} \mathbb{E}[X_i Y_1] + \frac{1}{n}\mathbb{E}[X_1 Y_1] - \mathbb{E}[X_1]\mathbb{E}[Y_1].
\end{aligned}$$

By independence between X_i and Y_1, for all $i > 1$, we have:

$$\begin{aligned}
\mathrm{Cov}[\tilde{X}_1, Y_1] &= \frac{1}{n}\sum_{i=2}^{n} \mathbb{E}[X_i]\mathbb{E}[Y_1] + \frac{1}{n}\mathbb{E}[X_1 Y_1] - \mathbb{E}[X_1]E[Y_1] \\
&= \frac{n-1}{n}\mathbb{E}[X_1]\mathbb{E}[Y_1] + \frac{1}{n}\mathbb{E}[X_1 Y_1] - \mathbb{E}[X_1]E[Y_1] \\
&= -\frac{1}{n}\mathbb{E}[X_1]\mathbb{E}[Y_1] + \frac{1}{n}\mathbb{E}[X_1 Y_1] \\
&= \frac{1}{n}\mathrm{Cov}[X_1, Y_1] \xrightarrow{n} 0.
\end{aligned}$$

The covariance between $\tilde{X}_1$ and Y_1 tends to zero when $n \to \infty$. This is equivalent for any $\mathrm{Cov}[\tilde{X}_i, Y_i]$ for $i = 1, \dots, n$, showing that the permutation of X tends to make the permuted variable uncorrelated to the others.

3.6.3 Estimation of Permutation Variable Importance

The estimation of the PVI values for independent random inputs is summarized in Algorithm 1.

The estimation of the PVI values for dependent random inputs is summarized in Algorithm 2.

Algorithme 1 : PVI values for independent random variables

Data : A model η, a random vector $\mathbf{X}$, a sample size n.

Generate a sample $\{\mathbf{X}_i\}_{i=1}^n$ from $\mathbf{X}$;

Evaluate the sample: $\{Y_i\}_{i=1}^n = \{\eta(\mathbf{X}_i)\}_{i=1}^n$;

Estimate the output variance $\hat{V}_Y$;

for $j = 1, \ldots, d$ **do**

Apply a permutation π among:

- the j-th variable: $\{\mathbf{X}_i^{(j)}\}_{i=1}^n = \{(X_{i,1}, \ldots, X_{\pi(i),j}, \ldots, X_{i,d})\}_{i=1}^n$
- all variables except the j-th: $\{\mathbf{X}_i^{(-j)}\}_{i=1}^n = \{(X_{\pi(i),1}, \ldots, X_j, \ldots, X_{\pi(i),d})\}_{i=1}^n$

Compute the permuted variable importance values:

$$\hat{\mathcal{I}}_\eta(X_j) = \frac{1}{n}\sum_{i=1}^n \left(Y_i - \eta(\mathbf{X}_i^{(j)})\right)^2$$

$$\hat{\mathcal{J}}_\eta(X_j) = \frac{1}{n}\sum_{i=1}^n \left(Y_i - \eta(\mathbf{X}_i^{(-j)})\right)^2$$

Estimate the Sobol' indices:

$$\widehat{ST}_j = \frac{\hat{\mathcal{I}}_\eta(X_j)}{2\hat{V}_Y} \tag{3.27}$$

$$\hat{S}_j = 1 - \frac{\hat{\mathcal{J}}_\eta(X_j)}{2\hat{V}_Y} \tag{3.28}$$

Algorithme 2 : PVI values for dependent random variables

Data : A model η, a random vector $\mathbf{X}$, a sample size n.

Generate a sample $\{\mathbf{X}_i\}_{i=1}^n$ from $\mathbf{X}$;

Evaluate the sample: $\{Y_i\}_{i=1}^n = \{\eta(\mathbf{X}_i)\}_{i=1}^n$;

Estimate the output variance $\hat{V}_Y$;

for $j = 1, \ldots, d$ **do**

a. Generate an uniform sample using a Rosenblatt transformation:

$$\{\mathbf{U}_i^j\}_{i=1}^n \xleftarrow{RT} \{(X_{i,j}, X_{i,j+1}, \ldots, X_{i,d}, X_{i,1}, \ldots, X_{i,j-1})\}_{i=1}^n ;$$

b. On the uniformed sample, apply a permutation π among

- the first variable: $\{\mathbf{U}_i^{j,(1)}\}_{i=1}^n = \{(U_{\pi(i),1}, X_{i,2}, \ldots, U_{i,d})\}_{i=1}^n$
- all variables except the first: $\{\mathbf{U}_i^{j,(-1)}\}_{i=1}^n = \{(U_{i,1}, X_{\pi(i),2}, \ldots, U_{\pi(i),d})\}_{i=1}^n$
- the d-th variable: $\{\mathbf{U}_i^{j,(d)}\}_{i=1}^n = \{(U_{i,1}, X_{i,2}, \ldots, U_{\pi(i),d})\}_{i=1}^n$
- all variables except the d-th: $\{\mathbf{U}_i^{j,(-d)}\}_{i=1}^n = \{(U_{\pi(i),1}, U_{\pi(i),2}, \ldots, U_{i,d})\}_{i=1}^n$

c. Inverse Rosenblatt transformation with reordering:

- $\{\mathbf{X}_i^{(j),full}\}_{i=1}^n \xleftarrow{IRT} \{\mathbf{U}_i^{j,(1)}\}_{i=1}^n$
- $\{\mathbf{X}_i^{(-j),full}\}_{i=1}^n \xleftarrow{IRT} \{\mathbf{U}_i^{j,(-1)}\}_{i=1}^n$
- $\{\mathbf{X}_i^{(j-1),ind}\}_{i=1}^n \xleftarrow{IRT} \{\mathbf{U}_i^{j,(d)}\}_{i=1}^n$
- $\{\mathbf{X}_i^{(-(j-1)),ind}\}_{i=1}^n \xleftarrow{IRT} \{\mathbf{U}_i^{j,(-d)}\}_{i=1}^n$

d. Compute the permuted variable importance values:

$$\hat{\mathcal{I}}_\eta^{full}(X_j) = \frac{1}{n}\sum_{i=1}^n \left(Y_i - \eta(\mathbf{X}_i^{(j),full})\right)^2$$

$$\hat{\mathcal{J}}_\eta^{full}(X_j) = \frac{1}{n}\sum_{i=1}^n \left(Y_i - \eta(\mathbf{X}_i^{(-j),full})\right)^2$$

$$\hat{\mathcal{I}}_\eta^{ind}(X_{j-1}) = \frac{1}{n}\sum_{i=1}^n \left(Y_i - \eta(\mathbf{X}_i^{(j-1),ind})\right)^2$$

$$\hat{\mathcal{J}}_\eta^{ind}(X_{j-1}) = \frac{1}{n}\sum_{i=1}^n \left(Y_i - \eta(\mathbf{X}_i^{(-(j-1)),ind})\right)^2$$

Estimate the Sobol' indices:

$$\widehat{ST}_j^{full} = \frac{\hat{\mathcal{I}}_\eta^{full}(X_j)}{2\hat{V}_Y} \tag{3.29}$$

$$\hat{S}_j^{full} = 1 - \frac{\hat{\mathcal{J}}_\eta^{full}(X_j)}{2\hat{V}_Y} \tag{3.30}$$

$$\widehat{ST}_{j-1}^{ind} = \frac{\hat{\mathcal{I}}_\eta^{ind}(X_{j-1})}{2\hat{V}_Y} \tag{3.31}$$

$$\hat{S}_{j-1}^{ind} = 1 - \frac{\hat{\mathcal{J}}_\eta^{ind}(X_{j-1})}{2\hat{V}_Y} \tag{3.32}$$

Chapter 4

Shapley effects for sensitivity analysis with dependent inputs: bootstrap and kriging-based algorithms

Abstract. In global sensitivity analysis, the well-known Sobol' sensitivity indices aim to quantify how the variance in the output of a mathematical model can be apportioned to the different variances of its input random variables. These indices are based on the functional variance decomposition and their interpretation becomes difficult in the presence of statistical dependence between the inputs. However, as there are dependencies in many application studies, this drawback enhances the development of interpretable sensitivity indices. Recently, the Shapley values that were developed in the field of cooperative games theory have been connected to global sensitivity analysis and present good properties in the presence of dependencies. Nevertheless, the available estimation methods do not always provide confidence intervals and require a large number of model evaluations. In this chapter, a bootstrap resampling is implemented in existing algorithms to assess confidence intervals. We also propose to consider a metamodel in substitution of a costly numerical model. The estimation error from the Monte Carlo sampling is combined with the metamodel error in order to have confidence intervals on the Shapley effects. Furthermore, we compare the Shapley effects with existing extensions of the Sobol' indices in different examples of dependent random variables.

Contents

4.1 Introduction . . . 80
4.2 Sobol' sensitivity indices . . . 81
4.3 Shapley effects . . . 86
4.4 Examples in Gaussian framework: analytical results and relations between indices . . . 89
4.5 Numerical studies . . . 95
4.6 Kriging metamodel with inclusion of errors . . . 100
4.7 Numerical simulations with kriging model . . . 104
4.8 Conclusion . . . 107
4.9 Appendix . . . 109

4.1 Introduction

In the last decades, computational models have been increasingly used to approximate physical phenomenons. The steady improvement of computational means has led to the use of very complex numerical codes involving an increasing number of parameters. In many situations, the model inputs are uncertain, which result in uncertain outputs. In this case it is necessary to understand the global impact of input uncertainties on the output to validate the computer code and use it properly. Sensitivity Analysis methods aim at solving this range of issues by characterizing input-output relationships of computer codes.

Within Sensitivity Analysis, three kinds of methods can be distinguished. First, Screening methods aim to discriminate influential inputs from non-influential ones, especially when the inputs are numerous and the problem should be simplified. Secondly, local methods, based on partial derivatives, are used to assess the influence of input variables for small perturbations. Last, Global Sensitivity Analysis (GSA) methods aim at ranking input random variables according to their importance in the uncertainty of the output, and can also quantify the global influence of a particular input on the output. In this chapter we are specifically interested in Global Sensitivity Analysis. One can refer to Iooss and Lemaître [2015] for a comprehensive review of sensitivity analysis methods.

Among GSA methods, variance-based approaches are a class of probabilistic ones that measure the part of variance of the model output which is due to the variance of a particular input. These methods were popularized by Sobol [1993] who introduced the well-known first order Sobol' indices. Shortly after, the total Sobol' indices have been introduced by Homma and Saltelli [1996] also taking advantage of Jansen et al. [1994]. These sensitivity indices are based on the functional ANalysis Of VAriance (ANOVA), the decomposition of which is unique only if the input random variables are assumed independent. However this hypothesis is sometimes not verified in practice, making their interpretation much harder. Several works have been carried out to adress this difficulty and they extend the Sobol' indices to the case of a stochastic dependence between the input variables, such as Chastaing et al. [2012]; Mara and Tarantola [2012]; Mara

et al. [2015]; Kucherenko et al. [2012]. Nonetheless, the practical estimation of these sensitivity measures and their interpretation remain difficult.

Recently, Owen [2014] established a relation between the Shapley values [Shapley and Shubik, 1954] coming from the field of game theory and Sobol' indices. Song et al. [2016] proposed an algorithm to estimate these indices. Some studies also highlighted the potential of this kind of index in the case of correlated input, such as Owen and Prieur [2017]; Iooss and Prieur [2017]. In this last case, the Shapley effects can be a good alternative to the existing extensions of Sobol' indices mentioned above. Indeed, Shapley effects allow an apportionment of the interaction and dependences contributions between the input involved, making them condensed and easy-to-interpret indices.

Most estimation procedures of the Sobol' indices and Shapley effects are based on Monte Carlo sampling. These methods require large sample sizes in order to have a sufficiently low estimation error. When dealing with costly computational models, a precise estimation of these indices can be difficult to achieve or even unfeasible. Therefore, the use of a surrogate model (or metamodel) instead of the actual model can be a good alternative and dramatically decrease the computational cost of the estimation. Various kinds of surrogate models exist in literature, such as Fang et al. [2005]. In this chapter, we are interested in the use of kriging as metamodels (see for example Martin and Simpson [2004]). One particular approach, developped by Le Gratiet et al. [2014], proposed an estimation algorithm of the Sobol' indices using kriging models which also provides the meta-model and Monte Carlo errors.

In this paper, we draw a comparison between the Shapley effects and the *independent* and *full* Sobol' indices defined in Mara et al. [2015]. We also establish an extension of the Shapley estimation algorithm proposed in Song et al. [2016] by implementing a bootstrap sampling to catch the Monte Carlo error. Inspired by the work of Le Gratiet et al. [2014], we used a kriging model in substitution of the true model for the estimation of these indices. Thus, the kriging model error is associated to the Monte Carlo error in order to correctly catch the overall estimation error.

The paper's outline is as follows: Section 4.2 recalls the basic concept of Sobol' indices in the independent and dependent configuration; Section 4.3 introduces the Shapley values and their links with sensitivity analysis; Section 4.4 theoretically compares the Sobol' indices and the Shapley effects for two toy examples; Section 4.5 studies the quality of the estimated Shapley effects and their confidence intervals; Section 4.6 introduces the kriging model and how the kriging and Monte Carlo errors can be separated from the overall error; Section 4.7 compares the indice performances using a kriging model on two toy examples; finally, Section 4.8 synthesizes this work and suggests some perspectives.

4.2 Sobol' sensitivity indices

4.2.1 Sobol' indices with independent inputs

Consider a model $Y = \eta(\mathbf{X})$ with d random inputs denoted by $\mathbf{X}_{\mathcal{D}} = \{X_1, X_2, \dots, X_d\}$, where $\mathcal{D} = \{1, 2, \dots, d\}$, and $\mathbf{X}_{\mathcal{J}}$ indicates the vector of inputs corresponding to the index set $\mathcal{J} \subseteq \mathcal{D}$. $\eta : \mathbb{R}^d \to \mathbb{R}$ is a deterministic squared integrable function and $Y \in \mathbb{R}$ the model output random variable. The random vector $\mathbf{X}$ follows a distribution $p_{\mathbf{X}}$ and we suppose, in this section, that $p_{\mathbf{X}}$ follows a d-dimensional uniform distribution $\mathcal{U}([0,1]^d)$. However, these results can be extended to any marginal distributions. In particular, all inputs are independent and the distribution of $\mathbf{X}$ is only defined by its margins.

The Hoeffding decomposition introduced in Hoeffding [1948], also known as high dimensional model representation (HDMR) [Li et al., 2001], allows writing $\eta(\mathbf{X})$ in the following way:

$$\eta(\mathbf{X}) = \eta_\emptyset + \sum_{i=1}^{d} \eta_i(X_i) + \sum_{1 \leqslant i < j \leqslant d} \eta_{i,j}(X_i, X_j) + \cdots + \eta_{1,\dots,d}(\mathbf{X}), \tag{4.1}$$

for some $\eta_\emptyset, \eta_i, \dots, \eta_{1,\dots,d}$ set of functions. In this formula, η is decomposed into 2^d terms such as $\eta_\emptyset$ is a constant and the other terms are square integrable functions.

The decomposition (4.1) is not unique due to the infinite possible choices for these terms. The uniqueness condition is granted by the following orthogonality constraint:

$$\int_0^1 \eta_{i_1,i_2,\dots,i_s}(x_{i_1}, x_{i_2}, \dots, x_{i_s}) \mathrm{d}x_{i_w} = 0, \tag{4.2}$$

where $1 \leq i_1 < i_2 < \cdots < i_s \leq d$ and $i_w \in \{i_1, i_2, \dots, i_s\}$. The consequence of this condition is that the terms of (4.1) are orthogonal to one another. This property implies the independence of the random variables X_i in the stochastic configuration and allow to obtain the following expressions for the functions $\eta_{i_1,i_2,\dots,i_s}$ of (4.1) :

$$\eta_\emptyset = \mathbb{E}(Y), \tag{4.3}$$
$$\eta_i(X_i) = \mathbb{E}_{\mathbf{X}_{\sim i}}(Y|X_i) - \mathbb{E}(Y), \tag{4.4}$$
$$\eta_{i,j}(X_i, X_j) = \mathbb{E}_{\mathbf{X}_{\sim ij}}(Y|X_i, X_j) - \eta_i - \eta_j - \mathbb{E}(Y) \tag{4.5}$$

where $\mathbf{X}_{\sim i} = \mathbf{X}_{\mathcal{D}\setminus\{i\}}$ (the vector $\mathbf{X}$ without X_i), and similarly for higher orders. Thus, the functions $\{\eta_i\}_{i=1}^d$ are the main effects, the $\eta_{i,j}$ for $i < j = 1, \dots, d$ are the second-order interaction effects, and so on.

The representation (4.1) leads to the functional ANAlysis Of VAriance (ANOVA) which consists in expanding the global variance into a sum of partial variances such as

$$\mathrm{Var}(Y) = \sum_{i=1}^{d} \mathrm{Var}[\eta_i(X_i)] + \sum_{i=1}^{d} \sum_{i<j}^{d} \mathrm{Var}[\eta_{i,j}(X_i, X_j)] + \cdots + \mathrm{Var}[\eta_{1,\dots,d}(\mathbf{X})]. \tag{4.6}$$

The so-called Sobol' indices [Sobol, 1993] can be derived from (4.6) by dividing both sides with $\mathrm{Var}(Y)$. This operation results in the following property:

$$\sum_{i=1}^{d} S_i + \sum_{i=1}^{d}\sum_{i<j}^{d} S_{ij} + \cdots + S_{1,\ldots,d} = 1, \tag{4.7}$$

where S_i is a first-order sensitivity index, S_{ij} is a second-order sensitivity index and so on. Thus, sensitivity indices are defined as

$$S_i = \frac{\mathrm{Var}[\eta_i(X_i)]}{\mathrm{Var}(Y)}, \qquad S_{ij} = \frac{\mathrm{Var}[\eta_{i,j}(X_i, X_j)]}{\mathrm{Var}(Y)}, \qquad \ldots \tag{4.8}$$

The first-order index S_i measures the part of variance of the model output that is due to the variable X_i, the second-order S_{ij} measure the part of variance of the model output that is due to the interaction of X_i and X_j and so on for higher interaction orders.

Another popular variance based coefficient, called total Sobol' index by Homma and Saltelli [1996], gathers the first-order effect of a variable with all its interactions. This index is defined by

$$ST_i = S_i + \sum_{i\neq j} S_{ij} + \cdots + S_{1,\ldots,d} = 1 - \frac{\mathrm{Var}_{\mathbf{X}_{\sim i}}[\mathbb{E}_{X_i}(Y|\mathbf{X}_{\sim i})]}{\mathrm{Var}(Y)} = \frac{\mathbb{E}_{\mathbf{X}_{\sim i}}[\mathrm{Var}_{X_i}(Y|\mathbf{X}_{\sim i})]}{\mathrm{Var}(Y)}. \tag{4.9}$$

The property (4.7) does not always hold for the total indices as summing total indices for all variables introduces redundant interactions terms appearing only once in (4.7). Thus, in most cases $\sum_i^d ST_i \geq 1$. Note that both the first order and total Sobol' indices are normalized measures. We refer to Iooss and Lemaître [2015] for an exhaustive review on the sensitivity indices and their properties.

As mentioned earlier, (4.6) only holds if the random variables are independent. Different approaches exist to treat the case of dependent input and one of them is explained in the following section.

4.2.2 Sobol' indices with dependent inputs

In this section, we suppose $\mathbf{X} \sim p_{\mathbf{X}}$ with dependent random inputs. Thanks to the Rosenblatt Transformation (RT) [Rosenblatt, 1952], it is possible to transform $\mathbf{X}$ into a random vector $\mathbf{U} \sim \mathcal{U}^d(0,1)$ with independent and uniformly distributed entries. For the following ordering of the components of $\mathbf{X} = (X_1, \ldots, X_k, \ldots, X_d)$, let $\mathbf{u} = T(\mathbf{x})$ where T is a transformation defined by

$$T : \mathbb{R}^d \to [0,1]^d$$

$$\mathbf{x} \mapsto \mathbf{u} = \begin{pmatrix} F_1(x_1) \\ \vdots \\ F_{k|1,\ldots,k-1}(x_k|x_1,\ldots,x_{k-1}) \\ \vdots \\ F_{d|1,\ldots,d-1}(x_d|x_1,\ldots,x_{d-1}) \end{pmatrix}$$

where $F_{k|1,\dots,k-1}$ is the conditional cumulative distribution function of X_k conditioned by $X_1, \dots, X_{k-1}$.

However, several RT are possible due to the $d!$ different permutations of the elements of $\mathbf{X}$. Note that in this procedure, only the d Rosenblatt Transformations obtained after circularly reordering the elements of $\mathbf{X}$ are considered. We denote as $\mathbf{U}^i = (U_1^i, \dots, U_d^i)$ the random vector obtained from the RT of the set $(X_i, X_{i+1}, \dots, X_d, X_1, \dots, X_{i-1})$ such as

$$(X_i, X_{i+1}, \dots, X_d, X_1, \dots, X_{i-1}) \sim p_{\mathbf{X}} \xrightarrow{T} (U_1^i, \dots, U_d^i) \sim \mathcal{U}^d(0,1). \tag{4.10}$$

It is important to note that this RT corresponds to a particular i-th ordering. Changing this order results in another RT. Such a mapping is bijective and we can consider a function g_i such as $Y = \eta(\mathbf{X}) = g_i(\mathbf{U}^i)$. Because the elements of $\mathbf{U}^i$ are independent, the ANOVA decomposition is unique and can be established to compute sensitivity indices. Thus, we can write

$$g_i(\mathbf{U}^i) = g_\emptyset + \sum_{k=1}^{d} g_k(U_k^i) + \sum_{k=1}^{d}\sum_{k<l}^{d} g_{k,l}(U_k^i, U_l^i) + \dots + g_{1,\dots,d}(U_1^i, \dots, U_d^i) \tag{4.11}$$

where $g_\emptyset = \mathbb{E}[g_i(\mathbf{U}^i)]$. Because the summands in (4.11) are orthogonal, the variance based decomposition can be derived, such that

$$\mathrm{Var}(Y) = \mathrm{Var}[g_i(\mathbf{U}^i)] = \sum_{k=1}^{d} V_k + \sum_{k=1}^{d}\sum_{k<l}^{d} V_{k,l} + \dots + V_{1,\dots,d} \tag{4.12}$$

where $V_k = \mathrm{Var}[\mathbb{E}(g_i(\mathbf{U}^i)|U_k^i)]$, $V_{k,l} = \mathrm{Var}[\mathbb{E}(g_i(\mathbf{U}^i)|U_k, U_l)] - V_k - V_l$ and so on for higher orders. The Sobol' indices are defined by dividing (4.12) with the total variance such that,

$$S_k^i = \frac{\mathrm{Var}[\mathbb{E}[g_i(\mathbf{U}^i)|U_k^i]]}{\mathrm{Var}[g_i(\mathbf{U}^i)]}. \tag{4.13}$$

We also consider the total Sobol' indices which are the overall contribution of U_k^i on the model output including the marginal and interaction effects. They can be written as

$$ST_k^i = \frac{\mathbb{E}[\mathrm{Var}[g_i(\mathbf{U}^i)|\mathbf{U}_{\sim k}^i]]}{\mathrm{Var}[g_i(\mathbf{U}^i)]}, \tag{4.14}$$

where $\mathbf{U}_{\sim k}^i = \mathbf{U}_{\mathcal{D}\setminus\{k\}}^i$. We remind that (4.13) and (4.14) are derived from the RT of the ordered set $(X_i, X_{i+1}, \dots, X_d, X_1, \dots, X_{i-1})$. The RT in equation (4.10) determines the following mapping between $\mathbf{X}$ and $\mathbf{U}^i$:

$$\Big[(X_i), (X_{i+1}|X_i), \dots, (X_1|X_i, X_{i+1}, \dots, X_d), \dots, (X_{i-1}|\mathbf{X}_{\sim(i-1)})\Big] \longleftrightarrow (U_1^i, U_2^i, \dots, U_d^i),$$

where $U_1^i = F_i(X_i)$, $U_2^i = F_{i+1|i}(X_{i+1}|X_i)$ and so on for other variables. From here, we only consider the variables U_1^i and U_d^i because they present interesting properties. Indeed, the variable U_1^i is representative of the behavior of X_i taking into account the dependence with other variables. On the opposite, the variable U_d^i represents the effects of X_{i-1} that is not due to its dependence with other variables. As a consequence, Mara et al. [2015] introduced the following indices:

- the *full* Sobol' indices which describe the influence of a variable including its dependence with other variables

$$S_i^{full} = \frac{\mathrm{Var}[\mathbb{E}[g_i(\mathbf{U}^i)|U_1^i]]}{\mathrm{Var}[g_i(\mathbf{U}^i)]} = \frac{\mathrm{Var}[\mathbb{E}[\eta(\mathbf{X})|X_i]]}{\mathrm{Var}[\eta(\mathbf{X})]} \tag{4.15}$$

$$ST_i^{full} = \frac{\mathbb{E}[\mathrm{Var}[g_i(\mathbf{U}^i)|\mathbf{U}_{\sim 1}^i]]}{\mathrm{Var}[g_i(\mathbf{U}^i)]} = \frac{\mathbb{E}[\mathrm{Var}[\eta(\mathbf{X})|(\mathbf{X}_{\sim i}|X_i)]]}{\mathrm{Var}[\eta(\mathbf{X})]} \tag{4.16}$$

  where $\mathbf{X}_{\sim i}|X_i$ represent all components except X_i not taking account the dependence with the variable X_i.

- the *independent* Sobol' indices which describe the influence of variables without its dependence with other variables.

$$S_i^{ind} = \frac{\mathrm{Var}[\mathbb{E}[g_{i+1}(\mathbf{U}^{i+1})|U_d^{i+1}]]}{\mathrm{Var}[g_{i+1}(\mathbf{U}^{i+1})]} = \frac{\mathrm{Var}[\mathbb{E}[\eta(\mathbf{X})|(X_i|\mathbf{X}_{\sim i})]]}{\mathrm{Var}[\eta(\mathbf{X})]} \tag{4.17}$$

$$ST_i^{ind} = \frac{\mathbb{E}[\mathrm{Var}[g_{i+1}(\mathbf{U}^{i+1})|\mathbf{U}_{\sim d}^{i+1}]]}{\mathrm{Var}[g_{i+1}(\mathbf{U}^{i+1})]} = \frac{\mathbb{E}[\mathrm{Var}[\eta(\mathbf{X})|\mathbf{X}_{\sim i}]]}{\mathrm{Var}[\eta(\mathbf{X})]} \tag{4.18}$$

  where $X_i|\mathbf{X}_{\sim i}$ represent the component X_i not taking account the dependence with other variables and with the convention that $\mathbf{U}^{d+1} = \mathbf{U}^1$ and $g_{d+1} = g_1$.

Thanks to the RT, we can also define the sensitivity indices of $(X_i|X_u), i = 1, \ldots, d$ and $u \subset \mathcal{D}\backslash\{i\}, u \neq \emptyset$ via U_u^i which represent the effect of X_i without its mutual dependent contribution with X_u. These indices can be estimated with a Monte Carlo algorithm and the procedure is described in the next section.

4.2.3 Estimation

The estimation of (S_i^{full}, ST_i^{full}, S_{i-1}^{ind}, ST_{i-1}^{ind}) can be done with four samples using the "pick and freeze" strategy (see Saltelli et al. [2010]). The procedure is divided in two steps:

- Generate and prepare the samples:
  - generate two independent sampling matrices $\mathbf{A}$ and $\mathbf{B}$ of size $N \times d$ with $\mathcal{U}(0,1)^d$ rows,
  - creates $\mathbf{B}_{\mathbf{A}}^{(1)}$ (resp. $\mathbf{B}_{\mathbf{A}}^{(d)}$) in the following way: keep all columns from $\mathbf{B}$ except the 1-th (resp. d-th) column which is taken from $\mathbf{A}$.
- Compute the indices with a given estimator:

$$\widehat{S}_i^{full} = \frac{\frac{1}{N}\sum_{j=1}^N g_i(A)_j g_i(\mathbf{B}_{\mathbf{A}}^{(1)})_j - g_{i_0}^2}{\widehat{V}} \tag{4.19}$$

$$\widehat{ST}_i^{full} = 1 - \frac{\frac{1}{N}\sum_{j=1}^N g_i(B)_j g_i(\mathbf{B}_{\mathbf{A}}^{(1)})_j - g_{i_0}^2}{\widehat{V}} \tag{4.20}$$

$$\widehat{S}_{i-1}^{ind} = \frac{\frac{1}{N}\sum_{j=1}^N g_i(A)_j g_i(\mathbf{B}_{\mathbf{A}}^{(d)})_j - g_{i_0}^2}{\widehat{V}} \tag{4.21}$$

$$\widehat{ST}_{i-1}^{ind} = 1 - \frac{\frac{1}{N}\sum_{j=1}^N g_i(B)_j g_i(\mathbf{B}_{\mathbf{A}}^{(d)})_j - g_{i_0}^2}{\widehat{V}} \tag{4.22}$$

where g_{i_0} is the estimate of the mean and $\widehat{V} = \frac{1}{N}\sum_{j=1}^{N}(g_i(A)_j)^2 - g_{i_0}^2$

This procedure considers the estimator from Janon et al. [2014] and the overall cost is $4dN$ with N the number of samples. However, another estimator can be used to estimate the indices. See Saltelli et al. [2010] for a review of various estimators of sensitivity indices.

4.3 Shapley effects

The purpose of the Sobol' indices is to decompose $\mathrm{Var}(Y)$ and allocate it to *each subset* $\mathcal{J}$ whereas the Shapley effects decompose $\mathrm{Var}(Y)$ and allocate it to *each input* X_i. This difference allows to consider any variables regardless of their dependence with other inputs.

4.3.1 Definition

One of the main issues in cooperative games theory is to define a relevant way to allocate the earnings between players. A fair share of earnings of a d players coalition has been proposed in Shapley [1953]. Formally, in Song et al. [2016] a d-player game with the set of players $\mathcal{D} = \{1, 2, \ldots, d\}$ is defined as a real-valued function that maps a subset of $\mathcal{D}$ to its corresponding cost, i.e., $c : 2^{\mathcal{D}} \mapsto \mathbb{R}$ with $c(\emptyset) = 0$. Hence, $c(\mathcal{J})$ represents the cost that arises when the players in the subset $\mathcal{J}$ of $\mathcal{D}$ participate in the game. The Shapley value of player i with respect to $c(\cdot)$ is defined as

$$v^i = \sum_{\mathcal{J}\subseteq\mathcal{D}\setminus\{i\}} \frac{(d - |\mathcal{J}| - 1)!|\mathcal{J}|!}{d!} \left(c\left(\mathcal{J}\cup\{i\}\right) - c\left(\mathcal{J}\right)\right) \tag{4.23}$$

where $|\mathcal{J}|$ indicates the size of $\mathcal{J}$. In other words, v^i is the incremental cost of including player i in set $\mathcal{J}$ averaged over all sets $\mathcal{J} \subseteq \mathcal{D}\ \{i\}$.

This formula can be transposed to the field of global sensitivity analysis [Owen, 2014] if we consider the set of inputs of $\eta(\cdot)$ as the set of players $\mathcal{D}$. We then need to define a $c(\cdot)$ function such that for $\mathcal{J} \subseteq \mathcal{D}$, $c(\mathcal{J})$ measures the part of variance of Y caused by the uncertainty of the inputs in $\mathcal{J}$. To this aim, we want a cost function that verifies $c(\emptyset) = 0$ and $c(\mathcal{D}) = 1$.

Functions $\tilde{c}(\mathcal{J}) = \mathrm{Var}\left[\mathbb{E}\left[Y|\mathbf{X}_{\mathcal{J}}\right]\right]/\mathrm{Var}(Y)$ and $c(\mathcal{J}) = \mathbb{E}\left[\mathrm{Var}\left[Y|\mathbf{X}_{-\mathcal{J}}\right]\right]/\mathrm{Var}(Y)$ satisfy the two conditions above. Besides, according to Theorem 1 of Song et al. [2016], the Shapley values calculated using both cost functions $\tilde{c}(\mathcal{J})$ and $c(\mathcal{J})$ are the same.

However, for some reasons described at the end of the section 3.1 of the article Song et al. [2016], about the estimation of these two cost functions, it is better to define the Shapley effect of the i-th input, Sh^i, as the Shapley value obtained by applying the cost function c instead of $\tilde{c}$. We denote in the sequel the Shapley effect by Sh^i and a generic Shapley value by v^i. A valuable property of the Shapley effects defined in this way is that they are non-negative and they sum to one. Each one can therefore be interpreted as a measure of the part of the variance of Y related to the i-th input of η.

4.3.2 Estimation of the Shapley effects

An issue with the Shapley value is its computational complexity as all possible subsets of the players need to be considered. Castro et al. [2009] proposed an estimation method based on an alternative definition of the Shapley value.

Indeed, the Shapley value can also be expressed in terms of all possible permutations of the players. Let us denote by $\Pi(\mathcal{D})$ the set of all possible permutations with player set $\mathcal{D}$. Given a permutation $\pi \in \Pi(\mathcal{D})$, define the set $P_i(\pi)$ as the players that precede player i in π. Thus, the Shapley value can be rewritten in the following way :

$$v^i = \frac{1}{d!} \sum_{\pi \in \Pi(\mathcal{D})} \left[c\left(P_i(\pi) \cup \{i\}\right) - c\left(P_i(\pi)\right) \right] \tag{4.24}$$

From this formula, Castro et al. [2009] proposed to estimate v^i with $\widehat{v}^i$ by drawing randomly m permutations in $\Pi(\mathcal{D})$ and thus we have :

$$\widehat{v}^i = \frac{1}{m} \sum_{l=1}^{m} \Delta_i c(\pi_l) \tag{4.25}$$

with $\Delta_i c(\pi_l) = c\left(P_i(\pi) \cup \{i\}\right) - c\left(P_i(\pi)\right)$ and $c(\cdot)$ the cost function.

Section 4 of Song et al. [2016] proposed some improvements on the Castro's algorithm by including the Monte Carlo estimation $\widehat{c}$ of the cost function $c(\mathcal{J}) = \mathbb{E}\left[\mathrm{Var}\left[Y|\mathbf{X}_{-\mathcal{J}}\right]\right] / \mathrm{Var}(Y)$ to estimate the Shapley effects. The estimator writes:

$$\widehat{Sh}^i = \frac{1}{m} \sum_{l=1}^{m} \left[\widehat{c}\left(P_i(\pi_l) \cup \{i\}\right) - \widehat{c}\left(P_i(\pi_l)\right) \right] \tag{4.26}$$

where m refers to the number of permutations. Song et al. [2016] proposed the following two algorithms, the main features of which are spelled out below:

- The *exact permutation method* if d is small, one does all possible permutations between the inputs (i.e. $m = d!$);
- The *random permutation method* which consists in randomly sampling m permutations of the inputs in $\Pi(\mathcal{D})$.

For each iteration of this loop on the inputs' permutations, a conditional variance expectation must be computed. The cost C of these algorithms is the following $C = N_v + m(d-1)N_o N_i$ with N_v the sample size for the variance computation of Y, N_o the outer loop size for the expectation, N_i the inner loop size for the conditional variance of Y and m the number of permutations according to the selected method.

Note that the full first-order Sobol' indices and the independent total Sobol' indices can be also estimated by applying these algorithms, each one during only one loop iteration.

Based on theoretical results, Song et al. [2016] recommends to fix parameters at the following values to obtain an accurate approximation of Shapley effects that is computationally affordable:

- The *exact permutation method*: N_o as large as possible and $N_i = 3$;
- The *random permutation method*: $N_o = 1$, $N_i = 3$ and m as large as possible.

The choice of N_v is independent from these values and Iooss and Prieur [2017] have also illustrated the convergence of two numerical algorithms for estimating Shapley effects.

4.3.3 Confidence interval for the Shapley effects

In this section, we propose a methodology to compute confidence intervals for the Shapley effects, in order to quantify the Monte Carlo error (sampling error).

Exact permutation method: bootstrap

Concerning this algorithm, we will use the bias-corrected percentile method of the Bootstrap [Efron, 1981].

Let be $\widehat{\theta}(X_1, \ldots, X_n)$ be an estimator of a unknown parameter θ, function of n independent and identically distributed observations of law $\mathcal{F}$. In non-parametric Bootstrap, from a n-sample $(x_1, \ldots, x_n)$, we compute $\widehat{\theta}(x_1, \ldots, x_n)$. After, we draw with replacement a bootstrap sample $(x_1^*, \ldots, x_n^*)$ from the original sample $(x_1, \ldots, x_n)$ and compute $\theta^* = \widehat{\theta}(x_1^*, \ldots, x_n^*)$. We repeat this procedure B times and obtain B bootstrap replications $\theta_1^*, \ldots, \theta_B^*$ which allows the estimate of the following confidence interval of level $1 - \alpha$ for θ:

$$\left[\widehat{G}^{-1} \circ \Phi(2\widehat{z}_0 + z_{\alpha/2}) \quad ; \quad \widehat{G}^{-1} \circ \Phi(2\widehat{z}_0 - z_{\alpha/2})\right] \tag{4.27}$$

where

- Φ is the cdf of a standard normal distribution;
- $z_{\alpha/2}$ percentile of level $\alpha/2$ of $\mathcal{N}(0, 1)$;
- $\widehat{G}$ is the cdf of the bootstrap distribution for the estimator $\widehat{\theta}$;
- and $\widehat{z}_0 = \Phi^{-1} \circ \widehat{G}(\widehat{\theta})$ is a bias correction constant.

This confidence interval has been justified in Efron [1981] when there exists an increasing transformation $g(.)$ such that $g(\widehat{\theta}) - g(\theta) \sim \mathcal{N}(-z_0\sigma, \sigma^2)$ and $g(\widehat{\theta}^*) - g(\widehat{\theta}) \sim \mathcal{N}(-z_0\sigma, \sigma^2)$ for some constants $z_0 \in \mathbb{R}$ and $\sigma > 0$. In the sequel, we will see in our examples that $g(.)$ can be considered as identity.

Thus, we need independent observations to obtain this interval but in our case as there is conditioning in the Shapley effects (more exactly in the cost function), it is not possible. To overcome this problem and estimate correctly the cdf $\widehat{G}(.)$, we make a bootstrap by blocks (on the N_o blocks) in order to use independent observations and preserve the correlation within each one. This strategy allowed to develop Algorithm 3 in order to obtain the distribution of $\widehat{Sh}^i$ to calcule the confidence interval for Sh^i.

Algorithme 3 : Compute confidence intervals for Sh^i

1 Generate a sample $\mathbf{x}^{(1)}$ of size N_v from the random vector $\mathbf{X}$;
2 Compute $\mathbf{y}^{(1)}$ from $\mathbf{x}^{(1)}$ to estimate $\mathrm{Var}(Y)$;
3 Generate a sample $\mathbf{x}^{(2)}$ of size $m(d-1)N_oN_i$ from the different conditional laws necessary to estimate $\mathbb{E}\left[\mathrm{Var}\left[Y|\mathbf{X}_{-\mathcal{J}}\right]\right]$;
4 Compute $\mathbf{y}^{(2)}$ from $\mathbf{x}^{(2)}$;
5 Compute $\widehat{Sh}^i$ thanks to Equation (4.26) ;
6 **for** $b = 1, \ldots, B$ **do**
7 Sample with replacement a realization $\tilde{\mathbf{y}}^{(1)}$ of $\mathbf{y}^{(1)}$ to compute $\mathrm{Var}(Y)$;
8 Sample by bloc with replacement a realization $\tilde{\mathbf{y}}^{(2)}$ of $\mathbf{y}^{(2)}$;
9 Compute $\widehat{Sh}_b^i$ thanks to Equation (4.26). ;
10 Compute confidence intervals for Sh^i with (4.27).

It is worth mentioning that confidence intervals for the Shapley effects can also be calculated from the Central Limit Theorem (CLT) on the outer loop (Monte Carlo sample of size No) as looss and Prieur [2017] performed it. However, it is also necessary to establish a method based on the Bootstrap in order to design in the sequel an algorithm which allows to correctly distinguish the metamodel and Monte Carlo errors.

Random permutation method: CLT

For the random permutation method, we have two options to calculate confidence intervals.

- The first one is to use the CLT like looss and Prieur [2017]. Indeed, in Castro et al. [2009] the CLT gives us:

$$\widehat{Sh}^i \xrightarrow[m\to\infty]{\mathcal{L}} \mathcal{N}\left(Sh^i, \frac{\sigma^2}{m}\right) \tag{4.28}$$

with $\sigma^2 = \dfrac{\mathrm{Var}\left(\Delta_i c(\pi_l)\right)}{\mathrm{Var}(Y)^2}$.

Thus, by estimating σ by $\widehat{\sigma}$ we have the following $1-\alpha$ asymptotic confidence interval for the Shapley effects :

$$Sh^i \in \left[\widehat{Sh}^i + z_{\alpha/2}\frac{\widehat{\sigma}}{\sqrt{m}} \quad ; \quad \widehat{Sh}^i - z_{\alpha/2}\frac{\widehat{\sigma}}{\sqrt{m}}\right]$$

with $z_{\alpha/2}$ percentile of level $\alpha/2$ of $\mathcal{N}$ (0,1).

- The second one is we can estimate the confidence interval doing a bootstrap on the permutations. We describe in Algorithm 4 the procedure allowing to do that.

Algorithme 4 : Compute confidence intervals for Sh^i

Generate a sample $\mathbf{x}^{(1)}$ of size N_v from the random vector $\mathbf{X}$;
Compute $\mathbf{y}^{(1)}$ from $\mathbf{x}^{(1)}$ to estimate $\mathrm{Var}(Y)$;
Draw randomly m permutations in $\Pi(\mathcal{D})$;
Generate a sample $\mathbf{x}^{(2)}$ of size $m(d-1)N_oN_i$ from the different conditional laws necessary to estimate $\mathbb{E}\left[\mathrm{Var}\left[Y|\mathbf{X}_{-\mathcal{J}}\right]\right]$;
Compute $\mathbf{y}^{(2)}$ from $\mathbf{x}^{(2)}$;
Compute $\widehat{Sh}^i$ thanks to Equation (4.26) ;
for $b = 1, \ldots, B$ **do**
    Sample with replacement a realization $\tilde{\mathbf{y}}^{(1)}$ of $\mathbf{y}^{(1)}$ to compute $\mathrm{Var}(Y)$;
    Sample with replacement m permutations from the original sample and retrieve in $\mathbf{y}^{(2)}$ those corresponding to drawn bootstrap permutations ;
    Compute $\widehat{Sh}_b^i$ thanks to Equation (4.26). ;
Compute confidence intervals for Sh^i with (4.27).

4.4 Examples in Gaussian framework: analytical results and relations between indices

In this section, we compare and interpret the analytic results of the studied indices for two different Gaussian models: an interactive and a linear model. We study the variation of the indices by varying the correlation between the input random variables.

4.4.1 Interactive model with two inputs

Let us consider a interactive model

$$Y = (\beta_1 X_1) \times (\beta_2 X_2) \tag{4.29}$$

with $\mathbf{X} \sim \mathcal{N}(0, \Sigma)$. We consider two cases: a model with independent variables and another with dependent variables. So we have the two following covariance matrices:

$$\Sigma = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

with $-1 \leq \rho \leq 1, \sigma_1 > 0, \sigma_2 > 0$.

From the definition of sensitivity indices, for $j = 1, 2$, we get for these models the results presented in Table 4.1.

In the independent model, the independent and full first-order Sobol indices are null because there is no dependence and the inputs have no marginal contribution. Thus, the independent

Independent model	Dependent model
Model variance	
$\sigma^2 = \mathrm{Var(Y)} = \beta_1^2\beta_2^2\sigma_1^2\sigma_2^2$	$\sigma^2 = \mathrm{Var(Y)} = (1+\rho^2)\beta_1^2\beta_2^2\sigma_1^2\sigma_2^2$
Independent first-order Sobol'indices	
$S_1^{ind} = 0$	$S_1^{ind} = 0$
$S_2^{ind} = 0$	$S_2^{ind} = 0$
Independent total Sobol'indices	
$\sigma^2 ST_1^{ind} = \beta_1^2\beta_2^2\sigma_1^2\sigma_2^2$	$\sigma^2 ST_1^{ind} = (1-\rho^2)\beta_1^2\beta_2^2\sigma_1^2\sigma_2^2$
$\sigma^2 ST_2^{ind} = \beta_1^2\beta_2^2\sigma_1^2\sigma_2^2$	$\sigma^2 ST_2^{ind} = (1-\rho^2)\beta_1^2\beta_2^2\sigma_1^2\sigma_2^2$
Full first-order Sobol'indices	
$S_1^{full} = 0$	$\sigma^2 S_1^{full} = 2\rho^2\beta_1^2\beta_2^2\sigma_1^2\sigma_2^2$
$S_2^{full} = 0$	$\sigma^2 S_2^{full} = 2\rho^2\beta_1^2\beta_2^2\sigma_1^2\sigma_2^2$
Full total Sobol'indices	
$\sigma^2 ST_1^{full} = \beta_1^2\beta_2^2\sigma_1^2\sigma_2^2$	$\sigma^2 ST_1^{full} = (1+\rho^2)\beta_1^2\beta_2^2\sigma_1^2\sigma_2^2$
$\sigma^2 ST_2^{full} = \beta_1^2\beta_2^2\sigma_1^2\sigma_2^2$	$\sigma^2 ST_1^{full} = (1+\rho^2)\beta_1^2\beta_2^2\sigma_1^2\sigma_2^2$
Shapley effects	
$\sigma^2 Sh^1 = \frac{1}{2}\beta_1^2\beta_2^2\sigma_1^2\sigma_2^2$	$\sigma^2 Sh^1 = \frac{1}{2}(1+\rho^2)\beta_1^2\beta_2^2\sigma_1^2\sigma_2^2$
$\sigma^2 Sh^2 = \frac{1}{2}\beta_1^2\beta_2^2\sigma_1^2\sigma_2^2$	$\sigma^2 Sh^2 = \frac{1}{2}(1+\rho^2)\beta_1^2\beta_2^2\sigma_1^2\sigma_2^2$

Table 4.1 *Sensitivity indices of independent and dependent Gaussian models*

and full total Sobol indices represent the variability in the model, which is due to interactions only. These ones are each equal to the variance model, i.e. each input is fully responsible of the model uncertainty, due to its interaction with the other variable. In contrast, the Shapley effects award fairly, i.e. half of the interaction effect to each input, which is more logical.

About the dependent model, $S_j^{ind} = 0, j = 1,2$ are still null because the inputs have no uncorrelated marginal contribution. But now, $S_j^{full} \neq 0, j = 1,2$ and represent marginal contribution due to the dependence. We see in these terms that the dependence effect $(\rho^2\beta_1^2\beta_2^2\sigma_1^2\sigma_2^2)$ is counted two times in comparison with the total variance. Concerning the independent and full total Sobol' indices, the interaction effect $(\beta_1^2\beta_2^2\sigma_1^2\sigma_2^2)$ of these indices is still allocated as half in Shapley effects. Besides, for the full total Sobol indices, each term is equal to the variance model, whereas the interaction and dependence effects are equally distributed for the Shapley effects which sum up to the total variance.

This example supports the idea mentioned in Iooss and Prieur [2017] whereby *a full Sobol index of an input comprises the effect of another input on which it is dependent*. We can add that, whether the model is independent or not, the phenomenon is similar for the interaction effect about the independent and full total Sobol indices of an input, i.e. these indices comprise the effect of another input on which the input is interacting.

To clarify the objectives of a SA study, Saltelli and Tarantola [2002] and Saltelli et al. [2004] defined the SA settings:

- Factors Prioratization (FP) Setting, to know on which inputs the reduction of uncertainty leads to the largest reduction of the output uncertainty;
- Factors Fixing (FF) Setting, to assess which inputs can be fixed at given values without any loss of information in the model output;
- Variance Cutting (VC) Setting, to know which inputs have to be fixed to obtain a target value on the output variance;
- Factors Mapping (FM) Setting, to determine the inputs mostly responsible for producing realizations of Y in a given region.

In their article, Iooss and Prieur [2017] tell at which goals of the SA settings the four (full and independent) Sobol' indices as well as the Shapley effects apply.

According to them, a combined interpretation of the four Sobol indices would just allow to do the FP (Factor prioritization) setting. But we can add that these indices allow also to do the FF (Factor Fixing) setting only if a factor has both indices ST_i^{ind} and ST_i^{full} which are null. Indeed, if $ST_i^{ind} = \frac{\mathbb{E}[\mathrm{Var}(Y|\mathbf{X}_{\sim i})]}{\mathrm{Var}(Y)} = 0$ and $ST_i^{full} = \frac{\mathbb{E}[\mathrm{Var}(Y|(\mathbf{X}_{\sim i}|X_i))]}{\mathrm{Var}(Y)} = 0$ and as the variance is always a positive function, that implies $\mathrm{Var}(Y|\mathbf{X}_{\sim i}) = 0$ and $\mathrm{Var}\left(Y|\left(\mathbf{X}_{\sim i}|X_i\right)\right) = 0$. Thus, Y can be expressed only as a function of $\mathbf{X}_{\sim i}$ or $\mathbf{X}_{\sim i}|X_i$, where $\mathbf{X}_{\sim i}|X_i$ represent all components except X_i not taking account the dependence with the variable X_i.

About the Shapley effects, they would allow to do the VC (Variance Cutting) setting as the sum is equal to $\mathrm{Var}(Y)$ and the FF setting. Sure enough, if $Sh^i = 0$, then we have $\forall \mathcal{J} \subseteq \mathcal{D}\backslash\{i\}, \mathrm{Var}\left[Y|\mathbf{X}_{-(\mathcal{J}\cup\{i\})}\right] = \mathrm{Var}\left[Y|\mathbf{X}_{-\mathcal{J}}\right]$ and so express Y as a function of $\mathbf{X}_{-(\mathcal{J}\cup\{i\})}$ equates to express Y as a function of $\mathbf{X}_{-\mathcal{J}}$. Hence, X_i is not an influential input in the model and can be fixed. The FP setting is not achieved according to them because of the fair distribution of the interaction and dependence effects in the indice. However, this share allocation makes the Shapley effects easier to interpret than the Sobol' indices and might be a great alternative to the four Sobol' indices. Thus, in the sequel, we will compare the Sobol indices' and the Shapley effects on a basic example to see if they make correctly the factor prioritization.

4.4.2 Linear model with three inputs

Let us consider

$$Y = \beta_0 + \boldsymbol{\beta}^\intercal \mathbf{X} \tag{4.30}$$

with the constants $\beta_0 \in \mathbb{R}$, $\boldsymbol{\beta} \in \mathbb{R}^3$ and $\mathbf{X} \sim \mathcal{N}(0, \Sigma)$ with the following covariance matrix :

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \alpha\sigma_1\sigma_2 & \rho\sigma_1\sigma_3 \\ \alpha\sigma_1\sigma_2 & \sigma_2^2 & \gamma\sigma_2\sigma_3 \\ \rho\sigma_1\sigma_3 & \gamma\sigma_2\sigma_3 & \sigma_3^2 \end{pmatrix}, -1 \leq \alpha, \rho, \gamma \leq 1, \sigma_1 > 0, \sigma_2 > 0, \sigma_3 > 0.$$

We obtain the following analytical results.

$$\sigma^2 = Var(Y) = \beta_1^2\sigma_1^2 + \beta_2^2\sigma_2^2 + \beta_3^2\sigma_3^2 + 2\gamma\beta_2\beta_3\sigma_2\sigma_3 + 2\beta_1\sigma_1(\alpha\beta_2\sigma_2 + \rho\beta_3\sigma_3)$$

- For $j = 1, 2, 3$, from the definition of independent Sobol indices, we have:

$$\begin{aligned}
\sigma^2 S_1^{ind} = \sigma^2 ST_1^{ind} &= \frac{\beta_1^2\sigma_1^2\left(-1+\alpha^2+\gamma^2+\rho^2-2\alpha\gamma\rho\right)}{\gamma^2-1} \\
\sigma^2 S_2^{ind} = \sigma^2 ST_2^{ind} &= \frac{\beta_2^2\sigma_2^2\left(-1+\alpha^2+\gamma^2+\rho^2-2\alpha\gamma\rho\right)}{\rho^2-1} \\
\sigma^2 S_3^{ind} = \sigma^2 ST_3^{ind} &= \frac{\beta_3^2\sigma_3^2\left(-1+\alpha^2+\gamma^2+\rho^2-2\alpha\gamma\rho\right)}{\alpha^2-1}
\end{aligned}$$

- For $j = 1, 2, 3$, from the definition of full Sobol indices, we have:

$$\begin{aligned}
\sigma^2 S_1^{full} = \sigma^2 ST_1^{full} &= (\beta_1\sigma_1 + \alpha\beta_2\sigma_2 + \rho\beta_3\sigma_3)^2 \\
\sigma^2 S_2^{full} = \sigma^2 ST_2^{full} &= (\alpha\beta_1\sigma_1 + \beta_2\sigma_2 + \gamma\beta_3\sigma_3)^2 \\
\sigma^2 S_3^{full} = \sigma^2 ST_3^{full} &= (\rho\beta_1\sigma_1 + \gamma\beta_2\sigma_2 + \beta_3\sigma_3)^2
\end{aligned}$$

In both cases, full and independent Sobol indices, the first order index is equal to the total order index because the model is linear, i.e., there is no interaction between the inputs.

- For $j = 1, 2, 3$, **in this example** we obtain the following decomposition for the Shapley effects :

$$\begin{aligned}
Sh^1 &= \frac{1}{3}\left(S_1^{full} + \frac{1}{2}ST_{1|2} + \frac{1}{2}ST_{1|3} + ST_1^{ind}\right) \\
Sh^2 &= \frac{1}{3}\left(S_2^{full} + \frac{1}{2}ST_{2|1} + \frac{1}{2}ST_{2|3} + ST_2^{ind}\right) \\
Sh^3 &= \frac{1}{3}\left(S_3^{full} + \frac{1}{2}ST_{3|1} + \frac{1}{2}ST_{3|2} + ST_3^{ind}\right)
\end{aligned}$$

So, for the **linear Gaussian model** we found a relation between the Shapley effects and the sensitivity indices obtained with the RT method. For more details about the calculation of Shapley effects, we refer the readers to the Appendix 4.9.1.
About the results, as the formula is similar regardless the input, we analyse it with the first input. We observe that the Shapley effect Sh^1 is in some way the average of all possible contributions of the input X_1 in the model. Indeed, S_1^{full} represents the full marginal contribution of X_1. Then, we have the total contributions of X_1 without its correlative contribution with each element of the set $\mathcal{D} = \{1, 2, 3\}\backslash\{1\} = \{2, 3\}$. Sure enough, $ST_{1|2}$ is the total contribution of X_1 without its correlative contribution with X_2, i.e. ones just look at the total effect with its dependence with X_3 ; $ST_{1|3}$ is the total contribution of X_1 without its correlative contribution with X_3, i.e. ones just look at the total effect with its dependence with X_2 and finally the uncorrelated total contribution of X_1 via $ST_1^{ind} = ST_{1|2,3}$. As in $\{2, 3\}$, there are two elements of size one, we find the coefficients $1/2$ before $ST_{1|2}$ and $ST_{1|3}$ and 1 for ST_1^{ind}. We then find the fair allocation of the Shapley effects.

Particular cases

Now, we will consider several particular cases of correlation in order to compare the prioritization obtained with the Sobol' indices and the Shapley effects. We will take in the following examples $\beta_0 = 0$; $\beta_1 = \beta_2 = \beta_3 = 1$ and $\sigma_1 = 0.2, \sigma_2 = 0.6, \sigma_3 = 1$. By making this choice, we define implicitly the most influential variables and we want to observe how the correlation affects the indices. Besides, for each considered case, we verify that the covariance matrix is positive definite.

	$\alpha = \rho = \gamma = 0$			$\alpha = \rho = \gamma = 0.5$			$\alpha = \rho = 0.75, \gamma = 0.15$		
	X_1	X_2	X_3	X_1	X_2	X_3	X_1	X_2	X_3
S_i^{ind}	0.0286	0.2571	0.7143	0.0115	0.1034	0.2874	0.0004	0.0085	0.0236
ST_i^{ind}	0.0286	0.2571	0.7143	0.0115	0.1034	0.2874	0.0004	0.0085	0.0236
S_i^{full}	0.0286	0.2571	0.7143	0.4310	0.6207	0.8448	0.9515	0.3932	0.7464
ST_i^{full}	0.0286	0.2571	0.7143	0.4310	0.6207	0.8448	0.9515	0.3932	0.7464
Sh_i	0.0286	0.2571	0.7143	0.1715	0.3123	0.5163	0.4553	0.1803	0.3644

Table 4.2 *Sensitivity indices of linear model with different configurations of correlation*

As part of the independent linear model, the Shapley effects are equal to the Sobol' indices as proved in Iooss and Prieur [2017] and thus, all the indices carry out to the same ranking of the inputs.

In the second configuration with the symmetric case, we remark a decrease of the independent Sobol indices and an increase of the full Sobol indices with respect to the independent model ($\alpha = \rho = \gamma = 0$). As regards of the Shapley effects, it reduces for the third input, raises slightly for the second input and significantly for the first input. All these changes are due to the mutualization of uncertainties within the inputs because of the correlation but the individual contributions of the inputs are still well captured for all the indices. Indeed, in spite of the correlation, all the indices indicate the same ranking for the inputs.

In this last configuration, we have strongly correlated a non-influential variable (X_1 has a low variance) in the model with two very influential variables. The independent Sobol' indices give us as ranking: X_3, X_2, X_1. However, as the values of these indices are close to zero, we can suppose they are not significant and implicitly the ranking neither. We obtain with the full indices the following ranking X_1, X_3, X_2. X_1 is supposed to be a non-influential variable and turns out to explain 95% of the model variance. Which is logical because being highly correlated with X_2 and X_3, X_1 has a strong impact on these variables. Then, X_2 and X_3 are correlated in the same way with X_1 and weakly between them. Regardless of the correlations, X_3 is more influential than X_2 in the model, hence this second position taking account the correlation. Lastly, we obtain the same ranking as the full Sobol' indices with the Shapley effects. FP (Factors Prioritization) setting aims to find which factors would allow to have the largest expected reduction in the variance of the model output. Thus, if we follow the previous ranking, we should reduce the uncertainty on the first input. But we will make several tests by reducing the uncertainty of 20% one by one on each input and we get:

Setting $\alpha = \rho = 0.75, \gamma = 0.15$	Model variance
$\sigma_1 = 0.2, \sigma_2 = 0.6, \sigma_3 = 1$	2.06
$\sigma_1 = 0.16, \sigma_2 = 0.6, \sigma_3 = 1$	1.95
$\sigma_1 = 0.2, \sigma_2 = 0.48, \sigma_3 = 1$	1.86
$\sigma_1 = 0.2, \sigma_2 = 0.6, \sigma_3 = 0.8$	1.60

Table 4.3 *Model variance by reducing the uncertainty on each input one by one*

It is clearly observed that the largest expected reduction in the variance is obtained with the third input. These results conflict the obtained ranking with the full Sobol indices and the Shapley effects. Indeed, X_1 is an influential input only because of the strong correlation with X_2 and X_3, and these indices capture this trend. However, without this correlation, X_1 becomes a non-influential input and the independent Sobol indices are supposed to highlight **meaningfully** that the inputs X_2 and X_3 are the most influential without taking account the correlation between the inputs. Nevertheless, these indices hardly account for the uncorrelated marginal contributions of these inputs due to the small values we obtain.

Thus, on this basic example we can see that the combined interpretation of the four Sobol indices as well as the Shapley effects doesn't allow to answer correctly to the purpose of the Factor Prioritization (FP) setting, i.e. on which inputs the reduction of uncertainty leads to the largest reduction of the output uncertainty. We can make a factor prioritization with these indices but not for the goal defined at the outset.

4.5 Numerical studies

Optimal values for the parameters of the exact and random permutation methods were given by Song et al. [2016]. Using a toy example, we empirically study how the algorithm settings can influence the estimation of the indices. We compare the accuracy of the estimations of the Sobol' indices obtained from the Shapley algorithm or from the RT method.

4.5.1 Parametrization of the Shapley algorithms

As defined in Section 4.3.2, three parameters of the Shapley algorithm govern the estimation accuracy: N_v, N_o and N_i. The first one, is the sample-size for the output variance estimation of Y. The second, is the number of outer loop for the sample-size of the expectation and the third one is the number of inner loop which controls the sample-size for the variance estimation of each conditioned distribution.

Theses variances are estimated through Monte Carlo procedures. The output variance $\mathrm{Var}[Y]$ is computed from a sample $\{Y_j = \eta(\mathbf{X}^{(j)})\}_{j=1\ldots,N_v}$. Because N_v is a small proportion of the overall cost $C = N_v + m(d-1)N_oN_i$, especially when the d is large, we can select N_v as large as

possible in order to reach the smallest possible estimation error of $\mathrm{Var}[Y]$. However, it is more difficult to chose N_o and N_i to estimate the conditioned variances. These choices also depend on the used algorithm: exact or random permutations.

Therefore, we empirically show the influence of N_o and N_i on the estimation error and the coverage probability. The Probability Of Coverage (POC) is defined as the probability to have the true indice value inside the confidence intervals of the estimation. We consider the three dimensional linear Gaussian model of Section 4.4.2 as a toy example with independent inputs, $\beta_1 = \beta_2 = \beta_3 = 1$, $\sigma_1 = \sigma_2 = 1$ and $\sigma_3 = 2$. The POC is estimated with 100 independent algorithm runs and for a 90 % confidence interval. When the bootstrap procedure is considered, the confidence intervals are estimated with 500 bootstrap sampling. We also set a large value of $N_v = 10000$ for all the experiments.

First experiments aim to show the influence of N_o on the estimation accuracy and the POC for the exact permutation algorithm. The Figure 4.1 shows the variation of the POC (solid lines) and the absolute error (dashed lines), averaged over the three indices, in function of the product $N_o N_i m$, where only N_o is varying and for three values of N_i at 3, 9 and 18. Because the errors are computed for 100 independent runs, we show in color areas the 95% quantiles.

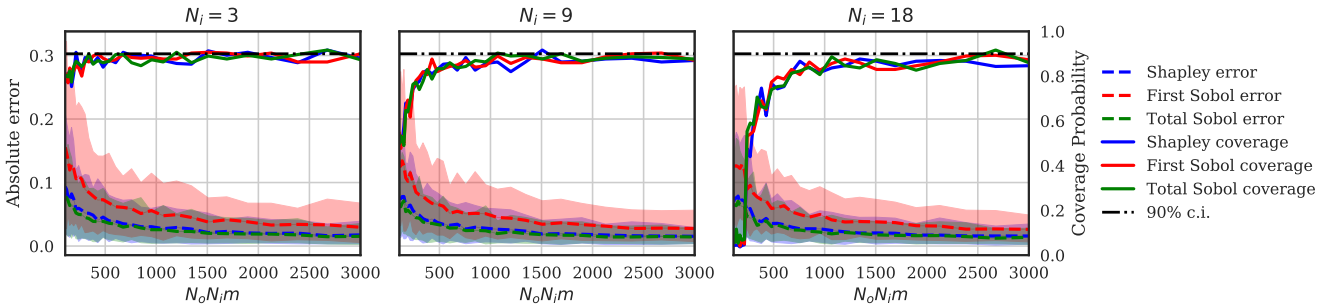


Figure 4.1 *Variation of the absolute error and the POC with N_o for three values of $N_i = 3, 9, 18$ for the exact permutation algorithm ($m = d! = 6$).*

We observe that the estimation error is similar for the three different values of N_i and decrease to 0 at the same rate. The true difference is for the POC which tends, at different rates, to the true probability: 90 %. For a same computational cost $N_o N_i m$, the smaller the value of N_i and the larger the value of N_o. Thus, these results show that, in order to have a correct confidence interval it is more important to have a large value of N_o instead of N_i. Indeed, exploring multiple conditioned variances with a lower precision (large N_o and low N_i) is more important than having less conditioned variances with a good precision (low N_o and large N_i).

The Figure 4.2 is similar to Figure 4.1 but for the random permutation algorithm and by fixing $N_o = 1$ and by varying the number of permutations.

As for the exact permutation algorithm, we can see that the estimation errors are similar for the three values of N_i and the difference is shown for the POC. We observe that the lower N_i and the faster the POC converges to the true probability. Indeed, for a same computational cost,

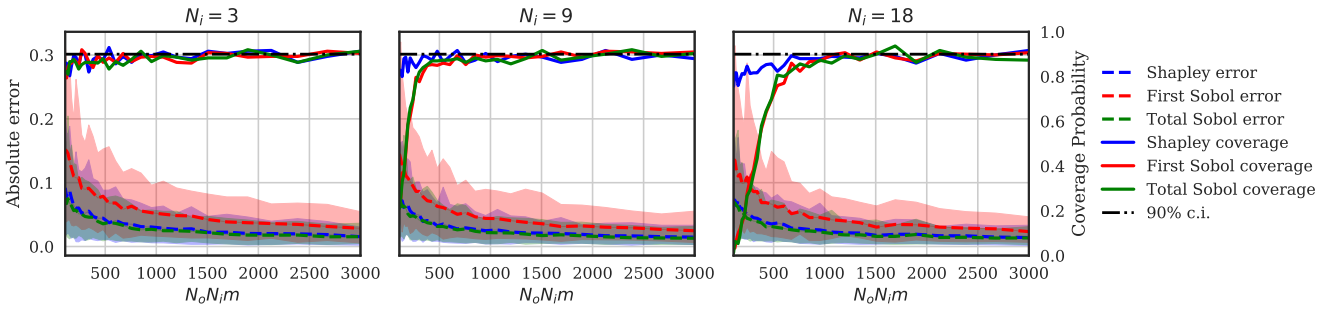


Figure 4.2 *Variation of the absolute error and the POC with m for three values of $N_i = 3, 9, 18$ and $N_o = 1$ for the random permutation algorithm.*

the lower N_i and the larger the number of permutations m can be.

To show the influence of N_o with the random permutation algorithm, the Figure 4.3 is the same as Figure 4.2 but with $N_o = 3$. We observe that the convergence rates of the POC are slower than the ones for $N_o = 1$. Thus, it shows that having a lower value of N_o and a large value of m is more important to have consistent confidence intervals.

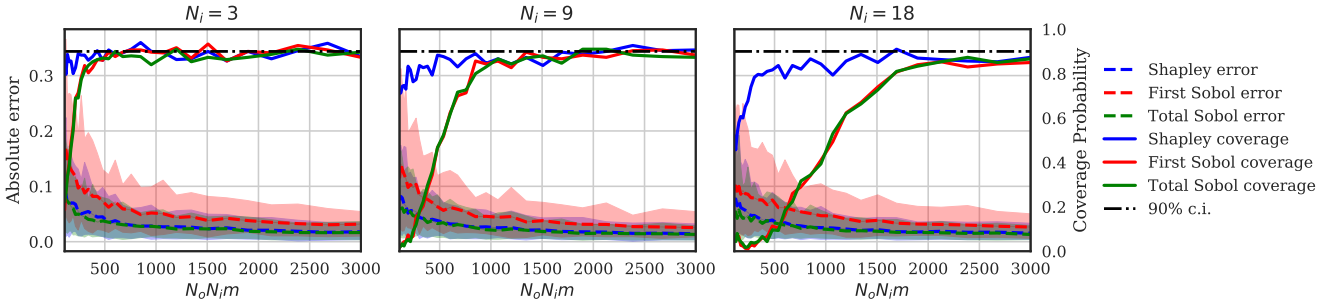


Figure 4.3 *Variation of the absolute error and the POC with m for three values of $N_i = 3, 9, 18$ and $N_o = 3$ for the random permutation algorithm.*

From these experiments, we can conclude that the parametrization does not significantly influence the estimation error but has a strong influence on the POC. Moreover, these experiments were established on different toy examples (Ishigami model defined in Section 4.7.2 and interactive model) and the same conclusion arises. Therefore, in order to have consistent confidence intervals, we can suggest:

- for the exact algorithm to consider $N_i = 3$ and to take N_o as large as possible,
- for the random permutation algorithm to consider $N_i = 3$, $N_o = 1$ and take m as large as possible.

This conclusion confirms the proposed parametrization of Song et al. [2016] explained in 4.3.2

and the suggestion analyzed in Iooss and Prieur [2017].

4.5.2 Minor bias observed

At the start of this section, we chose to establish these experiments for independent random variables. This choice was justified by unexpected results obtained for correlated variables. The Figure 4.4 illustrates the same experiment as Figure 4.1 but with a correlation of $\gamma = 0.9$ between X_2 and X_3. We observed that the POC of the total Sobol' indice starts to tend to the true probability (at 90%) before slowly decreasing. Thus, it seems that the confidence intervals are underestimated or the indice estimation is biased.

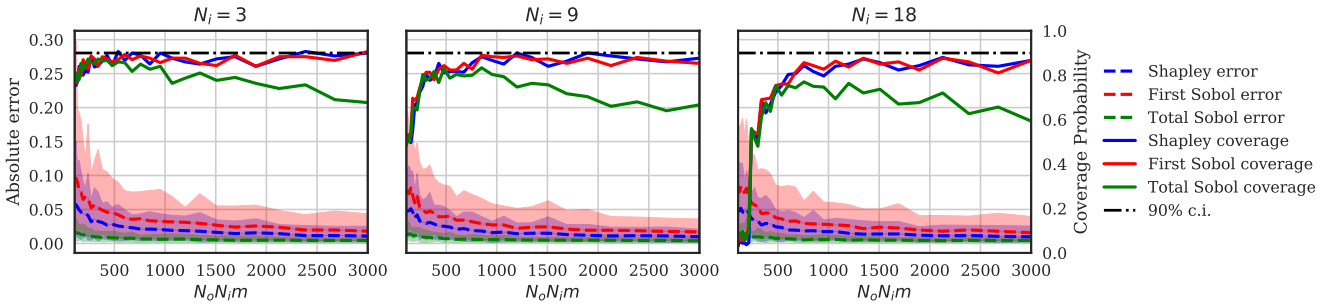


Figure 4.4 *Variation of the absolute error and the POC with N_o for three values of $N_i = 3, 9, 18$ for the exact permutation algorithm and a correlation $\gamma = 0.9$ between X_2 and X_3.*

To verify this observation, Figure 4.5 shows the estimation of the total Sobol' indice for $N_v = 20000$, $N_o = 10000$, $N_i = 3$ with the histogram from the bootstrap sampling in blue, the estimated indice ST_i in red line and the true indice in green line. It is clear that the true value for X_2 and X_3 is outside of estimated distribution. This explains why the coverage probability is decreasing in Figure 4.4. Moreover, this phenomenon only happens to the indices of X_2 and X_3, which are correlated and it seems that this bias increases with the correlation strength for this example. Therefore, the reasons of this slight bias should be investigated in future works

4.5.3 Comparing Sobol' index estimation using Shapley algorithm and RT method

An interesting result of the Shapley algorithm is that it gives the full first-order Sobol' indices and the independent total Sobol' indices in addition to the Shapley effects. We compare the estimation accuracy of the Sobol' indices obtained from the Shapley algorithm and the ones from the RT method. We consider the same example as in Section 4.5.1 but with dependent random variables. In this section, only the pair X_2-X_3 is correlated with parameter γ.

A first experiment aims to validate the confidence intervals estimated from the bootstrap sampling of the RT method by doing the same experiments as in Section 4.5.1 by increasing the

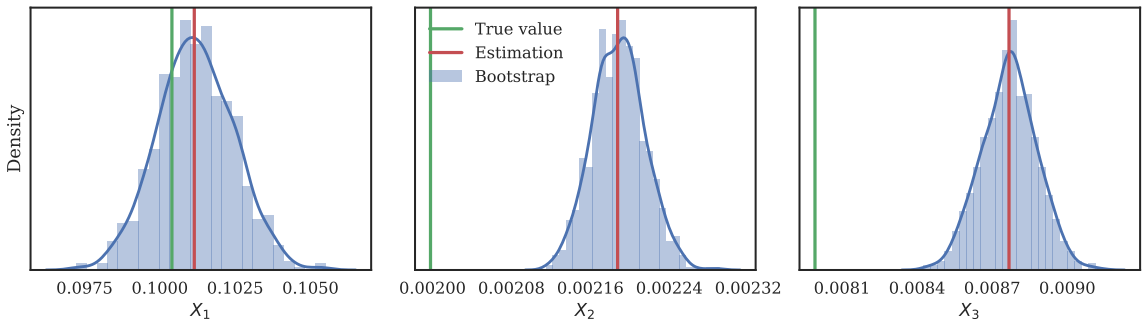


Figure 4.5 *Estimated bootstrap estimations of the total Sobol' indices from the exact Shapley algorithm with a correlation of 0.99 between X_2 and X_3.*

sample-size N. The Figure 4.6 shows the absolute error and the POC with the computational cost $(4 \times N \times d)$ for the full first-order Sobol' indices and the independent total Sobol' indices for $\gamma = 0.5$. As we can see the error tends to zero and the POC converges quickly to the true probability. Thus, we can see that the confidence intervals correctly catch the Monte Carlo error.

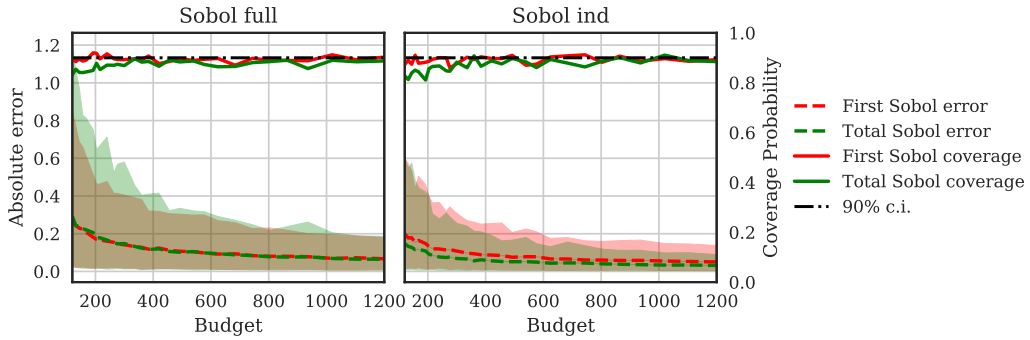


Figure 4.6 *Variation of the absolute error and the POC with the computational cost for the RT method.*

We recall from Section 4.3 that the full first-order Sobol' indices are equivalent to the classical first-order Sobol' indices and the independent total indices are the classical total indices. The Figure 4.7 shows the estimated indices with $\gamma = 0.5$ from the Shapley algorithm and the RT method for similar computational costs. We observe that both algorithms seem to correctly estimate the Sobol' indices for a low computational cost. However, in this example, the estimation errors from the RT method is much larger than the ones from the Shapley algorithm.

We recall from Section 4.2.3 that RT method used the Janon estimator from Janon et al. [2014]. The accuracy of the Sobol' estimator depends on the values of the target indices and the Janon estimator is less accurate for low value indices. Changing with another estimator, such as the one from Mara et al. [2015], can lead to another estimation variance as shown in Figure 4.8.

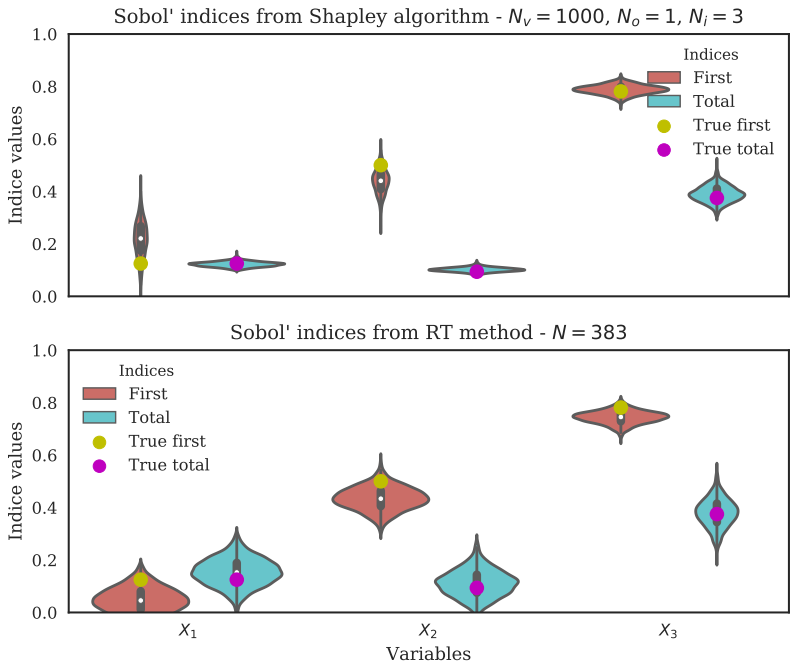


Figure 4.7 *Sobol' indices estimation from the exact permutation method of the Shapley algorithm (top) and the RT method (bottom) using the Janon estimator for similar number of evaluation:* $N_v + N_o N_i m(d-1) = 4Nd = 4600$.

We observed that the estimation errors from the RT method depends of the used estimator and this error is lower using estimator from Figure 4.8 than the one from Figure 4.7.

The Figure 4.9 shows the Sobol' indices for the exact Shapley algorithm and the RT method in function of the correlation γ between X_2 and X_3. The lines shows the true values of the indices and the areas are the 95% confidence intervals of the indices.

This experiment shows that the estimation of the Sobol' indices from the Shapley algorithm gives satisfactory estimations of the first full and total ind Sobol' indices. Note that the error of estimation is similar for both the exact or random permutation algorithm if we consider the same computational cost.

4.6 Kriging metamodel with inclusion of errors

Shapley effects are a suitable tool for performing global sensitivity analysis. However, their estimates require an important number of simulations of the costly function $\eta(\mathbf{x})$ and often cannot be processed under reasonable time constraint. To handle this problem, we use $\tilde{\eta}(\mathbf{x})$ an approximating function of the numerical model under study $\eta(\mathbf{x})$ [Fang et al., 2005]. Its main advantage is obvioulsy to be much faster-to-calculate than the original one. In addition, if one

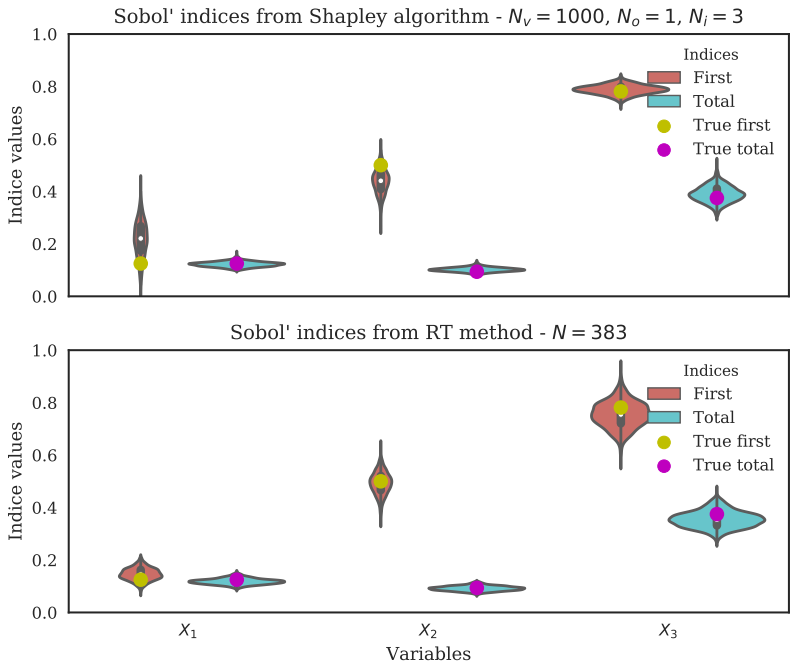


Figure 4.8 *Sobol' indices estimation from the exact permutation method of the Shapley algorithm (top) and the RT method (bottom) using the estimator from Mara et al. [2015] for similar number of evaluation:* $N_v + N_o N_i m(d-1) = 4Nd = 4600$.

uses a kriging method [Sacks et al., 1989] to build this $\tilde{\eta}(\mathbf{x})$ surrogate model, a quantification of the approximation uncertainty can be easily produced. The Shapley effects can then be calculated using the metamodel $\tilde{\eta}(\mathbf{x})$ instead of $\eta(\mathbf{x})$ with a control on the estimation error.

We present in this section a methodology for estimating the Shapley effects through a kriging surrogate model taking into account both the Monte Carlo error and the surrogate model error.

4.6.1 Introduction to the kriging model

Kriging, also called *metamodeling by Gaussian process*, is a method consisting in the use of an emulator of a costly computer code for which the interpolated values are modeled by a Gaussian process. More precisely, it is based on the assumption that the $\eta(x)$ function is the realization of a Gaussian random process. The data is then used to infer characteristics of this process, allowing a joint modelization of the code itself and the uncertainty about the interpolation on the domain. In general, one assumes a particular parametric model for the mean function of the process and for its covariance. The parameters of these two functions are called "hyperparameters" and are estimated using the data. The Gaussian hypothesis then provides an explicit formula for the law of the process conditionaly to the value taken by η on a design of experiments $\mathbf{D}$.

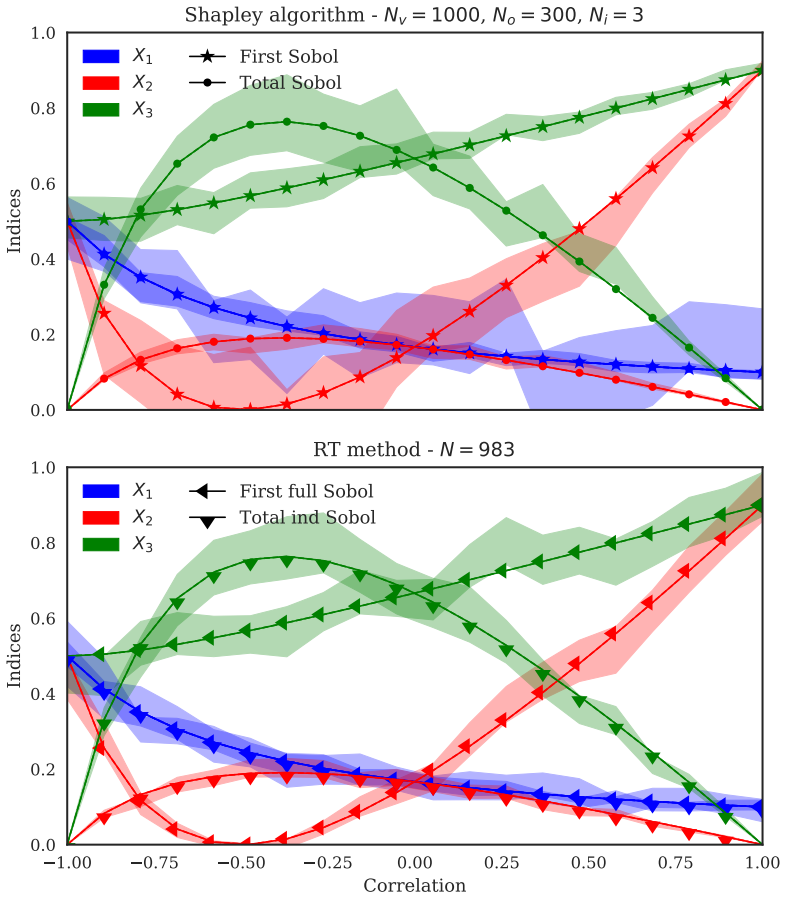


Figure 4.9 *Sobol' indices estimations from the exact permutation method of the Shapley algorithm (top) and the RT method (bottom) in fonction of γ.*

Thus, we consider that our expensive function $\eta(x)$ can be modeled by a Gaussian process $H(x)$ which's mean and variance are such that $\mathbb{E}[H(x)] = \mathbf{f}'(x)\boldsymbol{\beta}$ and $Cov(H(x), H(\tilde{x})) = \sigma^2 r(x, \tilde{x})$, where $r(x, \tilde{x})$ is the covariance kernel (or the correlation function) of the process.

Then, $\eta(x)$ can be easily approximated by the conditional Gaussian process $H_n(x)$ having the predictive distribution $[H(x)|H(\mathbf{D}) = \boldsymbol{\eta}^n, \sigma^2]$ where $\boldsymbol{\eta}^n$ are the known values of $\eta(x)$ at points in the experimental design set $\mathbf{D} = \{x^1, \ldots, x^n\}$ and σ^2 is the variance parameter. Therefore, we have

$$H_n(x) \sim GP\left(m_n(x), s_n^2(x, \tilde{x})\right), \tag{4.31}$$

where the mean $m_n(x)$ is given by

$$m_n(x) = \mathbf{f}'(x)\widehat{\boldsymbol{\beta}} + \mathbf{r}'(x)\mathbf{R}^{-1}\left(\boldsymbol{\eta}^n - \mathbf{f}\widehat{\boldsymbol{\beta}}\right),$$

where $\mathbf{R} = [r(x_i, x_j)]_{i,j=1,\ldots,n}$, $\mathbf{r}'(x) = [r(x, x_i)]_{i=1,\ldots,n}$, $\mathbf{f} = [\mathbf{f}'(x_i)]_{i=1,\ldots,n}$, and

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{f}'\mathbf{R}^{-1}\mathbf{f}\right)^{-1}\mathbf{f}'\mathbf{R}^{-1}\boldsymbol{\eta}^n.$$

The variance $s_n^2(x, \tilde{x})$ is given by

$$s_n^2(x,\tilde{x}) = \sigma^2 \left(1 - \begin{pmatrix} \mathbf{f}'(x) & \mathbf{r}'(x) \end{pmatrix} \begin{pmatrix} 0 & \mathbf{f}' \\ \mathbf{f} & \mathbf{R} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{f}(\tilde{x}) \\ \mathbf{r}(\tilde{x}) \end{pmatrix} \right)$$

The variance parameter σ^2 can be estimated with a restricted maximum likelihood method.

4.6.2 Kriging based Shapley effects and estimation

Inspired by the idea used in Le Gratiet et al. [2014] for the Sobol indices, we substitute the true function $\eta(\mathbf{x})$ with $H_n(\mathbf{x})$ in (4.24) which leads to

$$Sh_n^i = \frac{1}{d!} \sum_{\pi \in \Pi(\mathcal{D})} \left[c_n \left(P_i(\pi) \cup \{i\} \right) - c_n \left(P_i(\pi) \right) \right] \tag{4.32}$$

where the exact function $Y = \eta(\mathbf{X})$ is replaced by the Gaussian process $H_n(\mathbf{X})$ in the cost function such as $c_n(\mathcal{J}) = \mathbb{E}\left[\mathrm{Var}\left[H_n(\mathbf{X}) | \mathbf{X}_{-\mathcal{J}} \right] \right]$.

Therefore, if we denote by $(\Omega_H, \mathcal{F}_H, \mathbb{P}_H)$ the probability space where the Gaussian process $H(x)$ lies, then the index Sh_n^i lies in $(\Omega_H, \mathcal{H}, \mathbb{P}_H)$ (it is hence random).

Then, for estimating Sh_n^i, we use the same estimator (4.26) developed by Song et al. [2016] in which we remplace Y by the Gaussian process $H_n(\mathbf{X})$ in the cost function to obtain :

$$\widehat{Sh}_n^i = \frac{1}{m} \sum_{l=1}^{m} \left[\widehat{c}_n \left(P_i(\pi_l) \cup \{i\} \right) - \widehat{c}_n \left(P_i(\pi_l) \right) \right] \tag{4.33}$$

where $\widehat{c}_n$ is the Monte Carlo estimator of c_n.

4.6.3 Estimation of errors : Monte Carlo and surrogate model

The estimator (4.33) above integrates two sources of uncertainty : the first one is related to the metamodel approximation, and the second one is related to the Monte Carlo integration. So, in this part, we quantify both by decomposing the variance of $\widehat{Sh}_n^i$ as follows :

$$\mathrm{Var}(\widehat{Sh}_n^i) = \mathrm{Var}_H \left(\mathbb{E}_X \left[\widehat{Sh}_n^i | H_n(x) \right] \right) + \mathrm{Var}_X \left(\mathbb{E}_H \left[\widehat{Sh}_n^i | \left(\mathbf{X}_{\kappa_l} \right)_{l=1,\dots,B} \right] \right)$$

where $\mathrm{Var}_H \left(\mathbb{E}_X \left[\widehat{Sh}_n^i | H_n(x) \right] \right)$ is the contribution of the metamodel on the variability of $\widehat{Sh}_n^i$ and $\mathrm{Var}_X \left(\mathbb{E}_H \left[\widehat{Sh}_n^i | \left(\mathbf{X}_{\kappa_l} \right)_{l=1,\dots,B} \right] \right)$ is that of the Monte Carlo integration.

In section 4 of the article Le Gratiet et al. [2014], they proposed the algorithm (5) we adapted here to estimate each of these contributions.

Algorithme 5 : Evaluation of the distribution of $\widehat{Sh}^i_{\kappa,n}$

Build $H_n(x)$ from the n observations $\boldsymbol{\eta}^n$ of $\eta(x)$ at points in $\mathbf{D}$;
Generate a sample $\mathbf{x}^{(1)}$ of size N_v from the random vector $\mathbf{X}$;
Generate a sample $\mathbf{x}^{(2)}$ of size $m(d-1)N_oN_i$ from the different conditional laws necessary to estimate $\mathbb{E}\left[\mathrm{Var}\left[Y|\mathbf{X}_{-\mathcal{J}}\right]\right]$;
Set N_H as the number of samples for $H_n(x)$ and B the number of bootstrap samples for evaluating the uncertainty due to Monte Carlo integration ;
for $k = 1, \ldots, N_H$ **do**
    Sample a realization $\{\mathbf{y}^{(1)}, \mathbf{y}^{(2)}\} = \eta_n(\mathbf{x})$ of $H_n(\mathbf{x})$ with $\mathbf{x} = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}\}$;
    Compute $\widehat{Sh}^i_{n,k,1}$ thanks to (4.32) from $\eta_n(\mathbf{x})$;
    for $l = 2, \ldots, B$ **do**
        Sample with replacement a realization $\tilde{\mathbf{y}}^{(1)}$ of $\mathbf{y}^{(1)}$ to compute $\mathrm{Var}(Y)$;
        Sample by bloc with replacement a realization $\tilde{\mathbf{y}}^{(2)}$ of $\mathbf{y}^{(2)}$;
        Compute $\widehat{Sh}^i_{n,k,l}$ thanks to the equation (4.32) from $\{\tilde{\mathbf{y}}^{(1)}, \tilde{\mathbf{y}}^{(2)}\}$;
return $\left(\widehat{Sh}^i_{n,k,l}\right)_{\substack{k=1,\ldots,N_H \\ l=1,\ldots,B}}$

The output $\left(\widehat{Sh}^i_{n,k,l}\right)_{\substack{k=1,\ldots,N_H \\ l=1,\ldots,B}}$ of the algorithm (5) is a sample of size $N_H \times B$ representative of the distribution of $\widehat{Sh}^i_n$ and takes into account both the uncertainty of the metamodel and that of the Monte Carlo integration.

From this algorithm and some theoretical results, Le Gratiet et al. [2014] proposed estimators in section 4.2 to estimate each of these contributions.

4.7 Numerical simulations with kriging model

This section aims at estimating the studied indices using a surrogate model in substitution of the true and costly computational code. The previous section explained the theory behind the Gaussian processes to emulate a function. The Section 4.6.3 explained that the kriging error can be estimating through a large number of realization of the Gaussian Process in addition to the Monte Carlo error estimated through a bootstrap sampling. In this section, we illustrate the decomposition of the overall error from the estimation of the indices and we consider as examples the additive Gaussian framework and the Ishigami function.

4.7.1 Gaussian framework

We use the same configuration as in the Section 4.5.3 with a correlation coefficient $\rho = 0.7$. To illustrate the influence of the kriging model in the estimation of the indices, we show in Figure 4.10 the distribution of the estimators of the indices with the procedure using the true function (top figure) and using the surrogate function (bottom figure). We took $N_v = 1000, N_o = 100$ and $N_i = 3$ for the two graphics.

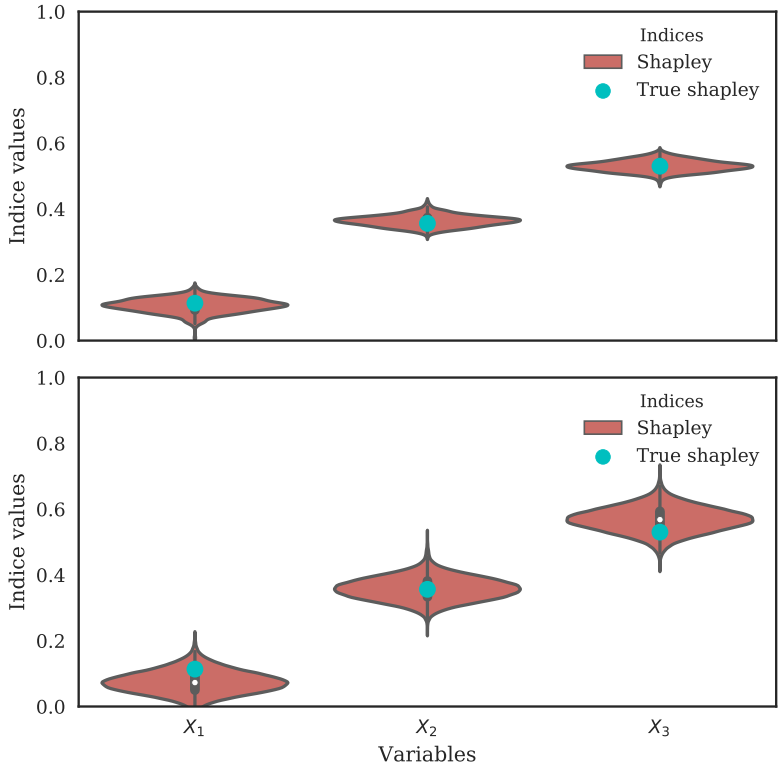


Figure 4.10 *Estimation of the Shapley effects with the exact permutation algorithm. The top and bottom figures respectively show the estimation results with the true function and the kriging model with* $Q^2 = 0.90$.

The kriging model is built with 10 points using a LHS sampling (at independence) and a Matern kernel with a linear basis, leading to a Q^2 of 0.90 and the kriging error is estimated with $N_H = 300$ realizations. We intentionally took low values for the algorithm parameters in order to have a relatively high variance. If we compare the violinplots of the two figures, we observe that the variance of the estimation is larger for the kriging configuration. This is due to the additional error from the kriging model. The Figure 4.11 allows to distinguish which part the overall error is due to the kriging. We see immediately what the kriging error is larger than the Monte Carlo error and it is normal that this error feeds through to the quality of the estimations as observed in Figure 4.10.

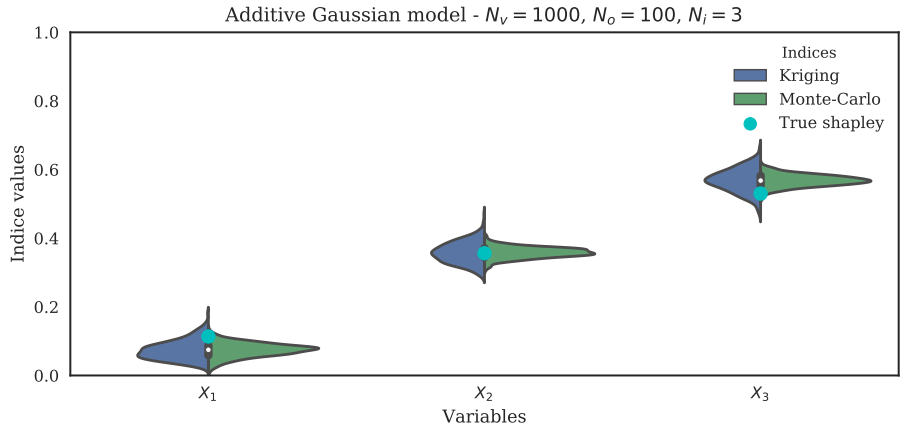


Figure 4.11 *Separation of the uncertainty from the Monte Carlo estimation and the kriging model approximation.*

4.7.2 Ishigami Function

Introduced in Ishigami and Homma [1990], the Ishigami function is typically used as a benchmarking function for uncertainty and sensitivity analysis. It is interesting because it exhibits a strong non-linearity and has interactions between variables. For any variable $\mathbf{x} = (x_1, x_2, x_3) \in [-\pi, \pi]^3$, the model function can be written as

$$\eta(\mathbf{x}) = \sin(x_1) + 7\sin^2(x_2) + 0.1x_3^4 \sin(x_1). \tag{4.34}$$

In this example, we consider that the random variable $\mathbf{X}$ follows a distribution $p_{\mathbf{X}}$ with uniform margins $\mathcal{U}[-\pi, \pi]$ and a multivariate Gaussian copula $C_{\boldsymbol{\rho}}$ with parameter $\boldsymbol{\rho} = (\rho_{12}, \rho_{13}, \rho_{23})$. Thanks to the Sklar Theorem [Sklar, 1959], the multivariate cumulative distribution function F of $\mathbf{X}$ can be written as

$$F(x_1, x_2, x_3) = C_{\boldsymbol{\rho}}\left(F_1(x_1), F_2(x_2), F_3(x_3)\right) \tag{4.35}$$

where F_1, F_2, F_3 are the marginal cumulative distribution functions of $\mathbf{X}$. In the independent case, analytical full first order and independent total Sobol' indices are derived as well as the Shapley effects. Unfortunately, no analytical results are available for the other indices. Thus, we place in the sequel in the independent framework.

Remind that the main advantage of the metamodel is to be much faster-to-calculate than the original function. Thus, we can use this characteristic in order to decrease the Monte Carlo error during the estimation of the indices by increasing the calculation budget.

In this example, the kriging model is built with 200 points using an optimized LHS sampling (at independence) and a Matern kernel with a linear basis, leading to a Q^2 of 0.98 and the kriging error will be estimated subsequently with $N_H = 300$ realizations.
To illustrate the influence of the kriging model in the estimation of the indices, we show in Figure 4.12 the distribution of the estimators of the indices obtained with the true function (top

figure) for $N_v = 1000, N_o = 100, N_i = 3$ and using the surrogate function (bottom figure) with $N_v = 5000, N_o = 600$ and $N_i = 3$. We intentionally took high values for the estimation with the metamodel in order to decrease the overall variance.

If we compare the violinplots of the two figures, we observe that the variance of the estimations is higher with the true function. For the true function, the uncertainty is only due to the Monte Carlo estimation. For the surrogate function, as observed in Figure 4.13, in spite of a slight metamodel error, this same Monte Carlo is obviously lower owing to a higher calculation budget. Hence, if the metamodel approximates correctly the true function, it is better to use it to estimate the sensitivity indices to gain accuracy on the distribution of the estimators.

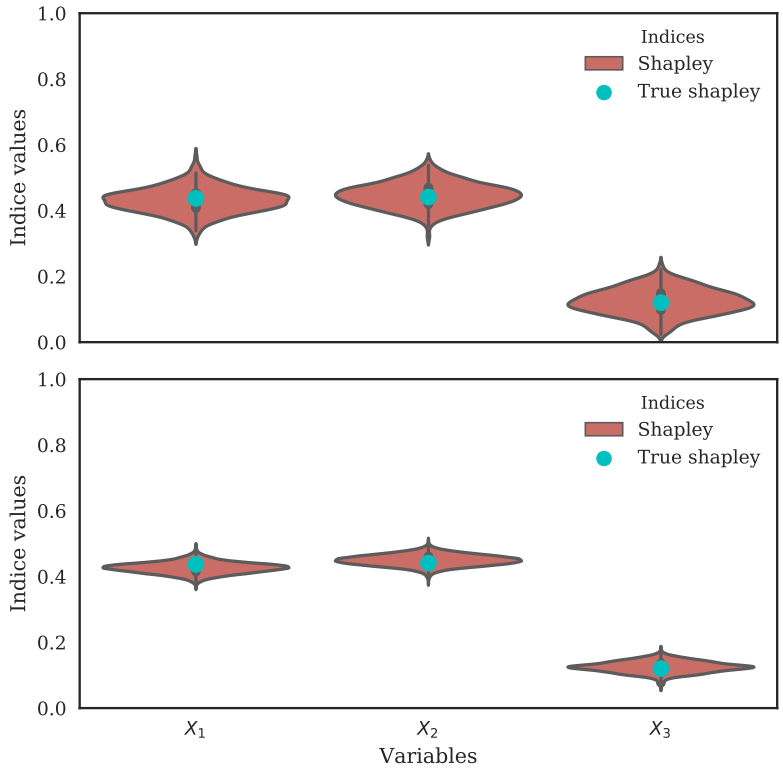


Figure 4.12 *Estimation of the Shapley effects with the exact permutation algorithm. The top and bottom figures respectively show the estimation results with the true function and the kriging model with* $Q^2 = 0.98$.

4.8 Conclusion

Throughout this article, we studied the Shapley effects and the *independent* and *full* Sobol' indices defined in Mara and Tarantola [2012] for the models with a dependence structure on the input variables. The comparison between these indices revealed that

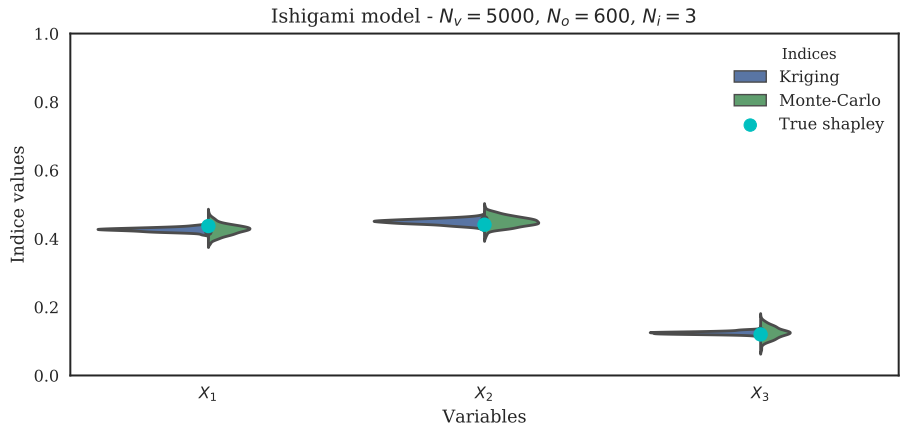


Figure 4.13 *Separation of the uncertainty from the Monte Carlo estimation and the kriging model approximation.*

- the full Sobol' index of an input includes the effect of another input on which it is dependent,
- the independent and full total Sobol' indices of an input includes the effect of another input on which it is interacting,
- the Shapley effects rationally allocate these different contributions for each input.

Each of these indices allows to answer certain objectives of the SA settings defined in Saltelli and Tarantola [2002] and Saltelli et al. [2004]. But, it is important to pay attention about the FP setting. This one can be made with the Shapley effects but not for the goal defined at the outset, i.e. prioritize the input variables taking account the dependence but not to find which would allow to have the largest expected reduction in the variance of the model output. Always about the FP setting, it was declared in conclusion of our example that the combined interpretation of the four Sobol' indices doesn't allow to answer correctly to the purpose of the FP setting due to the small values that have been obtained for the *independent* Sobol' indices. However, although these values were close to zero, the ranking that they had provided was correct to make FP setting. Hence, it could be investigated whether these values are significant or not.

A relation between the Shapley effects and the Sobol' indices obtained from the RT method was found for the linear Gaussian model. It would be interesting to see if this relation could be extended to a general linear model in the first instance and subsequently if a overall relation can be established between these indices for a global model.

About the estimation procedure of the Shapley effects, a major contribution of this article is the implementation of a bootstrap sampling to estimate the Monte Carlo error. The CLT can give confidence intervals but require large sample size in order to be consistent, which is rarely possible in practice for expensive computer codes. We confirmed that the parametrization of the Shapley algorithms proposed by Song et al. [2016] and analyzed by Iooss and Prieur [2017] is correct and optimal in order to have consistent confidence intervals. The numerical comparison of the Sobol' indices estimated from the Shapley algorithm and the RT method for a toy example showed that the estimations from the Shapley algorithm are a bit less accurate than the ones from

the RT method, but are very satisfying for an algorithm that is not design for their estimation.

A second contribution is the splitting of the metamodel and Monte Carlo errors when using a kriging model to substitute the true model. The numerical results showed that for a reasonable number of evaluations of a kriging model, one can estimate the Shapley effects, as well as the Sobol' indices and still correctly catch estimation error due to the metamodel or the Monte Carlo sampling. Unfortunately, the computational cost to generate a sample from a Gaussian Process realization increases significantly with the sample-size. Thus, because the Shapley algorithm becomes extremely costly in high dimension, the estimation of indices using this technique can be computationally difficult.

The Shapley algorithm from Song et al. [2016] is efficient, but is extremely costly in high dimension. The cost is mainly due to the estimation of the conditional variances. A valuable improvement of the algorithm would be the use of a Kernel estimation procedure in order to significantly reduce the number of evaluation. The Polynomial Chaos Expension are good to compute the Sobol' indices analytically from the polynomial coefficients [Crestaux et al., 2009]. It would be interesting to have such a decomposition for the Shapley effects.

Acknowledgments

The authors express their thanks to Gerard Biau (Sorbonne Université), Bertrand Iooss and Roman Sueur (EDF R&D) for their useful comments and advice during the writing of this article.

4.9 Appendix

4.9.1 Gaussian framework: linear model

Let us consider

$$Y = \beta_0 + \boldsymbol{\beta}^\intercal \mathbf{X} \tag{4.36}$$

with the constants $\beta_0 \in \mathbb{R}$, $\boldsymbol{\beta} \in \mathbb{R}^3$ and $\mathbf{X} \sim \mathcal{N}(0, \Sigma)$ with the following covariance matrix :

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \alpha\sigma_1\sigma_2 & \rho\sigma_1\sigma_3 \\ \alpha\sigma_1\sigma_2 & \sigma_2^2 & \gamma\sigma_2\sigma_3 \\ \rho\sigma_1\sigma_3 & \rho\sigma_2\sigma_3 & \sigma_3^2 \end{pmatrix}, -1 \leq \alpha, \rho, \gamma \leq 1, \sigma_1 > 0, \sigma_2 > 0, \sigma_3 > 0.$$

We obtained the following analytical results.

$$\sigma^2 = Var(Y) = \beta_1^2\sigma_1^2 + \beta_2^2\sigma_2^2 + \beta_3^2\sigma_3^2 + 2\gamma\beta_2\beta_3\sigma_2\sigma_3 + 2\beta_1\sigma_1(\alpha\beta_2\sigma_2 + \rho\beta_3\sigma_3)$$

- For $j = 1, 2, 3$, from the definition of full Sobol indices, we have:

$$\begin{aligned}
\sigma^2 S_1^{full} &= \sigma^2 ST_1^{full} = (\beta_1\sigma_1 + \alpha\beta_2\sigma_2 + \rho\beta_3\sigma_3)^2 \\
\sigma^2 S_2^{full} &= \sigma^2 ST_2^{full} = (\alpha\beta_1\sigma_1 + \beta_2\sigma_2 + \gamma\beta_3\sigma_3)^2 \\
\sigma^2 S_3^{full} &= \sigma^2 ST_3^{full} = (\rho\beta_1\sigma_1 + \gamma\beta_2\sigma_2 + \beta_3\sigma_3)^2
\end{aligned}$$

- We calculate also the full first order Sobol indices for the others subsets of $\mathcal{D}$ and we have :

$$\begin{aligned}
\sigma^2 S_{1,2}^{full} &= \beta_1^2\sigma_1^2 + \beta_2^2\sigma_2^2 + 2\gamma\beta_2\beta_3\sigma_2\sigma_3 + 2\beta_1\sigma_1(\alpha\beta_2\sigma_2 + \rho\beta_3\sigma_3) - \frac{\beta_3^2\sigma_3^2\left(\gamma^2 + \rho^2 - 2\alpha\gamma\rho\right)}{\alpha^2 - 1} \\
\sigma^2 S_{1,3}^{full} &= \beta_1^2\sigma_1^2 + \beta_3^2\sigma_3^2 + 2\gamma\beta_2\beta_3\sigma_2\sigma_3 + 2\beta_1\sigma_1(\alpha\beta_2\sigma_2 + \rho\beta_3\sigma_3) - \frac{\beta_2^2\sigma_2^2\left(\alpha^2 + \gamma^2 - 2\alpha\gamma\rho\right)}{\rho^2 - 1} \\
\sigma^2 S_{2,3}^{full} &= \beta_2^2\sigma_2^2 + \beta_3^2\sigma_3^2 + 2\gamma\beta_2\beta_3\sigma_2\sigma_3 + 2\beta_1\sigma_1\left(\alpha\beta_2\sigma_2 + \rho\beta_3\sigma_3\right) - \frac{\beta_1^2\sigma_1^2\left(\alpha^2 + \rho^2 - 2\alpha\gamma\rho\right)}{\gamma^2 - 1} \\
\sigma^2 S_{\mathcal{D}}^{full} &= \sigma^2
\end{aligned}$$

- We calculate also the total Sobol indices for the variables $(X_i|X_u), i = 1, \ldots, 3$ and $u \subset \mathcal{D}\backslash\{i\}, u \neq \emptyset$ and we have :

$$\sigma^2 ST_{1|2} = -\frac{\left(\beta_1\sigma_1\left(\alpha^2 - 1\right) + \beta_3\sigma_3(\alpha\gamma - \rho)\right)^2}{\alpha^2 - 1} \quad \sigma^2 ST_{1|3} = -\frac{\left(\beta_1\sigma_1\left(\rho^2 - 1\right) + \beta_2\sigma_2(\gamma\rho - \alpha)\right)^2}{\rho^2 - 1}$$

$$\sigma^2 ST_{2|1} = -\frac{\left(\beta_2\sigma_2\left(\alpha^2 - 1\right) + \beta_3\sigma_3(\alpha\rho - \gamma)\right)^2}{\alpha^2 - 1} \quad \sigma^2 ST_{2|3} = -\frac{\left(\beta_2\sigma_2\left(\gamma^2 - 1\right) + \beta_1\sigma_1(\gamma\rho - \alpha)\right)^2}{\gamma^2 - 1}$$

$$\sigma^2 ST_{3|1} = -\frac{\left(\beta_3\sigma_3\left(\rho^2 - 1\right) + \beta_2\sigma_2(\alpha\rho - \gamma)\right)^2}{\rho^2 - 1} \quad \sigma^2 ST_{3|2} = -\frac{\left(\beta_3\sigma_3\left(\gamma^2 - 1\right) + \beta_1\sigma_1(\alpha\gamma - \rho)\right)^2}{\gamma^2 - 1}$$

- For $j = 1, 2, 3$, from the definition of Shapley effects, we have:

$$\begin{aligned}
Sh_1 &= \frac{1}{3}\left((\tilde{c}(1) - \tilde{c}(\emptyset)) + \frac{1}{2}\left(\tilde{c}(1,2) - \tilde{c}(2)\right) + \frac{1}{2}\left(\tilde{c}(1,3) - \tilde{c}(3)\right) + \left(\tilde{c}(1,2,3) - \tilde{c}(2,3)\right)\right) \\
&= \frac{1}{3}\left(S_1^{full} + \frac{1}{2}\left(S_{1,2}^{full} - S_2^{full}\right) + \frac{1}{2}\left(S_{1,3}^{full} - S_3^{full}\right) + \left(S_{1,2,3}^{full} - S_{2,3}^{full}\right)\right) \\
&= \frac{1}{3}\left(S_1^{full} + \frac{1}{2}ST_{1|2} + \frac{1}{2}ST_{1|3} + ST_1^{ind}\right)
\end{aligned}$$

$$\begin{aligned}
Sh_2 &= \frac{1}{3}\left((\tilde{c}(2) - \tilde{c}(\emptyset)) + \frac{1}{2}\left(\tilde{c}(1,2) - \tilde{c}(1)\right) + \frac{1}{2}\left(\tilde{c}(2,3) - \tilde{c}(3)\right) + \left(\tilde{c}(1,2,3) - \tilde{c}(1,3)\right)\right) \\
&= \frac{1}{3}\left(S_2^{full} + \frac{1}{2}\left(S_{1,2}^{full} - S_1^{full}\right) + \frac{1}{2}\left(S_{2,3}^{full} - S_3^{full}\right) + \left(S_{1,2,3}^{full} - S_{1,3}^{full}\right)\right) \\
&= \frac{1}{3}\left(S_2^{full} + \frac{1}{2}ST_{2|1} + \frac{1}{2}ST_{2|3} + ST_2^{ind}\right)
\end{aligned}$$

$$
\begin{aligned}
Sh_3 &= \frac{1}{3}\left((\tilde{c}(3)-\tilde{c}(\emptyset))+\frac{1}{2}\left(\tilde{c}(1,3)-\tilde{c}(1)\right)+\frac{1}{2}\left(\tilde{c}(2,3)-\tilde{c}(2)\right)+\left(\tilde{c}(1,2,3)-\tilde{c}(1,2)\right)\right)\\
&= \frac{1}{3}\left(S_3^{full}+\frac{1}{2}\left(S_{1,3}^{full}-S_1^{full}\right)+\frac{1}{2}\left(S_{2,3}^{full}-S_2^{full}\right)+\left(S_{1,2,3}^{full}-S_{1,2}^{full}\right)\right)\\
&= \frac{1}{3}\left(S_3^{full}+\frac{1}{2}ST_{3|1}+\frac{1}{2}ST_{3|2}+ST_3^{ind}\right)
\end{aligned}
$$

Chapter 5

Detecting and modeling worst-case dependence structures between random inputs of computational reliability models

Abstract. Uncertain information on input parameters of reliability models is usually modeled by considering these parameters as random, and described by marginal distributions and a dependence structure of these variables. In numerous real-world applications, while information is mainly provided by marginal distributions, typically from samples, little is really known on the dependence structure itself. Faced with this problem of incomplete or missing information, risk studies are often conducted by considering independence of input variables, at the risk of including irrelevant situations. This approach is especially used when reliability functions are considered as black-box computational models. Such analyses remain weakened in absence of in-depth model exploration, at the possible price of a strong risk misestimation. Considering the frequent case where the reliability output is a quantile, this article provides a methodology to improve risk assessment, by exploring a set of pessimistic dependencies using a copula-based strategy. In dimension greater than two, a greedy algorithm is provided to build input regular vine copulas reaching a minimum quantile to which a reliability admissible limit value can be compared, by selecting pairwise components of sensitive influence on the result. The strategy is tested over toy models and a real industrial case-study. The results highlight that current approaches can provide non-conservative results, and that a nontrivial dependence structure can be exhibited to define a worst-case scenario.

Contents

5.1 Introduction 112
5.2 Minimization of the quantile of the output distribution 114
5.3 A preliminary study of the copula influence on quantile minimization 120
5.4 Quantile minimization and choice of penalized correlation structure 124
5.5 Applications 131
5.6 Conclusion and discussion 137
5.7 Proof of the consistency result 138

5.1 Introduction

Many industrial companies, like energy producers or vehicle and aircraft manufacturers, have to ensure a high level of safety for their facilities or products. In each case, the structural reliability of certain so-called critical components plays an essential role in overall safety. For reasons related to the fact that these critical components are highly reliable, and that real robustness tests can be very expensive or even hardly feasible, structural reliability studies generally use simulation tools [de Rocquigny et al., 2008; Lemaire et al., 2010]. The physical phenomenon of interest being reproduced by a numerical model η (roughly speaking, a *computer code*), such studies are based on the calculation of a reliability indicator based on the comparison of $y = \eta(\mathbf{x})$ and a safety margin, where $\mathbf{x}$ corresponds to a set of input parameters influencing the risk. In the framework of this article, such models are considered as black box and can be explored only by simulation means.

While the problems of checking the validity of η and selecting inputs $\mathbf{x} \in \chi \subseteq \mathbb{R}^d$ are addressed by an increasing methodological corpus [Bayarri et al., 2007; Council, 2012], a perennial issue is the modeling of $\mathbf{x}$. Differing from the specification of η itself, this input vector is known with uncertainty, either because the number of experiments to estimate is limited, or because some inputs reflect intrinsically variable phenomena [Nilsen and Aven, 2003]. In most cases, these epistemic and aleatory uncertainties are jointly modeled by probability distributions [Helton, 2011]. Consecutively, the reliability indicator is often defined as the probability that y be lower than a threshold (*failure probability*), or a limit quantile for y. This article focuses on this last indicator, which provides an upper or lower bound of the mean effect of the output variable uncertainty.

Therefore the modeling of $\mathbf{x}$ stands on the assessment of a joint probability distribution with support χ, divided between marginal and dependencies features. Though information on each dimension of $\mathbf{x}$ can often be accessible experimentally or using physical or expert knowledge [Bedford et al., 2006], the dependence structure between the component of $\mathbf{x}$ remains generally unknown. Typically, statistical data are only available per dimension, but not available for two or more dimensions simultaneously. For this reason, most of robustness studies are conducted by sampling within independent marginal distributions. Doing so, reliability engineers try to

capture input situations that minimize the reliability indicator. Such situations are defined as so-called *worst cases*. However, the assumption of independence between inputs has been severely criticized since the works by Grigoriu and Turkstra [1979] and Thoft-Christensen and Sørensen [1982], who showed that output failure probabilities of industrial systems can significantly vary and be underestimated if the input dependencies are neglected. More generally, Tang et al. [2013, 2015] showed that tail dependencies between inputs can have major expected effects on the uncertainty analysis results.

Returning to a probabilist framework, and beyond structural reliability, the problem of defining a worst-case scenario by selecting a joint input distribution, from incomplete information, is a topical issue encountered in many fields. In decision-making problems, Scarf et al. [1958] proposed a general definition of the worst case distribution as the minimizer of an excepted cost among a set of possible distributions. More recently, Agrawal et al. [2012] extended this approach to account for incomplete dependence information. These theoretical works, that propose selection rules over the infinite set of all possible joint distributions, remain hard to apply in practice. Recent applied works made use of copulas [Nelsen, 2007] to model dependencies between stochastic inputs [Tang et al., 2013, 2015], following other researchers confronted to similar problems in various fields: finance [Cherubini et al., 2004], structural safety [Goda, 2010], environmental sciences [Schoelzel and Friederichs, 2008] or medicine [Beaudoin and Lakhal-Chaieb, 2008]. These studies mainly consider bivariate copulas, which makes theses analysis effective only when two random variables are correlated. Cases where a greater number of variables is involved were explored by Jiang et al. [2015], who used *vine copulas* to approach complex multidimensional correlation problems in structural reliability. A vine copula is a graphical representation of the *pair-copula construction* (PCC), proposed by Joe [1996], which defines a multidimensional dependence structure using conditional bivariate copulas. Various class of vines exist (see Czado [2010] for a review), and among them the regular vines (R-vines) introduced by Bedford and Cooke [2001, 2002] are known for their appealing computational properties, while inference on PCC is usually demanding [Dissmann et al., 2013; Haff, rway].

R-vine parametric copulas seem promising to improve the search for a worst-case dependence between stochastic inputs, while keeping the benefits of a small number of parameters, as favoring inference and conducting simple sensitivity analyses *a posteriori*. To our knowledge, however, no practical methodology has been yet proposed to this end for which the notion of worst case is defined by the minimization of an output quantile. This is the subject of this article. More precisely, the aim of this research is to determine a parametric copula over $\mathbf{x}$, close to the worst case dependence structure, which is associated to a minimum value of the quantile of the distribution of y. Given a vine structure defined by a parameter vector, the optimization problem involves to conduct empirical quantile estimations for each value of this vector in a finite set of interest (chosen as a grid). The proposed methodology stands on an encompassing greedy algorithm exploring copula structures, which integrates several sub-algorithms of increasing complexity and is based on some simplifying assumptions. These algorithms are made available in the Python library `dep-impact` [Benoumechiara, 2018].

The article is therefore organized as follows. Section 5.2 introduces the framework and studies the consistency of a statistical estimation of the minimum quantile, given an input copula family and a growing sequence of grids. A preliminary study of the influence of the dependence

structure, specific to quantile minimization, is conducted in Section 5.3 as a first application of this statistical optimization. The wider problem of selecting copulas in high-dimensional settings using a sequence of quantile minimization is considered in Section 5.4. While the choice of R-vines is defended, a sparsity hypothesis is made to diminish the computational burden, according to which only a limited number of pairwise dependencies is influent on the result. A greedy algorithm is proposed to carry out the complete procedure of optimization and modeling. This heuristic is tested in Section 5.5 over toy examples, using simulation, and a real industrial case-study. The results highlight that worst-case scenarios produced by this algorithm are often bivariate copulas reaching the Fréchet-Hoeffding bounds [Hoeffding, 1940; Frechet, 1951] (describing perfect dependence between variables), as it could be expected in monotonic frameworks, but that other nontrivial copulas can be exhibited in alternative situations. Results and avenues for future research are extensively discussed in the last section of this article. We also refer to Appendix 5.7 and 5.7.3 for supplementary material on consistency proofs, on R-vine copulas and on R-vine iterative construction.

5.2 Minimization of the quantile of the output distribution

This section introduces a general framework for the calculation of the minimum quantile of the output distribution of a computational model, when the input distribution can be taken from a large family of distributions, each one corresponding to a particular choice of dependencies between the input variables.

5.2.1 A general framework for the computation of the minimum quantile

To be general, let us consider a computer code which takes a vector $\mathbf{x} \in \chi \subseteq \mathbb{R}^d$ as an input and produces a real quantity y in output. This code is represented by a deterministic function $\eta : \mathbb{R}^d \to \mathbb{R}$ such that $\eta(\mathbf{x}) = y$. The sets $\mathbb{R}$ and $\mathbb{R}^d$ are endowed with their Borel sigma algebras and we assume that η is measurable. The general expression of the function η is unknown but for some vector $\mathbf{x} \in \mathbb{R}^d$ it is assumed that the quantity $\eta(\mathbf{x})$ can always be computed. In particular, the derivatives of η, when they exist, are never assumed to be known. Let $P_1, \ldots, P_d$ be a fixed family of d distributions, all supported on $\mathbb{R}$. We introduce the set $\mathcal{D}(P_1, \ldots, P_d)$ of all multivariate distributions P on $\mathbb{R}^d$ such that the marginal distributions of P are all equal to the $(P_j)_{j=1\ldots d}$. Henceforth, we use the shorter notation $\mathcal{D}$ for $\mathcal{D}(P_1, \ldots, P_d)$.

For some $P \in \mathcal{D}$, let G be the cumulative distribution function of the model output. In other terms dG is the push-forward measure of P by η. For $\alpha \in (0, 1)$, let G^{-1} be the α-quantile of the output distribution:

$$G^{-1}(\alpha) := \inf\{y \in \mathbb{R} : G(y) \geq \alpha\}. \tag{5.1}$$

For the rest of this document, we denote as *output quantile* the α-quantile of the output distribution.

In many real situations, the function η corresponds to a known physical phenomenon. The input variables $\mathbf{x}$ of the model are subject to uncertainties and are quantified by the distribution

P. The propagation of these uncertainties leads to the calculation of the output quantile, which defines an overall risk. Due to the difficulties to gather information, it is common to have this distribution incompletely defined and only known through its marginal distributions. Therefore, the set $\mathcal{D}$ corresponds to all the possible distributions that are only known through their marginal distributions $(P_j)_{j=1\dots d}$. In a reliability study, it is essential to avoid underestimating the risk. In such a situation, we might consider a more pessimistic computation of the quantile. We define as the *worst quantile*, the minimum value of the quantile by considering all the possible input distributions $P \in \mathcal{D}$. This conservative approach consists in minimizing $G^{-1}(\alpha)$ over the family $\mathcal{D}$ such as

$$G^{-1\star}(\alpha) := \min_{P\in\mathcal{D}} G^{-1}(\alpha). \tag{5.2}$$

Since the function η has no closed form in general, it is not possible to give a simple expression of $G^{-1}(\alpha)$ in function of the distribution P, and consequently the minimum $G^{-1\star}(\alpha)$ does not have a simple expression too. In this chapter we propose to study a simpler problem than (5.2), by minimizing $G^{-1}(\alpha)$ over a subset of $\mathcal{D}$. This subset is a family of distributions $(P_{\boldsymbol{\theta}})_{\boldsymbol{\theta}\in\mathcal{D}}$ associated to a parametric family of copula $(C_{\boldsymbol{\theta}})_{\boldsymbol{\theta}\in\mathcal{D}}$, where $\boldsymbol{\Theta}$ is a compact set of $\mathbb{R}^p$ and p is the number of copula parameters.

5.2.2 Copula-based approach

We introduce the real-values random vector $\mathbf{X} = (X_1, \dots, X_d) \in \mathbb{R}^d$ associated to the distribution $P_{\boldsymbol{\theta}}$. Each component X_j, for $j = 1, \dots, d$, is a real-value random variable with distribution P_j. A copula describes the dependence structure between a group of random variables. Formally, a copula is a multidimensional continuous cumulative distribution function (CDF) linking the margins of $\mathbf{X}$ to its joint distribution. Sklar's Theorem [Sklar, 1959] states that every joint distribution $F_{\boldsymbol{\theta}}$ associated to the measure $P_{\boldsymbol{\theta}}$ can be written as

$$F_{\boldsymbol{\theta}}(\mathbf{x}) = C_{\boldsymbol{\theta}}\left(F_1(x_1), \dots, F_d(x_d)\right), \tag{5.3}$$

with some appropriate d-dimensional copula $C_{\boldsymbol{\theta}}$ with parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and the marginal CDF's $F_j(x_j) = \mathbb{P}[X_j \leq x_j]$. If all marginal distributions are continuous functions, then there exists an unique copula satisfying

$$C_{\boldsymbol{\theta}}(u_1, \dots, u_d) = F_{\boldsymbol{\theta}}(F_1^{-1}(u_1), \dots, F_d^{-1}(u_d))$$

where $u_j = F_j(x_j)$. For $F_{\boldsymbol{\theta}}$ absolutely continuous with strictly increasing marginal distributions, one can derive (5.3) to obtain the joint density of $\mathbf{X}$:

$$f_{\boldsymbol{\theta}}(\mathbf{x}) = c_{\boldsymbol{\theta}}\left(F_1(x_1), \dots, F_d(x_d)\right) \prod_{j=1}^{d} f_j(x_j), \tag{5.4}$$

where $c_{\boldsymbol{\theta}}$ denotes the copula density function of $C_{\boldsymbol{\theta}}$ and $f_j(x_j)$ are the marginal densities of $\mathbf{X}$. Numerous parametric copula families are available and are based on different dependence structures. Most of these families have bidimensional dependencies, but some can be extended to higher dimensions. However, these extensions have a lack of flexibility and cannot describe all types of dependencies [Nelsen, 2007]. To overcome these difficulties, tools like vine copulas [Joe,

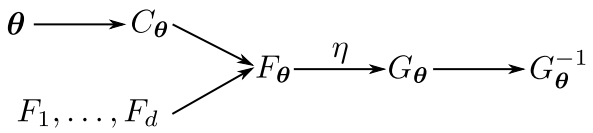


Figure 5.1 *Illustration of the link between the dependence parameter $\boldsymbol{\theta}$ and the quantile function $G_{\boldsymbol{\theta}}^{-1}$. The joint CDF $F_{\boldsymbol{\theta}}$ is obtained using (5.3) from a copula $C_{\boldsymbol{\theta}}$ and marginal CDF's $(F_j)_{j=1}^d$. The push-forward of $F_{\boldsymbol{\theta}}$ through the model η leads to the CDF $G_{\boldsymbol{\theta}}$ and quantile function $G_{\boldsymbol{\theta}}^{-1}$ of the output distribution.*

1994] (described in Section 5.4) combine bivariate copulas, from different families, to create a multidimensional copula.

Let $G_{\boldsymbol{\theta}}$ and $G_{\boldsymbol{\theta}}^{-1}$ be respectively the CDF and quantile function of the push-forward distribution of $P_{\boldsymbol{\theta}}$ by η (see Figure 5.1). For a given parametric family of copula $(C_{\boldsymbol{\theta}})_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}$ and a given $\alpha \in (0,1)$, the minimum output quantile for a given copula is defined by

$$G_C^{-1\star}(\alpha) := \inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} G_{\boldsymbol{\theta}}^{-1}(\alpha) \tag{5.5}$$

and if it exists, we consider a minimum

$$\boldsymbol{\theta}_C^* \in \operatorname*{argmin}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} G_{\boldsymbol{\theta}}^{-1}(\alpha). \tag{5.6}$$

We call this quantity the minimum quantile parameter or worst dependence structure.

Note that there is no reason for $G_{\boldsymbol{\theta}}^{-1}(\alpha)$ to be a convex function of $\boldsymbol{\theta}$. The use of gradient descent algorithms is thus not straightforward in this context. Moreover, the gradient of $\boldsymbol{\theta} \to G_{\boldsymbol{\theta}}^{-1}$ is unknown and only zero-order optimization methods can be applied to solve (5.6). For this reason, in the following of this section, we analyze the basic approach which consists in estimating $\boldsymbol{\theta}_C^*$ by approximating $\boldsymbol{\Theta}$ with a finite regular grid $\boldsymbol{\Theta}_N$ of cardinality N. Therefore, for a given parametric copula $(C_{\boldsymbol{\theta}})_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}$ and a given $\alpha \in (0,1)$, we restrict the problem (5.6) to

$$\boldsymbol{\theta}_N^* \in \operatorname*{argmin}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_N} G_{\boldsymbol{\theta}}^{-1}(\alpha). \tag{5.7}$$

5.2.3 Estimation with a grid search strategy

In the restricted problem (5.7), the greater N, the closer $\boldsymbol{\theta}_N^*$ to the minimum $\boldsymbol{\theta}^*$ of $\boldsymbol{\Theta}$; obviously the convergence rate should depend on the regularity of the function η and on the regularity of the quantile function $\boldsymbol{\theta} \mapsto G_{\boldsymbol{\theta}}^{-1}(\alpha)$. Because η has no closed form, the quantile function $G_{\boldsymbol{\theta}}^{-1}(\alpha)$ has no explicit expression. The minimizer $\boldsymbol{\theta}_N^*$ can be estimated by coupling the simulation of independent and identically distributed (i.i.d) data $(Y_1, \ldots, Y_n)$, defined as realizations of the model output random variable $Y := \eta(\mathbf{X})$ with distribution $\mathrm{d}G_{\boldsymbol{\theta}}$, with a minimization of the empirical quantile over $\boldsymbol{\Theta}_N$.

For $\boldsymbol{\theta}$ taking a value over the grid $\boldsymbol{\Theta}_N$, the empirical CDF of Y is defined for any $y \in \mathbb{R}$ by

$$\widehat{G}_{\boldsymbol{\theta}}(y) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{Y_i \leq y}. \tag{5.8}$$

The corresponding empirical quantile function $\widehat{G}_{\boldsymbol{\theta}}^{-1}(\alpha)$ is defined as in (5.1) by replacing G with its empirical estimate. For a given α, the worst quantile on the fixed grid $\boldsymbol{\Theta}_N$ is given by

$$\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_N} G_{\boldsymbol{\theta}}^{-1}(\alpha).$$

and can be estimated by replacing the quantile function with its empirical function:

$$\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_N} \widehat{G}_{\boldsymbol{\theta}}^{-1}(\alpha). \tag{5.9}$$

Finally the estimation of the minimum quantile parameter over the grid $\boldsymbol{\Theta}_N$ is denoted by

$$\widehat{\boldsymbol{\theta}}_N = \operatorname*{argmin}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_N} \widehat{G}_{\boldsymbol{\theta}}^{-1}(\alpha). \tag{5.10}$$

The construction of the grid $\boldsymbol{\Theta}_N$ can be difficult because $\boldsymbol{\Theta}$ can be unbounded (e.g. $\boldsymbol{\Theta} = [1, \infty]$ for a Gumbel copula). To tackle this issue, we chose to construct $\boldsymbol{\Theta}_N$ among a normalized space using a concordance measure, which is bounded in $[-1, 1]$ and does not rely on the marginal distributions. We chose the commonly used Kendall rank correlation coefficient (or Kendall's tau) [Kendall, 1938] as a concordance measure to create this transitory space. This non-linear coefficient $\tau \in [-1, 1]$ is related to the copula function as follows:

$$\tau = 4\int_{-1}^{1}\int_{-1}^{1} C_\theta(u_1, u_2)\mathrm{d}C(u_1, u_2) - 1.$$

For many copula families, this relation is much more explicit (see for instance Frees and Valdez [1998]). Therefore, the finite grid is created among $[-1, 1]^p$ and each element of this grid is converted to the copula parameter $\boldsymbol{\theta}$. Moreover, the use of concordance measures gives a normalized expression of the strength of dependencies for all pairs of variables, independently of the used copula families.

The consistency of estimators (5.9) and (5.10) is studied in next section, under general regularity and geometric assumptions on η and the functional $\boldsymbol{\theta} \mapsto P_{\boldsymbol{\theta}}$.

5.2.4 Consistency of worst quantile-related estimators

In this section, we give consistency results of the estimators $\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_N} \widehat{G}_{\boldsymbol{\theta}}^{-1}(\alpha)$ and $\widehat{\boldsymbol{\theta}}_N$, for a growing sequence of grids on the domain $\boldsymbol{\Theta}$. For easier reading, we skip some definitions needed for our assumptions. Section 5.7 in Appendix provides a more complete presentation, including the formal definition of the modulus of increase of the quantile function.

Let α be a fixed value in $(0, 1)$. To approximate $\mathcal{D}$, we consider a sequence of finite discrete grids $(\boldsymbol{\Theta}_N)_{N\geq 1}$ on $\mathcal{D}$ where N is the cardinal of $\boldsymbol{\Theta}_N$ and such that

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta},\, \boldsymbol{\theta}'\in\boldsymbol{\Theta}_N} \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2 \to 0 \quad \text{as } N \text{ tends to infinity.} \tag{5.11}$$

We first introduce technical hypotheses required for the consistency result which are commented further in the text.

Assumption A. *For all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, the distribution $P_{\boldsymbol{\theta}}$ admits a density $f_{\boldsymbol{\theta}}$ for the Lebesgue measure and the copula $C_{\boldsymbol{\theta}}$ admits a density $c_{\boldsymbol{\theta}}$ for the Lebesgue measure on $[0,1]^d$ such that*

$$\begin{aligned} \boldsymbol{\Theta} \times [0,1]^d &\longrightarrow \mathbb{R} \\ \boldsymbol{\theta} \times (x_1, \dots, x_d) &\longrightarrow c_{\boldsymbol{\theta}}(x_1, \dots, x_d) \end{aligned}$$

is a continuous function.

Assumption B. *For all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, $G_{\boldsymbol{\theta}}$ is a continuous function.*

Assumption C. *For all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, $G_{\boldsymbol{\theta}}$ is strictly increasing and the modulus of increase of $G_{\boldsymbol{\theta}}$ at $G_{\boldsymbol{\theta}}^{-1}(\alpha)$ is lower bounded by a positive function $\underline{\epsilon}_{\boldsymbol{\Theta}}$.*

Assumption D. *There exists an unique $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}$ minimizing $\boldsymbol{\theta} \mapsto G_{\boldsymbol{\theta}}^{-1}(\alpha)$.*

Let $(N_n)_{n\geq 1}$ be a sequence of integers such that $N_n \lesssim n^\beta$ for some $\beta > 0$. For every $n \geq 1$ we consider the grid $\boldsymbol{\Theta}_{N_n}$ and for every $\boldsymbol{\theta} \in \boldsymbol{\Theta}_{N_n}$ we compute the empirical quantile $\widehat{G}_{\boldsymbol{\theta}}^{-1}(\alpha)$ from a sample of n i.i.d variables $Y_1, \dots, Y_n$ with $Y_i = \eta(\mathbf{X}_i)$, where the $\mathbf{X}_i's$ are i.i.d. random vectors with distribution $P_{\boldsymbol{\theta}}$. We then introduce the extremum estimator

$$\hat{\boldsymbol{\theta}} := \widehat{\boldsymbol{\theta}}_{N_n}. \tag{5.12}$$

Theorem 3. *Under Assumptions A, B and C, for all $\varepsilon > 0$ we have*

$$P\left(\left|\widehat{G}_{\hat{\boldsymbol{\theta}}}^{-1}(\alpha) - G_C^{-1\star}(\alpha)\right| > \varepsilon\right) \xrightarrow{n\to\infty} 0. \tag{5.13}$$

Moreover, if Assumption D is also satisfied, then for all $h > 0$ we have

$$\mathbb{P}[|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_C^*| > h] \xrightarrow{n\to\infty} 0$$

(proof given in Appendix 5.7).

It would be possible to provide rates of convergence for this extremum quantile and for $\boldsymbol{\theta}^\star$ at the price of more technical proofs, by considering also the dimension metric of the domain $\boldsymbol{\Theta}$ and the modulus of increase of the function $\boldsymbol{\theta} \mapsto G_{\boldsymbol{\theta}}(\alpha)$ (see for instance the proofs of Theorems 1 and 2 in Chazal et al. [2015] for an illustration of such computations). It would be also possible to derive similar results for alternative extremum quantities. One first example, useful in many applications, would be to estimate some risk probability by determining an extremum $\inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} G_{\boldsymbol{\theta}}(y)$ of the CDF for a fixed y.

This consistency result could also be extended for regular functional of $G_{\boldsymbol{\theta}}$ or $G_{\boldsymbol{\theta}}^{-1}$, such that

$$\inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \int_{y\geq y_0} G_{\boldsymbol{\theta}}(y)\mathrm{d}y \quad \text{or} \quad \inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \int_{\alpha\geq\alpha_0} G_{\boldsymbol{\theta}}^{-1}(\alpha)\mathrm{d}y,$$

for some fixed values y_0 and α_0. Extending our results for such quantities is possible essentially because the Dvoretsky-Kiefer-Wolfowitz (DKW) inequality [Dvoretzky et al., 1956], used in the proof, gives an uniform control on the estimation of the CDF and the quantile function.

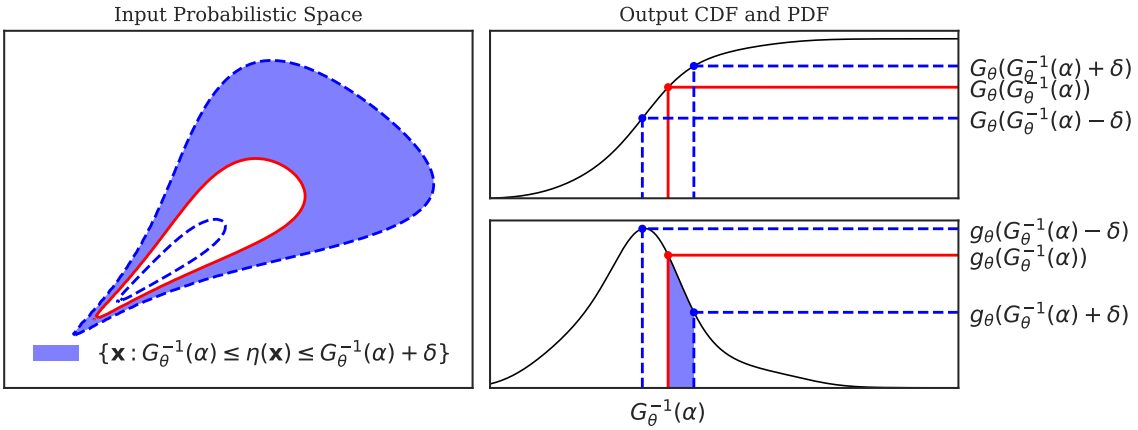


Figure 5.2 *Pre-image (left) and image (right) of a modulus of increase of $G_{\boldsymbol{\theta}}$ at the point $G_{\boldsymbol{\theta}}^{-1}(\alpha)$ for a deviation $\pm\delta$.*

We now discuss the three first assumptions and provide some geometric and probabilistic interpretations of them. Assumption **A** requires some regularity of the input distribution with respect to $\boldsymbol{\theta}$. This is indeed necessary to locate the minimum of the quantile. Assumption **B** and **C** ensure that the output quantile function $G_{\boldsymbol{\theta}}^{-1}$ has a regular behavior in a neighborhood of the computed quantile $G_{\boldsymbol{\theta}}^{-1}(\alpha)$. Assumption **B** ensures that the output distribution $\mathrm{d}G_{\boldsymbol{\theta}}$ has no Dirac masses whereas Assumption **C** ensures that there is no area of null mass inside the domain of $\mathrm{d}G_{\boldsymbol{\theta}}$.

Figure 5.2 illustrates Assumption **B** with a possible configuration of the input distribution. For $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ an $\delta > 0$, we consider a small neighborhood $[G_{\boldsymbol{\theta}}^{-1}(\alpha) - \delta, G_{\boldsymbol{\theta}}^{-1}(\alpha) + \delta]$ of $G_{\boldsymbol{\theta}}^{-1}(\alpha)$, and the pre-image of this neighborhood. The two right figures are the CDF $G_{\boldsymbol{\theta}}$ (top) and PDF $g_{\boldsymbol{\theta}}$ (bottom) of the output variable Y for a given $\boldsymbol{\theta}$. The figure at the left hand represents the contours of the pre-image in the input space. The red plain line is the level set $\eta^{-1}(G_{\boldsymbol{\theta}}^{-1}(\alpha))$ and the dot blue line is the perturbed level set $\eta^{-1}(G_{\boldsymbol{\theta}}^{-1}(\alpha) \pm \delta))$. The blue area in the right figure corresponds to $[G_{\boldsymbol{\theta}}^{-1}(\alpha) - \delta, G_{\boldsymbol{\theta}}^{-1}(\alpha) + \delta]$ and the pre-image of this neighborhood is the blue area in the left figure. Assumption **B** requires that the mass of the blue domain is lower bounded by a positive function $\varepsilon_{\boldsymbol{\Theta}}(\delta)$ that does not depend on $\boldsymbol{\theta}$.

It is possible to give sufficient conditions on the input distribution $F_{\boldsymbol{\theta}}$ and on the geometry of the code η to obtain Assumptions **B** and **C**. Using the definition of the modulus of continuity from Equation (5.7.1) in Appendix, it comes

$$\begin{aligned}
\epsilon_{G_{\boldsymbol{\theta}}}(\delta, G_{\boldsymbol{\theta}}^{-1}(\alpha)) &= \max\left[\int_{\{G_{\boldsymbol{\theta}}^{-1}(\alpha) \leq g \leq G_{\boldsymbol{\theta}}^{-1}(\alpha)+\delta\}} f_{\boldsymbol{\theta}}(\mathbf{x})\mathrm{d}\lambda(\mathbf{x}); \int_{\{G_{\boldsymbol{\theta}}^{-1}(\alpha)-\delta \leq g \leq G_{\boldsymbol{\theta}}^{-1}(\alpha)\}} f_{\boldsymbol{\theta}}(\mathbf{x})\mathrm{d}\lambda(\mathbf{x})\right] \\
&\geq \int_{\{G_{\boldsymbol{\theta}}^{-1}(\alpha) \leq g \leq G_{\boldsymbol{\theta}}^{-1}(\alpha)+\delta\}} f_{\boldsymbol{\theta}}(\mathbf{x})\mathrm{d}\lambda(\mathbf{x})
\end{aligned}$$

Assume that the code η is a Lipschitz and differentiable function with no null derivatives almost everywhere in the neighborhood of $G_{\boldsymbol{\theta}}^{-1}(\alpha)$. Then, using the *coarea formula* (see for instance

Evans and Gariepy [2015], Section 3.4.4, Proposition 3), we find that

$$\epsilon_{G_{\boldsymbol{\theta}}}(\delta, G_{\boldsymbol{\theta}}^{-1}(\alpha)) \geq \int_{G_{\boldsymbol{\theta}}^{-1}(\alpha)}^{G_{\boldsymbol{\theta}}^{-1}(\alpha)+\delta} \left[\int_{\eta^{-1}\{u\}} \frac{f_{\boldsymbol{\theta}}}{\|\nabla \eta\|} \mathrm{d}\mathcal{H}^{d-1} \right] \mathrm{d}u,$$

where $\mathcal{H}^{d-1}$ is the $d-1$ dimensional Hausdorff measure (see for instance Chapter 2 in Evans and Gariepy [2015]). If the copula and the code are such that there exists a constant I such that for any $\boldsymbol{\theta} \in \mathcal{D}$ and any u in the support of $\mathrm{d}G_{\boldsymbol{\theta}}$

$$\int_{\eta^{-1}\{u\}} f_{\boldsymbol{\theta}} \, \mathrm{d}\mathcal{H}^{d-1} \leq I,$$

then we find that

$$\epsilon_{G_{\boldsymbol{\theta}}}(\delta, G_{\boldsymbol{\theta}}^{-1}(\alpha)) \geq \delta \frac{I}{\|\nabla \eta\|_{\infty}}.$$

Note that $\|\nabla \eta\|_{\infty} < \infty$ since η is assumed to be Lipschitz. We have proved that Assumption **C** is satisfied in this context. Finally, by rewriting again the co-area formula for $G_{\boldsymbol{\theta}}(y)$, we find that Assumption **B** is satisfied as soon as the set of stationary points ($\|\nabla \eta(x)\| = 0$) of all level set $\eta^{-1}\{u\}$ has null mass for the Hausdorff measure.

In conclusion, we see that for smooth copulas, Assumptions **C** and **B** mainly depend on the regularity of the code, by requiring on one side that η does not oscillate to much and on the other side that the set of stationary points does not have a positive mass on the level sets of η.

5.3 A preliminary study of the copula influence on quantile minimization

This section is dedicated to a preliminary exploration of the influence of copula structure on the behavior of the worst quantile, illustrated with toy examples. Especially, while it could be expected that $G_{\boldsymbol{\theta}}^{-1}(\alpha)$ is a monotonic function with $\boldsymbol{\theta}$, and that the minimum can be reached for a trivial copula (i.e., reaching the Fréchet-Hoeffding bounds). Our experiments show that this behavior is not systematic.

5.3.1 About the copula choice

One of the most common approaches to model the dependence between random variables is to assume linear correlations feeding a Gaussian copula. In this case, the problem is reduced by determining the correlation matrix of $\mathbf{X}$ that minimizes $G_{\boldsymbol{\theta}}^{-1}(\alpha)$. However, the positive semi-definite constraint on the correlation matrix makes the exploration difficult and the minimization harder when the problem dimension increases. Moreover, such a Gaussian assumption is very restrictive and is inappropriate for simulating heavy tail dependencies [Malevergne et al., 2003]. Still in this elliptical configuration, the t-copulas [Demarta and McNeil, 2005] can be used to counterpart these problems. Nevertheless, tail dependencies are symmetric and with equal strengths for each pair of variables. Another alternative is to consider multivariate Archimedean copulas [McNeil

and Nešlehová, 2009] which are great tools to describe asymmetric tail dependencies. However, only one parameter governs the strength of the dependence among all the pairs, which is very restrictive and not flexible in high dimension. For a same correlation measure between two random variables, multiple copulas can be fitted and lead to a different distribution of Y.

It is clear that the copula choice of $\mathbf{X}$ has a strong impact on the distribution of Y (see for instance Tang et al. [2013]). Therefore, various copula types should be tested to determine the most conservative configuration. In the following, we may consider a flexible approach setting by modeling the input multivariate distribution using regular vine copulas (R-vines). The necessary basics of R-vines are introduced in Section 5.4.1 and detailed in Appendix 6.

5.3.2 About the monotony of the quantile

For many simple case studies case studies, the worst quantile is reached for perfect dependencies (Fréchet-Hoeffding bounds). More generally, when the function has a monotonic behavior with respect to many variables, it is likely that the minimum output quantile is reached at the boundary of $\mathbf{\Theta}$. This phenomenon is observed for various physical systems.

To illustrate this phenomenon, we consider a simplified academic model that simulates the overflow of a river over a dike that protects industrial facilities. The river overflow S is described by

$$S = H_d + C_b - Z_v - H \qquad \text{with} \qquad H = \left(\frac{Q}{BK_s \sqrt{\frac{Z_m - Z_v}{L}}} \right)^{0.6}, \tag{5.14}$$

such as, when $S < 0$, a flooding occurs. The involved parameters of (5.14) are physical characteristics of the river and the dike (e.g., flow rate, height of the dike) which are described by random variables with known marginal distributions. See Iooss and Lemaître [2015] for more information. For a given risk α, we aim at quantifying the associated overflow's height describe by the α-quantile of S. We extend this model by supposing that the friction (Strickler-Manning) coefficient K_s and the maximal annual flow rate Q are dependent with an unknown dependence structure. To show the influence of a possible correlation between K_s and Q on the quantile of S, we describe their dependence structure with multiple copula families.

Figure 5.3 shows the variation of the estimated quantile of S (with a large sample size) in function of the Kendall coefficient τ between K_s and Q for different copula families. We observe different slopes of variation for the different copula families, with lower quantile values for the copulas with heavy tail dependencies (i.e., Clayton, Joe). At independence ($\tau = 0$) and for the counter-monotonic configuration ($\tau = -1$),the quantile values of these families are obviously equivalent. This variation is slight and the quantile is still above zero, but this shows how the dependencies can influence the results of a reliability problem. This illustration shows that the minimum is reached at the boundary of the exploration space, where the two variables are perfectly correlated.

We can take advantage of this observation to speed up the algorithms presented in the next sections by exploring only the boundaries of $\mathbf{\Theta}$. However, assuming that the minimum is reached

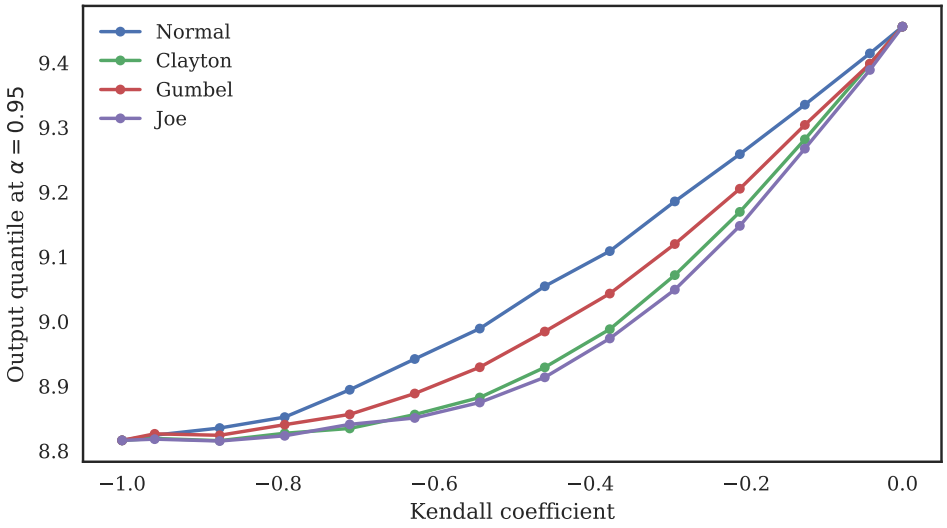


Figure 5.3 *Variation of the quantile of the overflow distribution with the Kendall coefficient τ for $\alpha = 95\%$ and different copula families (Gaussian, Clayton, Gumbel and Joe).*

on the boundary of $\boldsymbol{\Theta}$ is a strong assumption that can be unsatisfied in some applications. See Fallacy 3 of Embrechts et al. [2002] for a highlight of this pitfall.

To illustrate this statement, we now give a counter example in the bidimensional setting. We assume uniform marginal distributions for the input such that $X_1 \sim \mathcal{U}(-3, 1)$ and $X_2 \sim \mathcal{U}(-1, 3)$, and we consider the model function

$$\eta(x1, x2) = 0.58x_1^2x_2^2 - x_1x_2 - x_1^2 - x_2^2. \tag{5.15}$$

The same experience as for Figure 5.3 is established and the results are shown in Figure 5.4. The slopes of the quantile estimations with the Kendall coefficient, for each copula families, are quite different than the results of Figure 5.3. We observe that the quantile is not monotonic with the Kendall coefficient and its minimum is not reached at the boundary, but for $\tau \approx 0.5$. Moreover, the Gaussian copula is the family that minimizes the most the quantile. It shows that copulas with tail dependencies are not always the most penalizing.

A second example, inspired from Example 6 of Embrechts et al. [2002], also shows that the worst case dependence structure in an additive problem is not necessary for perfectly correlated variables. We consider a simple portfolio optimization problem with two random variables X_1 and X_2 with generalized Pareto distributions such as $F_1(x) = F_2(x) = \frac{x}{1+x}$. We aim at maximizing the profit of the portfolio, which is equivalent as minimizing the following additive model function

$$\eta(X_1, X_2) = -(X_1 + 10X_2). \tag{5.16}$$

We consider the median ($\alpha = 0.5$) of the output as an efficiency measure. Figure 5.5 shows the output median in function of the Kendall coefficient τ between X_1 and X_2. Just like the previous example, we observe a non-monotonic slope of the median in function of τ. The variation can

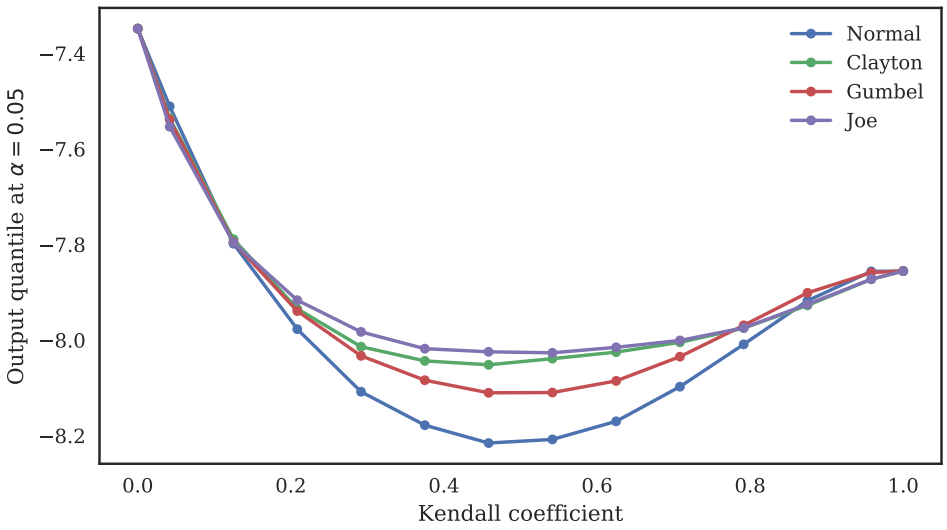


Figure 5.4 *Variation of the output quantile with the Kendall coefficient τ for $\alpha = 5\%$ and different copula families (Gaussian, Clayton, Gumbel and Joe).*

be significant and the minimum is obtained at $\tau \approx 0.53$ for the heavy tail copula families (i.e., Clayton and Joe). The phenomenon can be explained by the marginal distributions of the random variables, which are close Pareto distributions. A large correlation seems to diminish the influence of the tails, which gives a higher quantile value. This explains why the minimum is obtained for a dependence structure other that independence or the perfect dependence.

Therefore, these examples show that the worst quantile can be reached for other configurations than the perfect dependencies.

5.4 Quantile minimization and choice of penalized correlation structure

This section first provides a rationale for choosing the so-called R-vine structure as a preferential copula structure for modeling the variety of correlations between inputs. Then, the search for a minimum quantile is presented in two times. Subsection 5.4.2 proposes an exhaustive grid-search algorithm for estimating this quantile when the R-vine copula structure is fixed with a given pair-copula families and indexed by the parameter vector $\boldsymbol{\theta}$. Subsection 5.4.3 extends this rigid framework by permitting the search of particular sub-copula pairwise structures, such that the minimization be more significant. In each situation, examples are provided to demonstrate the feasibility of the approach.

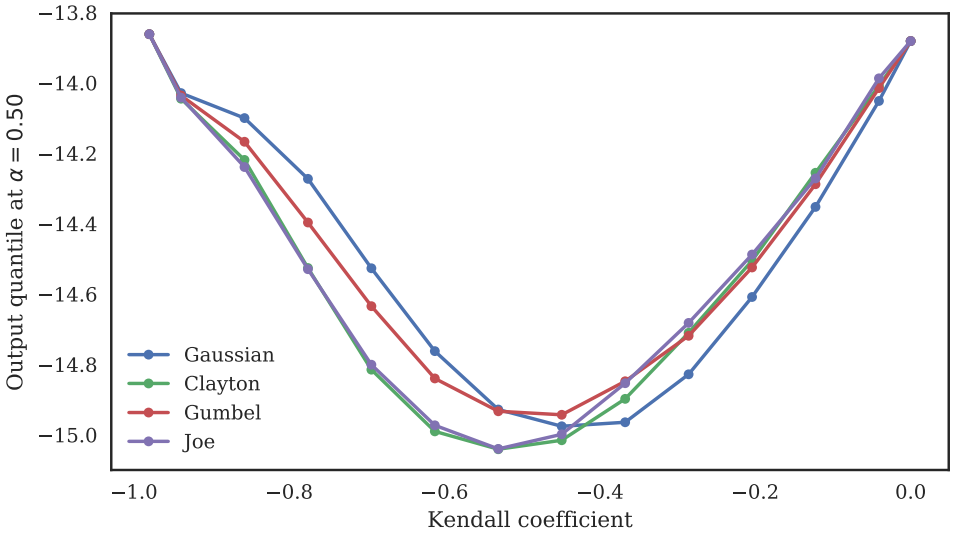


Figure 5.5 *Variation of the portfolio median with the Kendall coefficient τ for different copula families.*

5.4.1 A rationale for R-vine copula structures

Representing multi-dimensional dependence structures in high dimensional settings is a challenging problem. For the following definition, we simplify the expressions by omitting the use of $\boldsymbol{\theta}$: $f = f_{\boldsymbol{\theta}}$, $F = F_{\boldsymbol{\theta}}$ and $c = c_{\boldsymbol{\theta}}$. By recursive conditioning, the joint density can be written as a product of conditioning distributions such as

$$f(x_1, \ldots, x_d) = f_1(x_1) \cdot f_{2|1}(x_2|x_1) \cdot f_{3|1,2}(x_3|x_1, x_2) \cdots f_{d|1,2,\ldots d-1}(x_d|x_1, x_2, \ldots, x_{d-1}). \quad (5.17)$$

For clarity reason, we now simplify the expression with $f_{3|1,2} = f_{3|1,2}(x_3|x_1, x_2)$ and so on for other orders. From (5.4), the conditioning densities of (5.17) can be rewritten as products of conditioning copula and marginal densities. For example, in a case of three variables and using (5.17), one possible decomposition of the the joint density can be written as

$$f(x_1, x_2, x_3) = f_1 \cdot f_{2|1} \cdot f_{3|1,2}. \quad (5.18)$$

Using (5.4), the reformulation of $f_{3|1,2}$ leads to

$$f_{3|1,2} = \frac{f_{1,3|2}}{f_{1|2}} = \frac{c_{1,3|2} \cdot f_{1|2} \cdot f_{3|2}}{f_{1|2}} = c_{1,3|2} \cdot f_{3|2} \quad (5.19)$$

where $c_{1,3|2} = c_{1,3|2}(F_{1|2}(x_1|x_2), F_{3|2}(x_3|x_2))$. By developing $f_{3|1,2}$ in the same way, we find that

$$f_{3|1,2} = c_{1,3|2} \cdot c_{2,3} \cdot f_3. \quad (5.20)$$

Thus, by replacing the expression of $f_{3|1,2}$ in (5.18) and doing the same procedure for $f_{2|1}$, the joint density can be written as

$$f(x_1, x_2, x_3) = f_1 \cdot f_2 \cdot f_3 \cdot c_{1,2} \cdot c_{2,3} \cdot c_{1,3|2}. \quad (5.21)$$

This final representation of the joint density based on pair-copulas has been developed in Joe [1996] and is called the pair-copula construction (PCC). The resulting copula represented by the product of conditional copulas in (5.21) offers a very flexible way to construct high-dimensional copulas. However, it is not unique; indeed, (5.17) has numerous decomposition forms and it increases with the dimension.

To describe all such possible constructions in an efficient way, Bedford and Cooke [2001, 2002] introduced the vine models. This graphical tool, based on a sequence of trees, gives a specific way to decompose the multivariate probability distribution. Basically, a vine model is defined by

- a structure of trees which can be represented by a matrix [Morales Nápoles, 2010],
- a copula family for each pair of the structure,
- a parameter for each pair-copula.

A R-vine is the general construction of a vine model, but particular cases exists such as the D-vines and C-vines, described in Appendix 5.7.3. Vine models were deeply studied in terms of density estimation and model selection using maximum likelihood [Aas et al., 2009], sequential estimation [Kurowicka, 2011; Dissmann et al., 2013], truncation [Aas et al., 2012] and Bayesian techniques [Gruber and Czado, 2015]. Their popularity and well-known flexibility led us to use R-vines in this article, despite the fact that in our context we are looking for a conservative form and not to select the most appropriate form with given data, in absence of correlated observations.

5.4.2 Estimating a minimum quantile from a given R-vine

5.4.2.1 Grid-search algorithm

Let $\Omega = \{(i,j) : 1 \leq i,j \leq d\}$ be the set of all the possible pairs of $\mathbf{X}$, in a d-dimensional problem. The number of pairs p is associated to the size of Ω such as $p = |\Omega| = \binom{d}{2} = d(d-1)/2$. We define $\mathcal{V}$ as the vine structure and we consider fixed copula families for each pair. In this article, we only consider single parameter pair-copulas, such that the parameter $\boldsymbol{\theta}$ is a p-dimensional parameter vector with a definition space $\boldsymbol{\Theta} := \prod_{(i,j)\in\Omega} \Theta_{i,j}$ where $\Theta_{i,j}$ is the parameter space for the pair-copula of the pair (i,j). However, the methodology can easily be extended to multi-parameter pair-copulas. Note that a pair-copula can be conditioned to other variables, depending on its position in the vine structure $\mathcal{V}$. Thus, the input distribution $\mathrm{d}F_{\boldsymbol{\theta}}(\mathcal{V})$ is defined by the vine structure $\mathcal{V}$, the copula families and the parameter $\boldsymbol{\theta}$. Also note that the copula parameter $\boldsymbol{\theta}$ is associated to the R-vine structure $\mathcal{V}$ (i.e., $\boldsymbol{\theta} = \boldsymbol{\theta}_{\mathcal{V}}$), see Section 5.4.2.2. For the sake of clarity, we simplify the notation to $\boldsymbol{\theta}$ only.

The most direct approach to estimate the minimum quantile is the Exhaustive Grid-Search

algorithm, described by the following pseudo-code.

Algorithme 6 : Exhaustive Grid-Search algorithm to minimize the output quantile.

Data : A vine structure $\mathcal{V}$, a fixed grid $\mathbf{\Theta}_N$, a sample size n

for $\boldsymbol{\theta} \in \mathbf{\Theta}_N$ **do**

  1. Simulate a sample $\{\mathbf{X}_i\}_{i=1}^n$ according to $\mathrm{d}F_{\boldsymbol{\theta}}(\mathcal{V})$;

  2. Evaluate $\{Y_i = \eta(\mathbf{X}_i)\}_{i=1}^n$;

  3. Compute $\widehat{G}_{\boldsymbol{\theta}}^{-1}(\alpha)$: empirical quantile of $\{Y_i\}_{i=1}^n$;

Result : $\min_{\boldsymbol{\theta}\in\mathbf{\Theta}_N} \widehat{G}_{\boldsymbol{\theta}}^{-1}(\alpha)$

For a given vine structure $\mathcal{V}$, copula families, a grid $\mathbf{\Theta}_N$ and a sample size n, three steps are needed for each $\boldsymbol{\theta} \in \mathbf{\Theta}_N$. The first step simulates an input sample $\{\mathbf{X}_i\}_{i=1}^n$ according to the distribution $\mathrm{d}F_{\boldsymbol{\theta}}(\mathcal{V})$ for a given sample size n. The second evaluates the sample through the model η. The third estimates the output quantile from the resulting sample $\{Y_i = \eta(\mathbf{X}_i)\}_{i=1}^n$. The minimum quantile is took among the results of each loop.

5.4.2.2 Influence of the vine structure

Using R-vines, the dependence parameter $\boldsymbol{\theta}$ is associated to the vine structure $\mathcal{V}$. Due to the hierarchy of the vine structure, some pair-copulas are conditioned to other variables and thus for their parameters. As an illustration, let us consider two vine structures with the two following copula densities, with the same simplified expressions as for (5.21):

$$c_{\mathcal{V}_1}(x_1,x_2,x_3,x_4) = c_{\theta_{1,3}} \cdot c_{\theta_{1,2}} \cdot c_{\theta_{2,4}} \cdot c_{\theta_{2,3|1}} \cdot c_{\theta_{1,4|2}} \cdot c_{\theta_{3,4|1,2}} \tag{5.22}$$

$$c_{\mathcal{V}_2}(x_1,x_2,x_3,x_4) = c_{\theta_{1,3}} \cdot c_{\theta_{3,4}} \cdot c_{\theta_{2,4}} \cdot c_{\theta_{1,4|3}} \cdot c_{\theta_{2,3|4}} \cdot c_{\theta_{1,2|3,4}}. \tag{5.23}$$

The difference between these densities is the conditioning of some pairs, the dependence parameters of theses vines are $\boldsymbol{\theta}_{\mathcal{V}_1} = [\theta_{1,2}, \theta_{1,3}, \theta_{1,4|2}, \theta_{2,3|1}, \theta_{2,4}, \theta_{3,4|1,2}]$ and $\boldsymbol{\theta}_{\mathcal{V}_2} = [\theta_{1,2|3,4}, \theta_{1,3}, \theta_{1,4|3}, \theta_{2,3|4}, \theta_{2,4}, \theta_{3,4}]$. Applying the same grid for these two vines may give different results due to the conditioning order from the vine structure. For example, if the pair X_3-X_4 is very influential on minimizing the output quantile, it would be more difficult to find a minimum with $\mathcal{V}_1$ than $\mathcal{V}_2$ due to the conditioning of the pair with X_1 and X_2 in $\mathcal{V}_1$. However, if the grid is thin enough, the minimum from these two vines should be equivalent.

To counter this difficulty, one possible option consists in randomly permuting the indexes of the variables and repeating the algorithm several times to visit different vines structures.

5.4.2.3 Computational cost

For one given R-vine structure and one fixed copula family at each pair, the overall cost of the method is equal to nN. However, as explained in § 5.2.4, the finite grid $\mathbf{\Theta}_N$, should be thin enough to reasonably explore $\mathbf{\Theta}$. Therefore, N should increase with the number of dimensions d and more specifically with the number of pairs $p = \binom{d}{2}$. A natural form for N would be to write

it as $N = \gamma^p$, where $\gamma \in \mathbb{R}^+$. Thus, the overall cost of the exhaustive grid-search would be equal to $n\gamma^{\binom{d}{2}}$. The cost is in $\mathcal{O}(\gamma^{d^2})$ which makes the method hardly scalable when the dimension d increases.

5.4.3 Iterative search for a penalizing R-vine structure: a greedy heuristic based on pairwise copula

5.4.3.1 Going further in quantile minimization

With Algorithm 6, the previous subsection proposes an exhaustive grid-search strategy to determine a R-vine copula $C_{\tilde{\theta}}$ such that the associated output quantile $G^{-1}_{\tilde{\boldsymbol{\theta}}}(\alpha)$ be the smallest (and also the most conservative in a structural reliability context). This approach remains however limited in practice since $C_{\tilde{\theta}}$ for fixed pair-copula families (e.g., Archimedean or max-stable copulas) and $\mathcal{V}$ which is a member of the set $\mathcal{F}_d$ of all the possible $d-$dimensional R-vine structure. Intuitively, a more reliable approach to quantile minimization should be based on mixing this estimation method with a selection among all members of the finite set $\mathcal{F}_d$, as well for the copula families. It is indeed likely that searching within an associative class of copulas like Archimedean ones, allowing modeling dependence in arbitrarily high dimensions, be a too rigid choice for estimating the minimum $G^{-1}_{\tilde{\boldsymbol{\theta}}}(\alpha)$.

A minimum quantile can probably be found using a R-vine structure defined by conditional pairwise sub-copulas (according to (5.21)) that are not part of the same rigid structure. However, a brute force exploration of $\mathcal{F}_d$ would be conducted at an exponential cost increasing with d [Morales-Nápoles, 2011]. If we also consider the large computational cost of an exhaustive grid-search for a large number of dependent variables (as explained in § 5.4.2), this approach is not feasible in practice for high dimensions.

For this reason, it is proposed to extend Algorithm 6 by a greedy *heuristic* that dynamically selects the most influential correlations between variables while limiting the search to pairwise correlations. Doing so, minimizing the output quantile can be conducted in a reasonable computational time. Therefore the selected $d-$dimensional vine structure would be filled with independent pair-copulas except for the pairs that are influential on the minimization.

This working limitation, interpreted as a sparsity constraint, is based on the following assumption: it is hypothesized that only few pairs of variables have real influences on the minimization. It is close in spirit to the main assumption of global sensitivity analysis applied to computer models, according to which only a limited number of random variables has a major impact on the output [Saltelli et al., 2000; Iooss and Lemaître, 2015].

5.4.3.2 General principle

The method basically relies on an iterative algorithm exploring pairwise correlations between the uniform random variables $U_j = F_j^{-1}(X_j)$ and progressively building a non-trivial R-vine structure,

adding one pair of variable to the structure at each iteration. Starting at step $k = 0$ from the simple independent copula

$$C_{\boldsymbol{\theta}^{(0)}}(u_1, \ldots, u_d) = \prod_{j=1}^{d} u_j,$$

the algorithm finally stops at a given step $k = K$ while proposing a new copula $C_{\boldsymbol{\theta}^{(K)}}$ associated to a R-vine structure $\mathcal{V}_K$ mostly composed of independent pair-copulas.

At each iteration k, we denote by Ω_k the *selected pairs* which are considered non-trivial (non-independent) due to their influence on the quantile minimization. Let $\Omega_{-k} = \Omega \backslash \Omega_k$ be the *candidate pairs*, which were not the remaining pairs, which influence on the minimization is still to be tested and are still considered independent. We also consider $\mathbf{B}$ as a set of candidate copula families. The pseudo-code of Algorithm 7 shows in detail how this iterative exploration and building is conducted. More algorithms in Appendix 5.7.3.2 described how to construct a vine structure with a given list of indexed pairs of variable.

5.4.3.3 Example

Consider the four-dimensional $(d = 4)$ situation such as $\mathbf{X} = (X_1, \ldots, X_4)$ where, for to the sake of simplicity, all marginal distributions of $\mathbf{X}$ are assumed to be uniform on $[0, 1]$. We consider a simple additive model described by

$$\eta(\mathbf{X}) = 30X_1 + 10X_3 + 100X_4. \tag{5.24}$$

For an additive model and uniform margins, the output quantile is monotonic with the dependence parameters (see Section 5.3.2) which locates the minimum quantile at the edge of $\boldsymbol{\Theta}$. Thus, Step 1.b. of Algorithm 7 is simplified by considering only Fréchet-Hoeffding copulas in the exploration.

In this illustration we consider $\alpha = 0.1$ and we select $n = 300,000$ large enough in order to have a great quantile estimation and the algorithm stops at $K = 3$. Figure 5.6 shows, for each iteration k, the $p-k$ vine structures that have been created by the algorithm. The red nodes and edges are the candidate pairs $(i, j) \in \Omega_{-k}$ and the blue nodes and edges are the selected pairs Ω_k. At iteration $k = 0$, the selected pair is $(1, 4)$ with an estimated minimum quantile of -52.18. At iteration $k = 1$, the second selected pair is $(3, 4)$ with an estimated minimum quantile of -56.03. At iteration $k = 2$, the third selected pair is $(2, 4)$ with an estimated minimum quantile of -56.23.

We observe that X_4 appears in all the selected pairs. This is not surprising since X_4 is the most influential variable with the largest coefficient in (5.24). The algorithm considers D-vines by default, but this is important for the first iterations since most of the pairs are independent. When it is possible, the algorithm creates a vine such as the selected pairs and the candidate pair are in the first trees. For example, the fourth vine at iteration $k = 2$ with the candidate pair $(2, 4)$ shows a R-vine structure that respects the ranking of the listed pairs. However, the third vine at iteration $k = 2$ for the candidate pair $(1, 3)$ along with the selected pairs $\{(1, 4), (3, 4)$ could respect the ranking and set all the pairs in the first tree altogether. Thus, using Algorithm 8 in Appendix, a valid vine structure is determined by placing the candidate pair $(1, 3)$ in the next tree.

Algorithme 7 : Minimization of the output quantile and estimation of $\boldsymbol{\theta}^{(K)}$ over an increasing family of R-vine structures.

Initialization:
Iteration: $k = 0$;
Selected pairs: $\Omega_0 = \emptyset$;
Selected families: $\mathbf{B}_0 = \emptyset$;
while $k \leq K$ **do**
  Copula parameter space of the selected pairs: $\boldsymbol{\Theta}_k = \prod_{(i,j)\in\Omega_k} \Theta_{i,j}$;
  1. Explore the set of candidate pairs Ω_{-k};
  for $(i,j) \in \Omega_{-k}$ **do**
    a. Create a vine structure $\mathcal{V}_{(i,j)}$ using the procedure of Section 5.7.3.2 applied to the list $\Omega_k \cup (i,j)$;
    b. Explore the set of candidate families $\mathbf{B}$;
    for $\mathcal{B} \in \mathbf{B}$ **do**
      Apply Algorithm 6 with the pair-copula families $\mathcal{B} \cup \mathbf{B}_k$;
      (i) Define a $(k+1)-$dimensional grid $\boldsymbol{\Delta}_{i,j}$ of $\boldsymbol{\Theta}_k \times \Theta_{i,j}$ with cardinality N_k;
      (ii) Select the minimum over the grid $\boldsymbol{\Delta}_{i,j}$:

$$\hat{\boldsymbol{\theta}}_{\mathcal{B}} = \underset{\boldsymbol{\theta}_{\mathcal{B}}\in\boldsymbol{\Delta}_{i,j}}{\operatorname{argmin}} \left\{ \widehat{G}^{-1}_{\boldsymbol{\theta}_{\mathcal{B}}}(\alpha) \right\}.$$

    c. Select the minimum among $\mathbf{B}$

$$\begin{aligned} \mathcal{B}_{i,j} &= \underset{\mathcal{B}\in\mathbf{B}}{\operatorname{argmin}} \left\{ \widehat{G}^{-1}_{\hat{\boldsymbol{\theta}}_{\mathcal{B}}}(\alpha) \right\} \\ \hat{\boldsymbol{\theta}}_{i,j} &= \hat{\boldsymbol{\theta}}_{\mathcal{B}_{i,j}} \end{aligned}$$

  2. Select the minimum among Ω_{-k}

$$\begin{aligned} (i,j)^{(k)} &= \underset{(i,j)\in\Omega_{-k}}{\operatorname{argmin}} \left\{ \widehat{G}^{-1}_{\hat{\boldsymbol{\theta}}_{i,j}}(\alpha) \right\}, \\ \mathcal{V}^{(k)} &= \mathcal{V}_{(i,j)^{(k)}}, \\ \hat{\boldsymbol{\theta}}^{(k)} &= \hat{\boldsymbol{\theta}}_{(i,j)^{(k)}} \\ \mathcal{B}^{(k)} &= \mathcal{B}_{(i,j)^{(k)}} \end{aligned}$$

  3. Check the stopping condition;
  if $\widehat{G}^{-1}_{\hat{\boldsymbol{\theta}}^{(k)}}(\alpha) \geq \widehat{G}^{-1}_{\hat{\boldsymbol{\theta}}^{(k-1)}}(\alpha)$ **then**
    $K = k - 1$;
  else
    Extend the list of selected pairs: $\Omega_k = \Omega_k \cup (i,j)^{(k)}$ and families: $\mathbf{B}_k = \mathbf{B}_k \cup \mathcal{B}^{(k)}$;
    if $k < K$ **and** computational budget not reached **then**
      New iteration: $k = k + 1$;

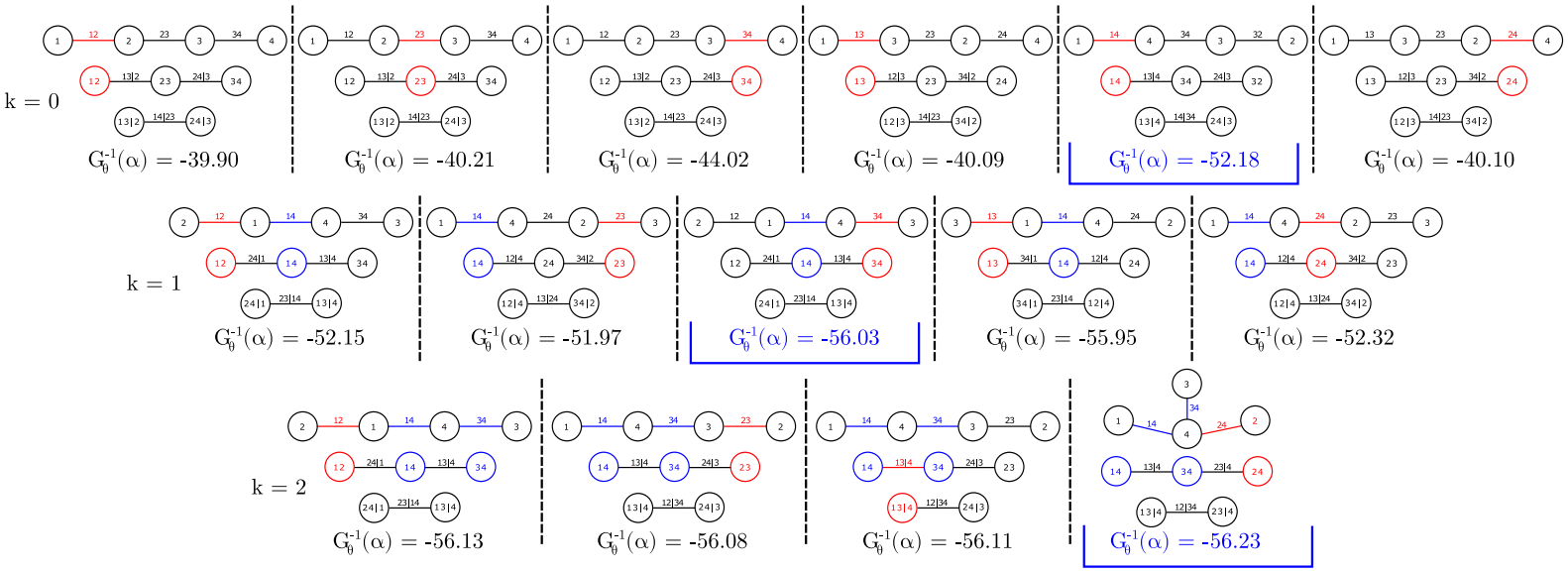


Figure 5.6 *Illustration of the vine structures created during the 3 iterations of the algorithm for the example of Section 5.4.3.3. The candidate and selected pairs are respectively represented in red and blue. The quantile associated to the selected pair of each iteration is written in blue.*

5.4.3.4 Computational cost

The number of model evaluations is influenced by several characteristics from the probabilistic model and from the algorithm. Let $|\mathbf{B}|$ be the number of family candidates. The total number of runs is

$$N = |\mathbf{B}|\frac{n}{2}\sum_{k=0}^{K} N_k \times (d(d-1) - 2k). \tag{5.25}$$

The sum corresponds to the necessary iterations to determine the influential pairs. The maximum possible cost is if all the pairs are equivalently influential (i.e., $K = p = d(d-1)/2$), which would be extremely high. The term nN_k is the cost from the grid-search quantile minimization at step **2.** of the algorithm. The greater N_k is and the better the exploration of $\mathbf{\Theta}_k \cup \Theta_{i,j}$. Because the dimension of $\mathbf{\Theta}_k$ increases at each iteration k, it is normal that N_k should also increases with k (e.g. $N_k = \gamma^{\beta k}$, where γ and β are constants). Also, the greater n is and the better the quantile estimations. The second term is the cost from the input dimension d which influences the number of candidate pairs Ω_{-k} at each iteration k.

Extensions can be implemented to reduce the computational cost such as removing from Ω, the pairs that are not sufficiently improving the minimization.

5.5 Applications

The previously proposed methodology is applied to a toy example and a real industrial case-study. It is worth to mention that these experiments (and future ones) can be conducted again using the Python library `dep-impact` [Benoumechiara, 2018], in which are encoded all the procedures of estimation and optimization presented here.

5.5.1 Numerical example

We pursue and extend the portfolio example considered in Section 5.3.2 and illustrated on Figure 5.5. The numerical model η is now defined by the weighted sum

$$Y = \eta(\mathbf{X}) = -\boldsymbol{\beta}\mathbf{X}^T = -\sum_{j=1}^{d} \beta_j X_j, \tag{5.26}$$

where the $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)$ is a vector of constant weights. The margins of the random vector $\mathbf{X}$ follow the same generalized Pareto distribution with scale σ and shape parameter ξ. Note that the bivariate example in Section 5.3.2 considered $\beta = 1$ and the distribution parameters as $\sigma = 1$ and $\xi = 1$. In the following examples, we aim at minimizing the median ($\alpha = 0.5$) of the output distribution. We chose to fix the marginal distribution's parameters at $\sigma = 10$ and $\xi = 0.75$, and we set the constant vector $\boldsymbol{\beta}$ to a base-10 increasing sequence such that $\boldsymbol{\beta} = (10^{1/d}, 10^{2/d}, \ldots, 10)$. This choice of weights aims to give more influence to the latest components of $\mathbf{X}$ on Y. Thus, some pairs of variables should be more important in the minimization of the output quantile, as required by the sparsity constraint. We also took n large enough to estimate the output quantile with high precision (i.e. $n = 300,000$).

For all these experiments the results from the different methods can be compared.

- Method 1: the grid-search approach with an optimized LHS sampling [McKay et al., 1979] inside $\boldsymbol{\Theta}$ and a random vine structure,
- Method 2: the iterative algorithm with an increasing grid-size of $N_k = 25^*(k+1)^2$.

The Method 1 is established with the same computational budget as Method 2.

5.5.1.1 Dimension 3

In a three dimensional problem, only three pairs of variables ($p = 3$) are involved in the dependence structure. The sampling size of $\boldsymbol{\Theta}$ in Method 1 is set to 400, which is great enough to explore a three dimensional space. The results are displayed on Figure 5.7: the estimated quantiles from Method 1 (blue dots) with a convex hull (blue dot line) and the quantile at independence (dark point) are provided. It also highlights the minimum estimated quantiles from Methods 1 and 2 which are respectively represented in blue and red points. We also show in green point, the minimum quantile by considering only the Fréchet-Hoeffding bounds. For each minimum, the 95 % bootstrap confidence intervals is displayed in dot lines.

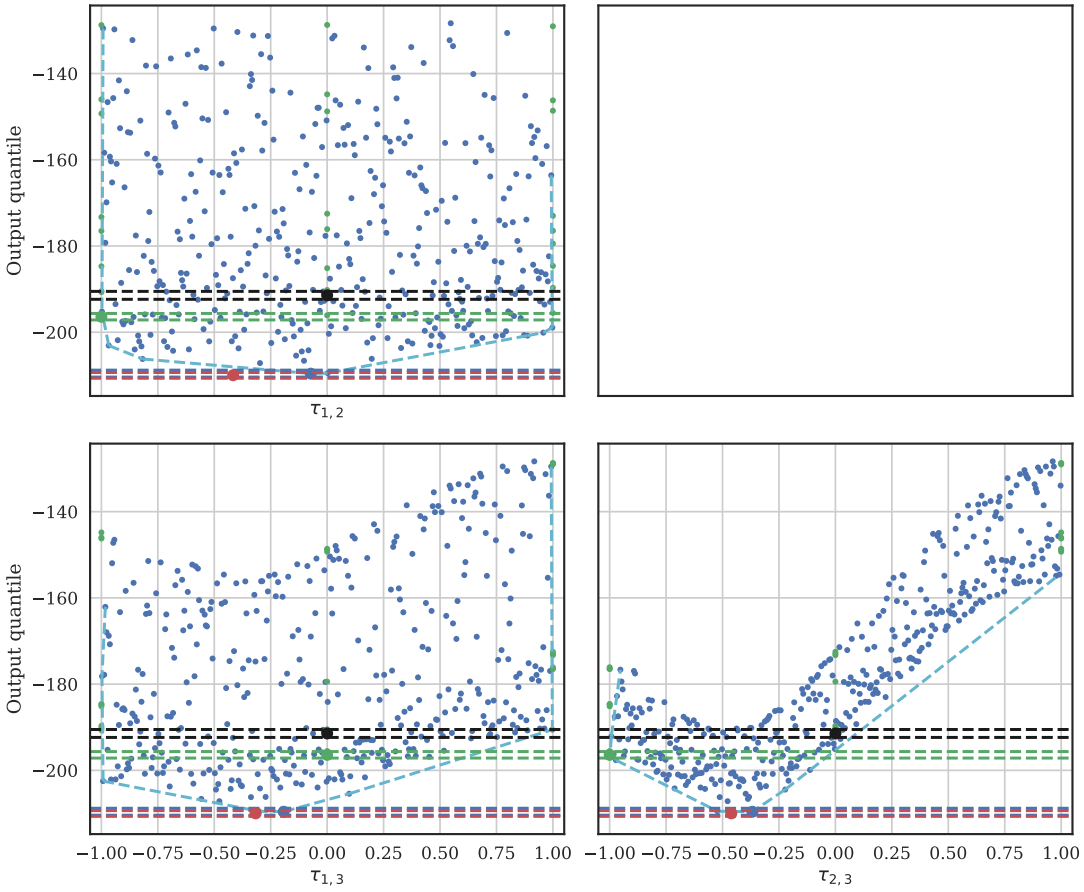


Figure 5.7 *Matrix plot of the output median in function of the Kendall coefficient of each pair. The blue dots represents the estimated quantiles of Method 1. The black point is the quantile at independence and the minimum of Method 1 and 2 are the red and blue points, which are equivalent here. The green point is the the minimum with only Fréchet-Hoeffding bounds. The 95 % bootstrap confidence intervals are displayed in dot lines.*

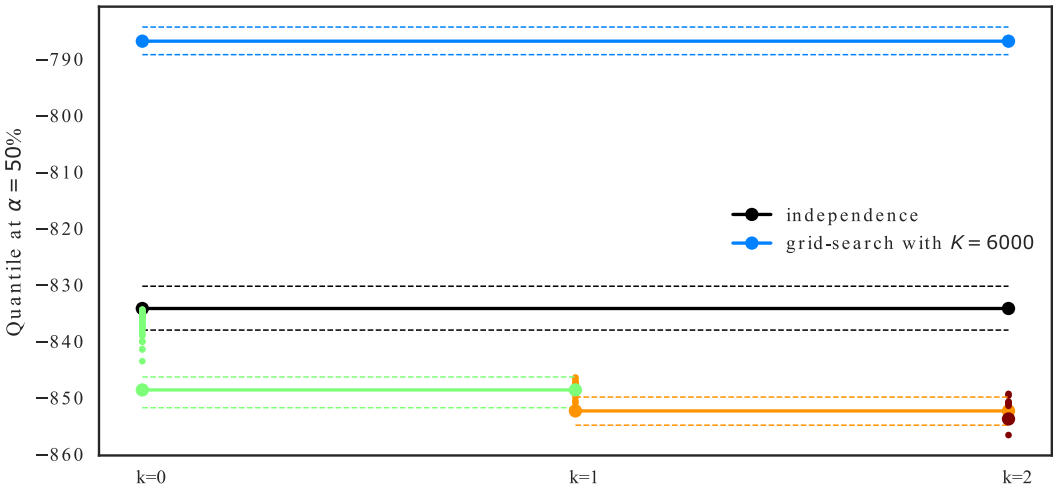


Figure 5.8 *Minimum quantile results with the iteration k of the iterative procedure. The quantile at independence is shown in dark line. The minimum quantiles from Method 1 is show in blue lines. The other lines and dots colors are the results from Method 2. For each iteration, the small dots are the estimated quantiles of all candidates and the point is the minimum. The 95 % bootstrap confidence intervals are also displayed for the independence and each minimums.*

This low dimensional problem confirms the non-monotonic form of the quantile with the dependence parameter, in particular for the variation of the quantile in function of $\tau_{2,3}$. As expected, the pair X_2-X_3 is more influential on the output quantile due to the large weights on X_2 and X_3. The minimum values obtained by each method are still lower that the results given by an independent configuration. The minimum using Fréchet-Hoeffding bounds is also provided to show that the minimum is not at the boundary of $\boldsymbol{\Theta}$. Method 1 and 2 have very similar minimum results.

5.5.1.2 Dimension 10

To illustrate the advantages of the iterative procedure, we now consider $d = 10$. In this example, we chose to only consider a Gaussian family for the set of pair-copula family candidates. The sampling size for the exploration of $\boldsymbol{\Theta}$ in Method 1 is set to $6,000$. Experimental results are summarized over Figure 5.8, by displaying the minimum quantiles in function of the iteration k of Method 2. The quantile at independence is shown in dark line, the minimum estimated quantile from Method 1 is shown in blue line and the other lines are the minimum quantiles at each iteration of the algorithm, all with their 95% bootstrap confidence interval. We display at each iteration the minimum quantiles of each candidate pair in small dots.

The minimum result from Method 1 is even higher than the quantile at independence. This is due to the very large number of pairs ($p = 45$) that makes the exploration of $\boldsymbol{\Theta}$ extremely difficult. On the other hand, Method 2 (iterative algorithm) is definitely better and significantly

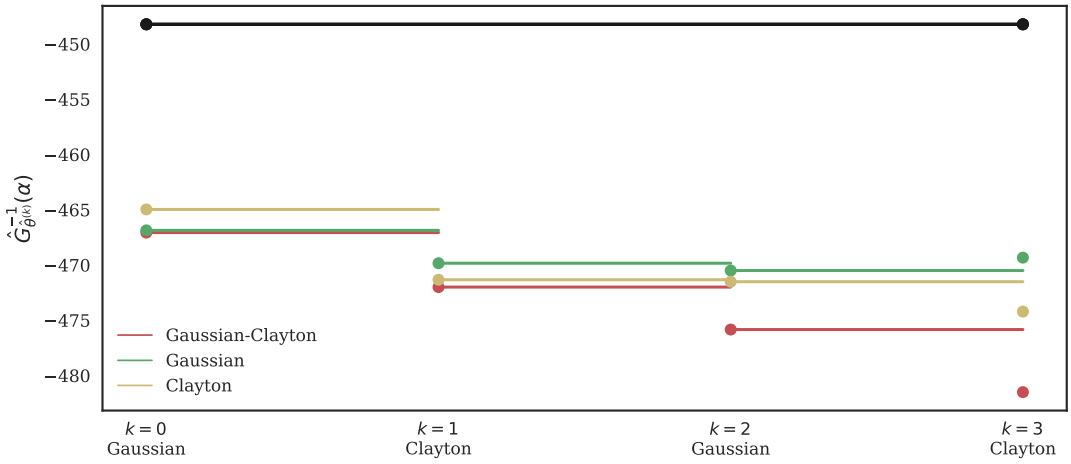


Figure 5.9 *Quantile minimization for different set of family candidates* $\mathbf{B}$. *The dark line shows the quantile at independence. The minimum at each iteration for the family candidates sets* $\mathbf{B}^1$, $\mathbf{B}^2$ *and* $\mathbf{B}^3$ *respectively in red, green and yellow.*

decreases the quantile value even at the first iteration (for only one dependent pair). The results are slightly improved with the iterations. We observe at the last iteration that the results from the candidate pairs are slightly higher than the minimum from the previous iteration. It is due to the choice of N_k which does not increases enough with the iterations to correctly explore $\mathbf{\Theta}_k$, which also increases with the iterations.

5.5.1.3 Using multiple pair-copula family candidates

To show the importance of testing multiple copula families, we consider $d = 6$ and three tests of Method 2 (iterative procedure). The Figure 5.9 shows the minimum from the iterative results using three sets of family candidates: a set of Gaussian and Clayton in red ($\mathbf{B}^1 = \{G, C\}$), Gaussian only in green ($\mathbf{B}^2 = \{G\}$), and Clayton only in yellow ($\mathbf{B}^3 = \{C\}$). We also display below the iteration number, the selected family for $\mathbf{B}^1$.

At iteration $k = 0$, the algorithm with the set $\mathbf{B}^1$ has selected the Gaussian copula as the selected pair and the result is as expected equivalent as for the set $\mathbf{B}^2$. At next iteration, a Clayton copula has been selected for algorithm with the set $\mathbf{B}^1$, which slightly improves the minimization compared to the others. The improvement start at iteration $k = 2$ where the Algorithm with the set $\mathbf{B}^1$ minimizes more the output quantile than the other sets with only one copula family. At the last iteration, the algorithm with set $\mathbf{B}^1$ selected a mix between Gaussian an Clayton families. This diversity seems to lead to better results than using only one family for every pairs. Testing multiple families is an interesting feature of the algorithm and is something that cannot be feasible for the grid-search approach. However, the cost for $\mathbf{B}^1$ is twice larger than for the other methods.

5.5.2 Industrial Application

5.5.2.1 Context

We consider an industrial component belonging to a production unit. This component must maintain its integrity even in case of an accidental situation. For this reason, it is subject to a justification procedure by regulation authorities, in order to demonstrate its ability to withstand severe operating conditions. This undesirable event consists in the concomitance of three different factors:

- the potential existence of small and undetectable manufacturing defects ;
- the exposition of the structure to an ageing phenomenon harming the material which progressively diminishes its mechanical resistance throughout its lifespan ;
- the occurrence of an accidental event generating severe constraints on the structure.

If combined, these three factors might lead to the initiation of a crack within the structure. Since no failure was observed until now, a structural reliability study should be conducted to check the safety of the structure. To do so a thermal-mechanical code $\eta : \mathbb{R}^d \to \mathbb{R}^+$ was used, which calculates the ratio between the resistance and the stress acting on the component during a simulated accident. The numerical model depends on parameters affected by uncertainties quantified throughout numerous mechanical tests. Nevertheless, these experiments are mostly established individually and only few experiments involves simultaneously two parameters.

5.5.2.2 Probabilistic model

For this problem, we introduce $d = 6$ random variables with predefined marginal distributions $(P_j)_{j=1\ldots d}$. The dependence structure is however unknown. From the 15 pairs of variables, only the dependencies of two pairs are known: one is independent and the other follows a Gumbel copula with parameter 2.27. Therefore, we consider $p = 13$ pairs of variables with unknown dependencies.

Given expert feedbacks, we restricted the exploration space $\mathbf{\Theta}$ by defining bounds for each pair of variables $(i, j) \in \Omega$ such that

$$T_{c_{i,j}}(\tau_{i,j}^-) \leq \theta_{i,j} \leq T_{c_{i,j}}(\tau_{i,j}^+),$$

where $\tau_{i,j}^-$ and $\tau_{i,j}^+$ are respectively the upper and lower kendall's correlation coefficient bounds for the dependence of the pair (i, j) and $T_{c_{i,j}}$ is the transformation from Kendall's tau value to the copula parameter for the associated copula $c_{i,j}$. This choice enables to explore only realistic dependence structures. For these experiments we only considered Gaussian copulas.

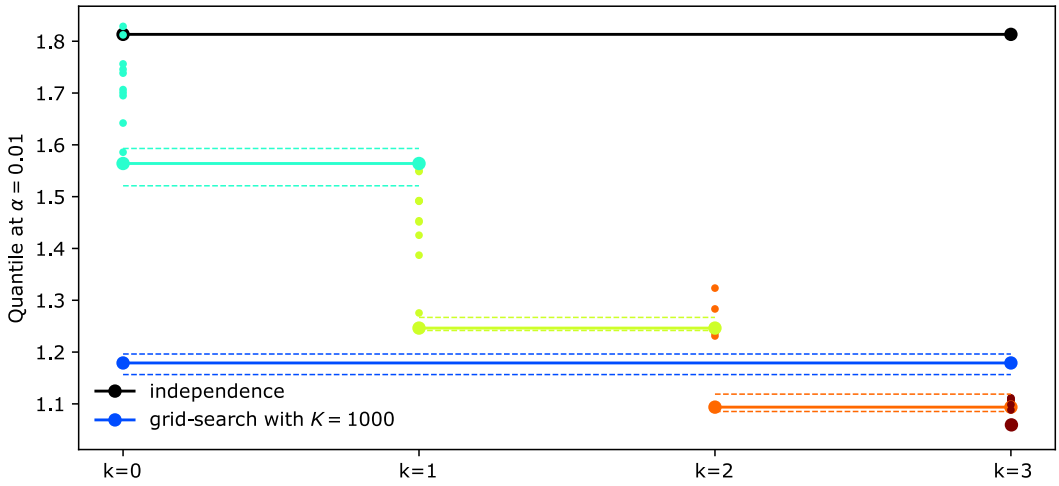


Figure 5.10 *Minimization of the output quantile using a grid-search and the iterative procedure for $\alpha = 1\%$. The description is the same as in Figure 5.8.*

5.5.2.3 Results

We consider the quantile at $\alpha = 0.01$ as a quantity of interest. A first experiment is established with the incomplete probability structure: only the two pairs with a known dependence structure and all others at independence. Two other experiments are established: an exhaustive grid-search approach with a given vine structure and an iterative procedure with a maximum budget equivalent to the grid-search. A grid-size of 1000 is chosen with $n = 20,000$.

The results are displayed in Figure 5.10. and has the same description as Figure 5.8. The quantile for the incomplete probability structure is approximately at 1.8. The grid-search and the iterative approaches found dependence structures leading to output quantile values close to 1.2 and 1.1 respectively. The minimum quantile from the iterative procedure is slightly lower than the grid-search approach. The problem dimension is not big enough to create make a significant difference between the methods. However, the resulting dependence structure from the iterative method is greatly simplified with only four pairs of variables, in addition to the already known pair.

This result highlights the risk of having an incomplete dependence structure in a reliability problem. In this application, the critical limit (*safety margin*) of the considered industrial component is 1. With the incomplete distribution of $\mathbf{X}$, the output quantile is very high compared to the critical limit and states a high reliability of the component. Unfortunately, if we consider worst-case dependence structures, the output quantile is significantly minimized and becomes closer to the critical limit. Thus, if the true dependence structure is close to the obtained worst case dependence structure, the risk of over estimating the output quantile can be important.

5.6 Conclusion and discussion

Incomplete information on inputs is an issue frequently encountered in structural reliability. Because safety analyses are mostly based on propagating random uncertainties through black-box computer models, the selection of a conservative dependence structure between input components appears as a requirement to define probabilistic worst cases. This article takes a first step towards such a methodology, by proposing a greedy, heuristic algorithm that explores a set of possible dependencies, taking advantage of the pair-copula construction (PCC) of multivariate probability distributions. Results of experiments conducted on toy and a real models illustrate the good behavior of the procedure: in situations where the monotonicity of the considered risk indicator (the output quantile) with respect to the inputs is postulated, a minimum value for the risk indicator is obtained using Fréchet-Hoeffding bounds. Nonetheless, it is possible to exhibit situations where the algorithm detect other and more conservative dependence structures. This first step required a number of hypotheses and approximations that pave the way for future research. Besides, some perspectives arise from additional technical results.

It would be interesting to improve the statistical estimation of the minimum quantile, given a dependence structure, by checking the hypotheses underlying the convergence results of Theorem 3. Checking and relaxing these hypotheses should be conducted in relation with expert knowledge on the computer model η and, possibly, a numerical exploration of its regularity properties. The grid search estimation strategy promoted in Section 5.2.3 arises from the lack of information about the convexity and the gradient of $\boldsymbol{\theta} \to G_{\boldsymbol{\theta}}^{-1}(\alpha)$. However, the method remains basic and stochastic recursive algorithms, such as the Robbins-Munro algorithm [Robbins and Monro, 1951], can be proposed and tested as possibly more powerful (faster) alternatives. Moreover, the empirical quantile was considered for theoretical reasons but can be costly for large α. Another estimator could be consider to reduce the estimation cost. Methods like importance sampling (see for instance Johns [1988]) are commonly used to estimate small failure probabilities but its use can be more challenging in the case of quantile estimation.

A significant issue, is the computational cost of the exploration of possible dependence structures. Reducing this cost while increasing the completeness of this exploration should be a main concern of future works. Guiding the exploration in the space of conditional bivariate copulas using enriching criteria and possible expert knowledge can facilitate the minimization. The Algorithm 7 can also be improved using nonparametric bootstrap. This would quantify the estimation quality of the selected minimum quantile of each iteration. Note however that a seducing feature of an iterative procedure is the a priori possibility of its adaptation to situations where the computational model η is time-consuming. In such cases, it is likely that Bayesian global optimization methods based on replacing the computer model by a surrogate model (e.g., a kriging-based meta-model) [Osborne et al., 2009] should be explored, keeping in mind that non-trivial conservative correlations – losses of quantile monotonicity– can be due to edge effects (e.g., discontinuities) characterizing the computational model itself.

We noticed in our experiments on real case-studies that expert knowledge remains difficult to incorporate otherwise that using association and concordance measures, mainly since we are lacking of representation tools (e.g., visual) of the properties of multivariate laws that provide intelligible diagnostics. A first step towards the efficient incorporation of expert knowledge could

be to automatize the visualization of the obtained vine structures, to simplify judgements about their realism.

Finally, another approach to consider could be to address the optimization problem (5.2) within the more general framework of optimal transport theory, and to take advantage of the many ongoing works in this area. Indeed, the problem (5.2) can be seen as a multi-marginal optimal transport problem (see Pass, Brendan [2015] for an overview). When $d = 2$, it corresponds respectively to the classical optimal transport problems of Monge and Kantorovich [Villani, 2008]. However, the multimarginal theory is not as well understood as for the bimarginal case, and developing efficient algorithms for solving this problem remains also a challenging issue [Pass, Brendan, 2015].

Acknowledgments

The authors express their thanks to Gerard Biau (Sorbonne Université), Bertrand Iooss and Roman Sueur (EDF R&D) for their useful comments and advice during the writing of this article.

5.7 Proof of the consistency result

The consistency of the estimator $\hat{\boldsymbol{\theta}}$ requires some regularity of the function $\boldsymbol{\theta} \mapsto G_{\boldsymbol{\theta}}$. This regularity can be also expressed in term of modulus of increase of the function $\boldsymbol{\theta} \mapsto G_{\boldsymbol{\theta}}^{-1}(\alpha)$, on which some useful definitions and connections with the modulus of continuity are reminded.

5.7.1 Modulus of increase of a cumulative distribution function

Let us recall that a modulus of continuity is any real-extended valued function $\omega : [0, \infty) \mapsto [0, \infty)$ such that $\lim_{\delta \to 0} \omega(x) = \omega(0) = 0$. The function $f : \mathbb{R} \mapsto \mathbb{R}$ admits ω as modulus of continuity if for any $(x, x') \in \mathbb{R}^2$,

$$|f(x) - f(x')| \leq \omega(|x - x'|).$$

Similarly, for some $x \in \mathbb{R}$, the function f admits ω as a *local* modulus of continuity if for any $x' \in \mathbb{R}^2$,

$$|f(x) - f(x')| \leq \omega(|x - x'|).$$

To control the deviation of the empirical quantile in the proof of Proposition 3 further, we consider the modulus of continuity of the quantile functions $G^{-1} : [0, 1] \to \mathbb{R}$ where G is a distribution function on $\mathbb{R}$. The quantile function being an increasing function, the *exact local modulus of continuity* of the quantile function G^{-1} at $\alpha \in (0, 1)$ can be defined as

$$\omega_{G^{-1}}(\epsilon, \alpha) := \max \left(G^{-1}(\alpha + \epsilon) - G^{-1}(\alpha), G^{-1}(\alpha) - G^{-1}(\alpha - \epsilon) \right).$$

In the proof of Proposition 3, we note that the continuity of a quantile function G^{-1} can be connected to the increase of the distribution function G (see also for instance Section A in Bobkov

and Ledoux [2014]). Using the fact that the distribution function is increasing, we introduce the *local modulus of increase* of the distribution function ϵ_G at $y = G^{-1}(\alpha) \in \mathbb{R}$ as:

$$\epsilon_G(\delta, y) := \min\left(G(y+\delta) - G(y), G(y) - G(y-\delta)\right).$$

5.7.2 Proofs

The estimator $\hat{\boldsymbol{\theta}}$ defined in (5.12) is an extremum-estimator (see for instance Section 2.1 of Newey and McFadden [1994]). The main ingredient to prove the consistency of this estimator is the uniform convergence in probability of the families of the empirical quantiles $(\widehat{G}_{\boldsymbol{\theta}}^{-1}(\alpha))_{\boldsymbol{\theta}\in\mathcal{D}_{K_n}}$ over the family of grids $\mathcal{D}_{K_n}$.

Proposition 3. *Let $\mathcal{D}_{K_n}$ be defined as in Theorem 3. Let assume that* **B** *and* **C** *are both satisfied. Then, for all $\epsilon > 0$,*

$$\mathbb{P}[\sup_{\boldsymbol{\theta}\in\mathcal{D}_{K_n}} |\widehat{G}_{\boldsymbol{\theta}}^{-1}(\alpha) - G_{\boldsymbol{\theta}}^{-1}(\alpha)| > \epsilon] \xrightarrow{n\to\infty} 0.$$

Proof of Proposition 3. We first make the connection between the local continuity of the quantile function $G_{\boldsymbol{\theta}}^{-1}$ and the local increase of the distribution function $G_{\boldsymbol{\theta}}$. According to Assumption **C**, we have that for any $\epsilon \in (0, \max((1-\alpha), \alpha))$, for any $\delta > 0$ and for any $\boldsymbol{\theta} \in \mathcal{D}$,

$$(*) : \begin{cases} G_{\boldsymbol{\theta}}^{-1}(\alpha+\epsilon) - G_{\boldsymbol{\theta}}^{-1}(\alpha) < \delta \\ G_{\boldsymbol{\theta}}^{-1}(\alpha) - G_{\boldsymbol{\theta}}^{-1}(\alpha-\epsilon) < \delta \end{cases} \Longrightarrow \begin{cases} G_{\boldsymbol{\theta}}\left(G_{\boldsymbol{\theta}}^{-1}(\alpha+\varepsilon)\right) < G_{\boldsymbol{\theta}}\left(G_{\boldsymbol{\theta}}^{-1}(\alpha)+\delta\right) \\ G_{\boldsymbol{\theta}}\left(G_{\boldsymbol{\theta}}^{-1}(\alpha)-\delta\right) < G_{\boldsymbol{\theta}}\left(G_{\boldsymbol{\theta}}^{-1}(\alpha-\varepsilon)\right) \end{cases}.$$

Next, using basic properties of quantile functions (see for instance point **ii** of Lemma 21.1 in Van der Vaart [2000]) together with Assumption **B**, we find that

$$G_{\boldsymbol{\theta}}\left(G_{\boldsymbol{\theta}}^{-1}(\alpha+\epsilon)\right) = \alpha + \varepsilon = G_{\boldsymbol{\theta}}\left(G_{\boldsymbol{\theta}}^{-1}(\alpha)\right) + \epsilon$$

and

$$G_{\boldsymbol{\theta}}\left(G_{\boldsymbol{\theta}}^{-1}(\alpha-\epsilon)\right) = \alpha - \varepsilon = G_{\boldsymbol{\theta}}\left(G_{\boldsymbol{\theta}}^{-1}(\alpha)\right) - \epsilon.$$

Thus,

$$(*) \Longrightarrow \begin{cases} G_{\boldsymbol{\theta}}\left(G_{\boldsymbol{\theta}}^{-1}(\alpha)+\delta\right) - G_{\boldsymbol{\theta}}\left(G_{\boldsymbol{\theta}}^{-1}(\alpha)\right) > \epsilon \\ G_{\boldsymbol{\theta}}\left(G_{\boldsymbol{\theta}}^{-1}(\alpha)\right) - G_{\boldsymbol{\theta}}\left(G_{\boldsymbol{\theta}}^{-1}(\alpha)-\delta\right) > \epsilon \end{cases}.$$

We have shown that any $\epsilon \in (0, \max((1-\alpha), \alpha))$, for any $\delta > 0$ and for any $\boldsymbol{\theta} \in \mathcal{D}$,

$$\omega_{G_{\boldsymbol{\theta}}^{-1}}(\epsilon, \alpha) > \delta \Longrightarrow \epsilon_{G_{\boldsymbol{\theta}}}(\delta, G_{\boldsymbol{\theta}}^{-1}(\alpha)) < \epsilon. \tag{5.27}$$

We now prove the proposition. For any $n \geq 1$ and any $\epsilon > 0$, we have

$$\begin{aligned} \mathbb{P}\left(\sup_{\boldsymbol{\theta}\in\mathcal{D}_{K_n}} |\widehat{G}_{\boldsymbol{\theta}}^{-1}(\alpha) - G_{\boldsymbol{\theta}}^{-1}(\alpha)| > \epsilon\right) &= \mathbb{P}\left(\bigcup_{\boldsymbol{\theta}\in\mathcal{D}_{K_n}} \{|\widehat{G}_{\boldsymbol{\theta}}^{-1}(\alpha) - G_{\boldsymbol{\theta}}^{-1}(\alpha)| > \epsilon\}\right) \\ &\leq \sum_{\boldsymbol{\theta}\in\mathcal{D}_{K_n}} P_{\boldsymbol{\theta}}\left(|\widehat{G}_{\boldsymbol{\theta}}^{-1}(\alpha) - G_{\boldsymbol{\theta}}^{-1}(\alpha)| > \epsilon\right). \end{aligned} \tag{5.28}$$

Let $\xi_1, \dots, \xi_n$ be n i.i.d. uniform random variables. The uniform empirical distribution function is defined by

$$\mathbb{U}(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\xi_i \leq t} \quad \text{for } 0 \leq t \leq 1.$$

The inverse uniform empirical distribution function is the function

$$\mathbb{U}_n^{-1}(u) = \inf\{t \,|\, \mathbb{G}_n(t) > u\} \quad \text{for } 0 \leq u \leq 1.$$

The empirical distribution function $\widehat{G}_{\boldsymbol{\theta}}$ can be rewritten as (see for instance Van der Vaart [2000]):

$$\widehat{G}_{\boldsymbol{\theta}}(y) \stackrel{\mathcal{L}}{=} \mathbb{U}_n(G_{\boldsymbol{\theta}}(y))$$

and as well for the quantile function,

$$\widehat{G}_{\boldsymbol{\theta}}^{-1}(\alpha) \stackrel{\mathcal{L}}{=} G_{\boldsymbol{\theta}}^{-1}(\mathbb{U}_n^{-1}(\alpha)).$$

From Inequality (5.28), we obtain

$$\sum_{\boldsymbol{\theta} \in \mathcal{D}_{K_n}} P_{\boldsymbol{\theta}} \left(|\widehat{G}_{\boldsymbol{\theta}}^{-1}(\alpha) - G_{\boldsymbol{\theta}}^{-1}(\alpha)| > \epsilon \right) = \sum_{\boldsymbol{\theta} \in \mathcal{D}_{K_n}} P_{\boldsymbol{\theta}} \left(|G_{\boldsymbol{\theta}}^{-1}(\mathbb{U}_n^{-1}(\alpha)) - G_{\boldsymbol{\theta}}^{-1}(\alpha)| > \epsilon \right) \tag{5.29}$$

By definition of the local modulus of continuity $\omega_{G_{\boldsymbol{\theta}}^{-1}}$ of the quantile function $G_{\boldsymbol{\theta}}^{-1}$ at α, we have

$$|G_{\boldsymbol{\theta}}^{-1}(\mathbb{U}_n^{-1}(\alpha)) - G_{\boldsymbol{\theta}}^{-1}(\alpha)| \leq \omega_{G_{\boldsymbol{\theta}}^{-1}}(|\mathbb{U}_n^{-1}(\alpha) - \alpha|, \alpha). \tag{5.30}$$

Therefore, by replacing (5.30) in (5.29) and using (5.27), we obtain

$$\begin{aligned} \sum_{\boldsymbol{\theta} \in \mathcal{D}_{K_n}} P_{\boldsymbol{\theta}} \left(\left| G_{\boldsymbol{\theta}}^{-1}(\mathbb{U}_n^{-1}(\alpha)) - G_{\boldsymbol{\theta}}^{-1}(\alpha) \right| > \epsilon \right) &\leq \sum_{\boldsymbol{\theta} \in \mathcal{D}_{K_n}} P_{\boldsymbol{\theta}} \left(\omega_{G_{\boldsymbol{\theta}}^{-1}}(|\mathbb{U}_n^{-1}(\alpha) - \alpha|, \alpha) > \epsilon \right) \\ &\leq \sum_{\boldsymbol{\theta} \in \mathcal{D}_{K_n}} P_{\boldsymbol{\theta}} \left(\epsilon_{G_{\boldsymbol{\theta}}}(\epsilon, G_{\boldsymbol{\theta}}^{-1}(\alpha)) < |\mathbb{U}_n^{-1}(\alpha) - \alpha| \right). \end{aligned}$$

Assumption $\mathbf{C}$ then yields

$$\mathbb{P}\left(\sup_{\boldsymbol{\theta} \in \mathcal{D}_{K_n}} |\hat{G}_{\theta}^{-1}(\alpha) - G_{\boldsymbol{\theta}}^{-1}(\alpha)| > \epsilon \right) \leq K_n \mathbb{P}\left(|\mathbb{U}_n^{-1}(\alpha) - \alpha| > \underline{\epsilon}_{\mathcal{D}}(\epsilon) \right). \tag{5.31}$$

The DKW inequality [Dvoretzky et al., 1956] gives an upper bound of the probability of an uniform empirical process $\{|\mathbb{U}_n(\alpha) - \alpha|\}$. As well for an uniform empirical quantile process $\{|\mathbb{U}_n^{-1}(\alpha) - \alpha|\}$ (see for example Section 1.4.1 of Csorgo [1983]), such as $\forall \lambda > 0$:

$$\mathbb{P}(\sup_{\alpha \in [0,1]} |\mathbb{U}_n^{-1}(\alpha) - \alpha|) \geq \lambda) \leq C \exp(-2n\lambda^2).$$

Moreover, Massart [1990] proved that one can take $C = 2$. Therefore, Equation (5.31) can be bounded using the DKW and

$$\mathbb{P}\left(\sup_{\boldsymbol{\theta} \in \mathcal{D}_{K_n}} |\widehat{G}_{\boldsymbol{\theta}}^{-1}(\alpha) - G_{\boldsymbol{\theta}}^{-1}(\alpha)| > \epsilon \right) \leq 2K_n \exp\left[-2n\underline{\epsilon}_{\mathcal{D}}^2(\epsilon) \right] \xrightarrow{n \to \infty} 0$$

since $K_N \lesssim n^{\beta}$. □

A second requirement to get the consistency of the extremum estimator is the regularity of $\boldsymbol{\theta} \mapsto G_{\boldsymbol{\theta}}^{-1}(\alpha)$. This is shown in the next proposition.

Proposition 4. *Under Assumptions* **A**, **B** *and* **C**, *the function*

$$\begin{aligned} \mathcal{D} &\longrightarrow \overline{Im(\eta)} \\ \boldsymbol{\theta} &\longrightarrow G_{\boldsymbol{\theta}}^{-1}(\alpha) \end{aligned}$$

is continuous in $\boldsymbol{\theta}$ *over* $\mathcal{D}$.

Proof of Proposition 4. According to Assumption **A**, for any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, the distribution $P_{\boldsymbol{\theta}}$ admits a density function $f_{\boldsymbol{\theta}}$ with respect to the Lebesgue measure on $\mathbb{R}^d$ such that

$$f_{\boldsymbol{\theta}}(x_1, \ldots, x_d) = c_{\boldsymbol{\theta}}\left(F_1(x_1, \ldots, x_d)\right) f_1(x_1) \ldots f_d(x_d),$$

where f_j is the marginal density function of X_j, for $j = 1, \ldots, d$ and the Lebesgue measure on $\mathbb{R}$. Moreover, for any $\mathbf{x} \in \mathbb{R}^d$, the function $\boldsymbol{\theta} \to f_{\boldsymbol{\theta}}(x)$ is continuous in $\boldsymbol{\theta}$ over $\mathcal{D}$.

The domain $\mathcal{D} \times [0,1]^p$ is a compact set and according to Assumption **A**, there exists a constant $\bar{c}$ such that $\forall (\boldsymbol{\theta}, \mathbf{u}) \in \mathcal{D} \times [0,1]^d$, $c_{\boldsymbol{\theta}}(\mathbf{u}) \leq \bar{c}$. Consequently, we have

$$|f_{\boldsymbol{\theta}}(x_1, \ldots, x_d)| \leq \bar{c} \prod_{i=1}^{d} f_i(x_i). \tag{5.32}$$

For $\boldsymbol{\theta} \in \mathcal{D}$ and for any $h > 0$, we denote $y_h = G_{\boldsymbol{\theta}+h}^{-1}(\alpha)$. According to Assumption **B** we have $\alpha = G_{\boldsymbol{\theta}+h}(y_h)$ and thus,

$$\begin{aligned} G_{\boldsymbol{\theta}}^{-1}(\alpha) - G_{\boldsymbol{\theta}+h}^{-1}(\alpha) &= G_{\boldsymbol{\theta}}^{-1}(\alpha) - y_h \\ &= G_{\boldsymbol{\theta}}^{-1}(G_{\boldsymbol{\theta}+h}(y_h)) - G_{\boldsymbol{\theta}}^{-1}(G_{\boldsymbol{\theta}}(y_h)) \end{aligned} \tag{5.33}$$

Now, using Assumption **C**, we have that $G_{\boldsymbol{\theta}}$ is strictly increasing in the neighborhood of $G_{\boldsymbol{\theta}}^{-1}(\alpha)$ and thus $G_{\boldsymbol{\theta}}^{-1}$ is continuous in the neighborhood of α. Note that

$$\begin{aligned} |G_{\boldsymbol{\theta}+h}(y_h) - G_{\boldsymbol{\theta}}(y_h)| &= \left| \int_{\mathbb{R}^d} \mathbb{1}_{\eta(\mathbf{x}) \leq y_h} \mathrm{d}F_{\boldsymbol{\theta}+h}(\mathbf{x}) - \int_{\mathbb{R}^d} \mathbb{1}_{\eta(\mathbf{x}) \leq y_h} \mathrm{d}F_{\theta}(\mathbf{x}) \right| \\ &= \left| \int_{\mathbb{R}^d} [f_{\boldsymbol{\theta}+h}(\mathbf{x}) - f_{\boldsymbol{\theta}}(\mathbf{x})] \mathbb{1}_{\eta(\mathbf{x}) \leq y_h} \mathrm{d}\lambda(\mathbf{x}) \right| \\ &\leq \int_{\mathbb{R}^d} |f_{\boldsymbol{\theta}+h}(\mathbf{x}) - f_{\boldsymbol{\theta}}(\mathbf{x})| \mathrm{d}\lambda(\mathbf{x}) \end{aligned}$$

We then apply a standard dominated convergence theorem using (5.32) to get that

$$G_{\boldsymbol{\theta}+h}(y_h) - G_{\boldsymbol{\theta}}(y_h) \xrightarrow{h \to 0} 0.$$

This, with (5.33) and with the continuity of $\boldsymbol{\theta} \mapsto G_{\boldsymbol{\theta}}$, shows that

$$G_{\boldsymbol{\theta}}^{-1}(\alpha) - G_{\boldsymbol{\theta}+h}^{-1}(\alpha) \xrightarrow{h \to 0} 0.$$

□

We are now in position to prove Theorem 3.

Proof of Theorem 3. Under Assumptions **B** and **C**, Proposition 3 directly gives that for any $\varepsilon > 0$,

$$P\left(\left|\inf_{\theta\in\mathcal{D}_{K_n}} \widehat{G}_{\boldsymbol{\theta}}^{-1}(\alpha) - \inf_{\theta\in\mathcal{D}_{K_n}} G_{\boldsymbol{\theta}}^{-1}(\alpha)\right| > \varepsilon\right) \xrightarrow{n\to\infty} 0$$

which means that

$$P\left(\left|\widehat{G}_{\hat{\boldsymbol{\theta}}}^{-1}(\alpha) - \inf_{\boldsymbol{\theta}\in\mathcal{D}_{K_n}} G_{\boldsymbol{\theta}}^{-1}(\alpha)\right| > \varepsilon\right) \xrightarrow{n\to\infty} 0. \tag{5.34}$$

If Assumption **A** is also satisfied, Proposition 4 together with (5.11) give that $\inf_{\boldsymbol{\theta}\in\mathcal{D}_{K_n}} G_{\boldsymbol{\theta}}^{-1}(\alpha)$ tends to $\inf_{\theta\in\mathcal{D}} G_{\boldsymbol{\theta}}^{-1}(\alpha)$ as n tends to infinity. Thus

$$\inf_{\boldsymbol{\theta}\in\mathcal{D}_{K_n}} G_{\theta}^{-1}(\alpha) \xrightarrow{n\to\infty} G_C^{-1\star}(\alpha) = G_{\theta_C^*}^{-1}(\alpha) \tag{5.35}$$

We then derive (5.13) from (5.34) and (5.35).

We now assume that Assumption **D** is also satisfied. Let $\boldsymbol{\theta}^*$ be the unique minimizer of $\boldsymbol{\theta} \mapsto G_{\boldsymbol{\theta}}^{-1}(\alpha)$. Let $h > 0$ such that $B(\boldsymbol{\theta}^*, h)^c := \{\boldsymbol{\theta} \in \mathcal{D} \ : \ \|\boldsymbol{\theta} - \boldsymbol{\theta}^\star\|_2 \geq \varepsilon\}$ is not empty. According to Proposition 4 and using the fact that $\mathcal{D}$ is compact, we have

$$\sup_{\boldsymbol{\theta}\in B(\theta^*,h)^c} |G_{\boldsymbol{\theta}}^{-1}(\alpha) - G_{\boldsymbol{\theta}^*}^{-1}(\alpha)| > 0. \tag{5.36}$$

Consequently, for any $\forall h > 0$ small enough, there exists $\epsilon > 0$ such that

$$|G_{\boldsymbol{\theta}}^{-1}(\alpha) - G_{\theta^*}^{-1}(\alpha)| \leq \epsilon \Longrightarrow |\boldsymbol{\theta} - \theta^*| < h \tag{5.37}$$

Let $h > 0$ and take ϵ such that (5.37) is satisfied for h. According to Proposition 3, $\widehat{G}_{\hat{\theta}}^{-1}(\alpha) - G_{\hat{\theta}}^{-1}(\alpha)$ tends to zero in probability as n tends to infinity. This, with (5.13), shows that

$$P\left(\left|G_{\hat{\theta}}^{-1}(\alpha) - G_{\theta^*}^{-1}(\alpha)\right| > \varepsilon\right) \xrightarrow{n\to\infty} 0.$$

We conclude using (5.37). □

5.7.3 Vine copulas

5.7.3.1 Definition

A vine model describes a d-dimensional pair-copula construction (PCC) and is a sequence of linked trees where the nodes and edges correspond to the $d(d-1)/2$ pair-copulas. According to Definition 6 from Bedford and Cooke [2001], a vine structure is composed of $d-1$ trees $T_1, \ldots, T_{d-1}$ with several conditions.

Definition 6 (R-vine)**.** *The sequence $\mathcal{V} = (T_1, \ldots, T_{d-1})$ is an R-vine on n elements if*

1. *T_1 is a tree with nodes $N_1 = \{1, \dots, d\}$ and a set of edges denoted E_1.*
2. *For $i = 2, \dots, d-1$, T_i is a tree with nodes $N_i = E_{i-1}$ and edges set E_i.*
3. *For $i =, \dots, d-1$ and $\{a, b\} \in E_i$ with $a = \{a_1, a_2\}$ and $b = \{b_1, b_2\}$ it must hold that $\#(a \cap b) = 1$ (proximity condition).*

Each tree T_i is composed of $d-i+1$ nodes which are linked by $d-i$ edges for $i = 1, \dots, d-1$. A node in a tree T_i must be an edge in the tree T_{i-1}, for $i = 2, \dots, d-1$. Two nodes in a tree T_i can be joined if their respective edges in tree T_{i-1} share a common node, for $i = 2, \dots, d-1$. The proximity condition, suggests that two nodes connected by an edge should share one variable from the conditioned set. The *conditioning set* and *conditioned set* are defined in Definition 7 along with the *complete union*. The complete union of an edge e is a set of all unique variables contained in e.

Definition 7 (Complete union, conditioning and conditioned sets of an edge)**.** *Let A_e be the complete union of an edge $e = \{a, b\} \in E_k$ in a tree T_k of a regular vine $\mathcal{V}$,*

$$A_e = \{v \in N_1 | \exists e_i \in E_i, i = 1, \dots, k-1, \text{such that } v \in e_i \in \cdots \in e_{k-1} \in e\}.$$

The conditioning set associated with edge $e = \{a, b\}$ is $D(e) := A_a \cap A_b$ and the conditioned sets associated with edge e are $i(e) := A_a \backslash D(e)$ and $j(e) := A_b \backslash D(e)$. Here, $A \backslash B := A \cap B^c$ and B^c is the complement of B.

The conditioned and conditioning sets of an edge $e = \{a, b\}$ are respectively the symmetric difference and the intersection of the complete unions of a and b. The conditioned and conditioning sets of all edges of $\mathcal{V}$ are collected in a set called *constraint set*. Each element of this set is composed of a pair of indices corresponding to the conditioned set and a set containing indices corresponding to the conditioning set, as shown in Definition 8.

Definition 8 (Constraint set)**.** *The constrain set for $\mathcal{V}$ is a set:*

$$\mathcal{CV} = \{(\{i(e), j(e)\}, D_e) | e \in E_i, e = \{a, b\}, i = 1, \dots, d-1\}$$

The pair-copula in the first tree characterize pairwise unconditional dependencies, while the pair-copula in higher order trees model the conditional dependency between two variables given a set of variables. The number of conditioning variables grows with the tree order. Note that a PCC where all trees have a path-like structure define the D-vine subclass while the star-like structures correspond to C-vine subclass. All other vine structures are called regular vines (R-vines) Bedford and Cooke [2001].

We illustrate the concept of a vine model with a $d = 5$ dimensional example. For clarity reasons, we use the same simplifications as in Section 5.4.2 which consider for instance $f_1 =$

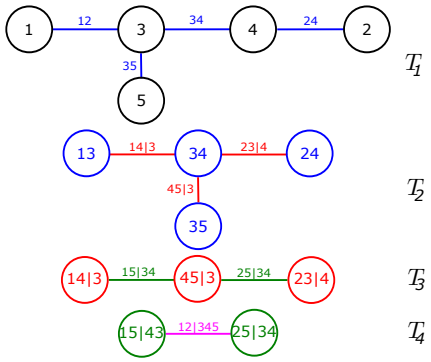


Figure 5.11 *R-vine structure for* $d = 5$.

$f_1(x_1)$, $f_2 = f_2(x_2)$ and so on for higher order and conditioning. One possible PCC can be written for this 5-dimensional configuration:

$$\begin{aligned} f(x_1, x_2, x_3, x_4, x_5) &= f_1 \cdot f_2 \cdot f_3 \cdot f_4 \cdot f_5 \text{ (margins)} \\ &\text{(unconditional pairs)} \times c_{12} \cdot c_{35} \cdot c_{34} \cdot c_{24} \\ &\text{(1st conditional pair)} \times c_{14|3} \cdot c_{23|4} \cdot c_{45|3} \\ &\text{(2nd conditional pair)} \times c_{15|34} \cdot c_{25|34} \\ &\text{(3rd conditional pair)} \times c_{12|345}. \end{aligned} \tag{5.38}$$

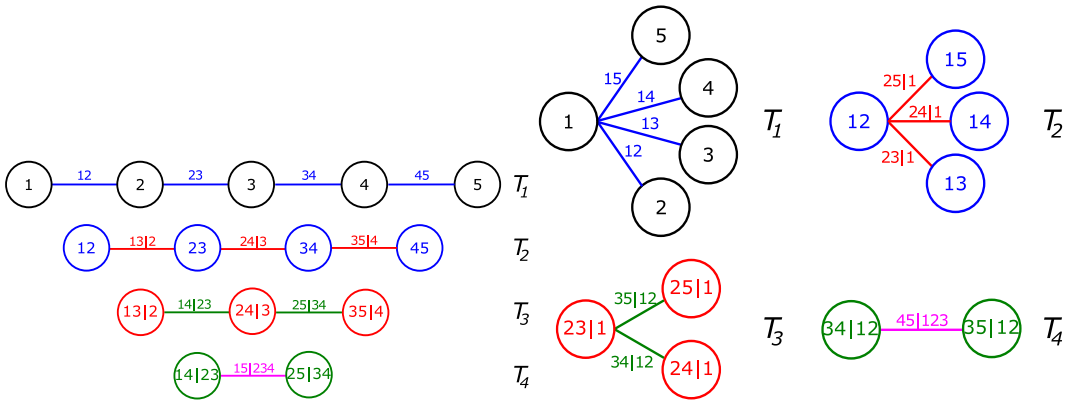
The vine structure associated to (5.38) is illustrated in Figure 5.11. This graphical model considerably simplify the understanding and we observe that this model is a R-vine because there is no specific constraints on the trees.

A re-labeling of the variables can lead to a large number of different PCC. Morales-Nápoles [2011] calculated the number of possible vine structures with the dimension d and shows that it becomes extremely large for high dimension problems. We illustrate below, using the same $d = 5$ dimensional example, two other PCC densities:

$$\begin{aligned} f_D &= f_1 \cdot f_2 \cdot f_3 \cdot f_4 \cdot f_5 \\ &\times c_{12} \cdot c_{23} \cdot c_{34} \cdot c_{45} \\ &\times c_{13|2} \cdot c_{24|3} \cdot c_{35|4} \\ &\times c_{14|23} \cdot c_{25|34} \\ &\times c_{15|234} \end{aligned} \tag{5.39}$$

$$\begin{aligned} f_C &= f_1 \cdot f_2 \cdot f_3 \cdot f_4 \cdot f_5 \\ &\times c_{12} \cdot c_{13} \cdot c_{14} \cdot c_{15} \\ &\times c_{23|1} \cdot c_{24|1} \cdot c_{25|1} \\ &\times c_{34|12} \cdot c_{35|12} \\ &\times c_{45|123} \end{aligned} \tag{5.40}$$

where (5.39) and (5.40) respectively correspond to D-vine and C-vine structures and are represented in Figures 5.12a and 5.12b. As we can see



(a) *D-vine structure for* $d = 5$. (b) *C-vine structure for* $d = 5$.

Figure 5.12 *D-vine and C-vine structure for* $d = 5$.

in these examples, the D-vine have a constraint on each tree that gives a path-like arrangement of the nodes. The C-vine on the other hand only has one node connected to all others for each tree.

An efficient way to store the information of a vine structure is proposed in Morales Nápoles [2010] and is called a R-vine array. The approach uses the specification of a lower triangular matrix where the entries belong to $1, \ldots, d$. Such matrix representation allows to directly derive the tree structure (or equivalently the associated PCC distribution). For more details, see Morales Nápoles [2010].

5.7.3.2 Generating R-vine from an indexed list of pairs

The iterative procedure proposed in Section 5.4.3, described by Algorithm 7, minimizes the output quantile by iteratively determining the pairs of variables that influences the most the quantile minimization. At each iteration of the algorithm (step 1.a), a new vine structure is created by considering the list of influential pairs. The specificity of this vine creation is to consider the ranking of the list by placing the most influential pairs in the first trees of the R-vine. Thus, we describe in this section how to generate vine structure with the constraint of a given list of indexed pairs to fill in the structure.

5.7.3.2.1 The algorithm We consider the same notation as in Algorithm 7. Creating a vine structure from a given indexed list of pairs Ω_k is not straightforward. The difficulties come from respecting the ranking of Ω_k and the respect of the R-vine conditions. Indeed, the pairs cannot be append in the structure easily. The vine structure must respect these conditions, which can be sometime very restrictive. The procedure we proposed is detailed by the pseudo-code of Algorithm 8 and can be greatly simplified in these few key steps:

1. fill $\mathcal{V}$ with the list Ω_k,

2. fill $\mathcal{V}$ with a permutation of Ω_{-k},
3. if $\mathcal{V}$ is not a R-vine, then permute Ω_k and restart at step 1.

In step 1 and 2, the *filling* procedure, detailed in Algorithm 9, successively *adds* the pairs of a list in the trees of a vine structure. Adding a pair (i, j) in a tree T_l associates (i, j) with the conditioned set and determine a possible conditioning set D from the previous tree such as a possible edge is $i, j|D$.

In step 2, because the ordering of Ω_{-k} is not important in the filling of $\mathcal{V}$, the permutation of Ω_{-k} aims at finding a ranking such as $\mathcal{V}$ leads to a R-vine.

In step 3, when the previous step did not succeeded and the resulting $\mathcal{V}$ is not a R-vine structure, then the ranking of Ω_k is not possible and must be changed. The permutation of some elements of Ω_k must be done such as the ranking of the most influential pairs remains as close as possible to the initial one.

Algorithme 8 : Generating a vine structure from a given list of indexed pairs Ω_k

Data : Ω_k, d
Result : A vine structure $\mathcal{V}$.

$\Omega_k^{init} = \Omega_k$;
$k = 1$;
do
  /* initialize $\mathcal{V}$ with a first empty tree */
  $N_1 = (1, \ldots, d)$;
  $E_1 = ()$;
  $\mathcal{V} = ((N_1, E_1))$;
  /* filling $\mathcal{V}$ with the list of selected pairs Ω_k */
  $\mathcal{V} = \texttt{Fill}(\mathcal{V},\ \Omega_k,\ d)$; // See Algorithm 9
  /* determining a permutation of Ω_{-k} that fills $\mathcal{V}$ */
  for $\Omega_{-k}^{\pi} \in \pi(\Omega_{-k})$ **do**
    /* filling $\mathcal{V}$ with the candidate pairs Ω_{-k}^{π} */
    $\mathcal{V}_\pi = \texttt{Fill}(\mathcal{V},\ \Omega_{-k}^{\pi},\ d)$; // See Algorithm 9
    if $\mathcal{V}_\pi$ is a R-vine **then**
      /* a permutation worked → we quit the loop */
      break
  $\mathcal{V} = \mathcal{V}_\pi$;
  if $\mathcal{V}$ is not a R-vine **then**
    /* filling did not work → permute initial list Ω_k^{init} */
    Get Ω_k by inverting pairs of $(\Omega_k^{init}$;
    $k = k + 1$;
while $\mathcal{V}$ is not a R-vine;

5.7.3.2.2 Example For illustration, let's create a $d = 5$ dimensional vine structure with the given list of pairs $\Omega_k = ((1, 2), (1, 3), (2, 3), (4, 5), (2, 4), (1, 5))$ using Algorithm 8. Using the

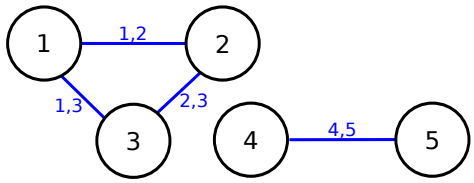


Figure 5.13 *Example: first tree of a non valid vine structure for $d = 5$ that does lead to a single connected tree.*

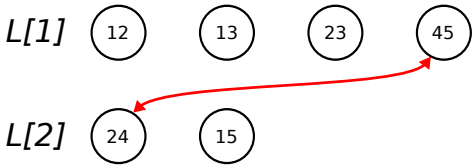


Figure 5.14 *Example: exchange of elements of Ω_k in order to lead to a valid vine structure.*

original list Ω_k, the `Fill` function may fail at line 7 of Algorithm 8, and more precisely, at line 15 of Algorithm 9. Indeed, the first tree of $\mathcal{V}$ does not validate the R-vine conditions. The tree is illustrated in Figure 5.13 and as we can see, the nodes are not all connected into one single tree. Therefore, we permuted the list Ω_k by exchanging the pairs $(2, 4)$ and $(4, 5)$, as shown in Figure 5.14. This permutation now leads to a vine structure that respects the new ranked list $\Omega_k = ((1, 2), (1, 3), (2, 3), (2, 4), (4, 5), (1, 5))$.

Algorithme 9 : Filling a vine structure with a given list

```
1  function Fill(V, Ω_k, d):
       /* V: an incomplete vine structure ,                                  */
       /* Ω_k:, a list of indexed pairs                                      */
       /* d: the input dimension.                                            */
2      l = |V|;                                          // number of existing trees
3      (T_1, ..., T_l) = V;
4      k = |T_l|;                                  // number of existing nodes in last tree
       /* loop over the list of pairs                                        */
5      for (i, j) ∈ Ω_k do
6          D = ∅;
7          if l >= 2 then
               /* conditioning set is only computed from T_2                 */
8              D = FindConditioningSet((i, j), N_{l-1});          // See Algorithm 10
9              if D = ∅ then
                   /* no conditioning set found → not possible               */
10                 return False
11         E_l = E_l ∪ i, j|D;                                  // add new edge in E_l
12         T_l = (N_l, E_l);                                    // update current tree
13         V = (T_1, ..., T_l);
14         if k ≥ d − l then
               /* if tree T_l is complete                                    */
15             if V does not fulfill the R-vine conditions then
                   /* the vine structure V is not valid                      */
16                 return False
17             k = 1;
18             l = l + 1;
19             N_l = E_{l-1};          // nodes of next tree are the edges of previous tree
20         else
21             k = k + 1;
22     return V
```

Algorithme 10 : Gets the conditioning set of a given conditioned set

```
1 function FindConditioningSet((i, j), N_):
      /* (i, j): the conditioned set,                                  */
      /* N_ : list of nodes from the previous tree.                    */
2     D = ∅;
3     for a, b ∈ N_, with a ≠ b do
4         if i ∈ a and j ∈ b then
5             if j ∉ A_a and i ∉ A_b then
                  /* See Definition 7                                 */
6                 D = A_a ∩ A_b;
7                 break;
8     return D
```

Chapter 6

Conclusion

6.1 Summary and main contributions

This work was intended to develop methodological tools address to reliability studies that aim to deal with dependencies. The two main themes covered in this manuscript are sensitivity analysis for dependent variables (when the dependence is known) and the assessment of a reliability risk (when the dependence is unknown). The first theme is covered in Chapter 3 and 4 and the second is treated in Chapter 5.

A state of the art of existing approaches and mathematical tools necessary for the whole thesis is provided in Chapter 2. We first presented the main sensitivity analysis methods in the case of independent and dependent inputs. We then introduced the notion of copula to model the dependence of probability distributions, as well as the Vine Copulas to model multivariate dependencies. Lastly, we presented two meta-modeling methods: random forests and the kriging models.

Chapter 3 first aims at studying the relations between the field of Sensitivity Analysis and Variable Importance in machine learning. Random forest algorithm is largely used in the machine learning field and provides Permutation based Variable Importance measures (PVI). When the true model η is considered, a first link between the total Sobol' indices and the PVI measures has been established [Gregorutti et al., 2015; Wei et al., 2015]. We extended the procedure to propose a new permutation based measure that is shown to measure the effect of the first-order Sobol' indices. In the case of dependent variables, the use of a Rosenblatt transformation can lead to the definition of new permutation values, which are equivalent to the full and independent Sobol' indices [Mara et al., 2015]. Using different numerical studies, we observed that the estimation of the Sobol' indices using permutation based values can be a suitable alternative the classical Monte Carlo procedure. When using an estimated model (such as a random forest model), the estimation of the PVI values is subject to a systematic bias. This chapter also aims at identifying this bias and studying its effect on the estimation of the PVI values. We observed that this bias is strongly related to the accuracy of the estimated model. Moreover, it increases significantly when some variables are correlated. In the latter case, the use of the PVI values using a Rosenblatt

transformation can be an alternative to deal with dependent variables. We observe that the systematic bias for these permutation values is less impacted by variable correlations. Moreover, they offer a different interpretation of the importance of variables that the classical PVI values.

Chapter 4 studied the Shapley effects and compares them to the independent and full Sobol' indices in the case of dependent input variables. The Shapley effects were shown to be interesting importance measure when dealing with dependencies [Owen and Prieur, 2017; Iooss and Prieur, 2017]. Our contribution is to implement a bootstrap sampling in the existing estimation algorithm [Song et al., 2016] in order to estimate confidence intervals from the Monte Carlo error for the exact and random permutation methods. For high dimensional problems, the estimation algorithm requires a large number of model evaluations. Thus, to reduce the number of model evaluations, one can substitute the true model with a metamodel. When using a kriging model, we proposed an extension of the algorithm such as it computes the overall estimation error from the kriging model error and the Monte Carlo error.

Chapter 5 aims at determining a worst-case dependence structure when the dependencies between input random variables are unknown in a reliability study. This chapter proposed a greedy heuristic algorithm that explores a set of possible dependencies by taking advantage of the pair-copula construction (PCC) of multivariate probability distributions. Results of experiments conducted on toy examples and a real model illustrate the good behavior of the procedure: in situations where the monotonicity of the considered risk indicator (the output quantile) with respect to the inputs is postulated, a minimum value for the risk indicator is obtained using Fréchet-Hoeffding bounds. Nonetheless, it is possible to exhibit situations where the algorithm detect other and more conservative dependence structures.

6.2 Perspectives

In Chapter 3, future work can be done in the analysis of identified bias. For example, the bias can be estimated through some bootstraping (for example in Wager et al. [2014]). We are convinced that the difficulties observed by Ishwaran and Lu [2018] in the evaluation of the confidence intervals for the PVI values can be linked to the observed bias. The use of a Rosenblatt transformation is a natural idea to transform strongly dependent samples into independent ones. However, this transformation requires knowing the whole input distribution which is not always available in practice. In that case, one can infer the input distribution through different techniques, or consider another transformation such as the procedure from Iman and Conover [1982] (also proposed in Mara et al. [2015]). However it can only be applied when the dependence structure is defined by a rank correlation matrix.

In Chapter 4, the estimation algorithm is efficient, but is extremely costly in high dimensions. A valuable improvement of the algorithm would be the use of a Kernel estimation procedure in order to significantly reduce the number of evaluation. The Polynomial Chaos Expansion are good to compute the Sobol' indices analytically from the polynomial coefficients [Crestaux et al., 2009]. It would probably be interesting to have such decomposition for the Shapley effects. Moreover, determining a theoretical connection between the Shapley effects and the permutation

based variable importance in random forest (Chapter 3) would be an interesting perspective.

In Chapter 5, it would be interesting to improve the statistical estimation of the minimum quantile, given a dependence structure, by checking the hypotheses underlying the convergence results of Theorem 3. The grid search estimation strategy promoted in Section 5.2.3 arises from the lack of information about the convexity and the gradient of the output quantile. However, the method remains basic and stochastic recursive algorithms, such as the Robbins-Munro algorithm Robbins and Monro [1951], can possibly be more powerful (faster) alternatives. A significant issue is the computational cost of the exploration of possible dependence structures. Reducing this cost while increasing the completeness of this exploration should be the main concern for future works. Guiding the exploration in the space of conditional bivariate copulas using enriching criteria and possible expert knowledge can facilitate the minimization. The Algorithm 7 can also be improved using nonparametric bootstraps. This would quantify the estimation quality of the selected minimum quantile of each iteration. Note, however, that a seducing feature of an iterative procedure is the a priori possibility of its adaptation to situations where the computational model η is time-consuming. In such cases, it is likely that Bayesian global optimization methods based on replacing the computer model by a surrogate model (e.g., a kriging-based meta-model) Osborne et al. [2009] should be explored. We noticed in our experiments on real case studies that expert knowledge remains difficult to incorporate otherwise that using association and concordance measures, mainly since we are lacking representation tools (e.g., visual) of the properties of multivariate laws that provide intelligible diagnostics. A first step towards the efficient incorporation of expert knowledge could be to automatize the visualization of the obtained vine structures, to simplify judgements about their realism. Finally, another approach to consider could be to address the optimization problem (5.2) within the more general framework of optimal transport theory, and to take advantage of the many ongoing works in this area. Indeed, the problem (5.2) can be seen as a multi-marginal optimal transport problem (see Pass, Brendan [2015] for an overview). When $d = 2$, it corresponds respectively to the classical optimal transport problems of Monge and Kantorovich Villani [2008]. However, the multimarginal theory is not as well understood as for the bimarginal case, and developing efficient algorithms for solving this problem also remains a challenging issue Pass, Brendan [2015].

Bibliography

Aas, K., Czado, C., and Brechmann, E. C. (2012). Truncated regular vines in high dimensions with application to financial data. *Canadian Journal of Statistics*, 40(1):68–85.

Aas, K., Czado, C., Frigessi, A., and Bakken, H. (2009). Pair-copula constructions of multiple dependence. *Insurance, Mathematics and Economics*, 44:182–198.

Agrawal, S., Ding, Y., Saberi, A., and Ye, Y. (2012). Price of correlations in stochastic optimization. *Operations Research*, 60(1):150–162.

Archer, G., Saltelli, A., and Sobol, I. (1997). Sensitivity measures, anova-like techniques and the use of bootstrap. *Journal of Statistical Computation and Simulation*, 58(2):99–120.

Baudin, M. and Martinez, J.-M. (2014). Introduction to sensitivity analysis with nisp. https://forge.scilab.org/index.php/p/nisp/downloads/701/.

Bayarri, M., Berger, J., Paulo, R., Sacks, J., Cafeo, J., Cavendish, J., Lin, C., and Tu, J. (2007). A framework for validation of computer models. *Technometrics*, 49:138–154.

Beaudoin, D. and Lakhal-Chaieb, L. (2008). Archimedean copula model selection under dependent truncation. *Statistics in medicine*, 27(22):4440–4454.

Bedford, T. and Cooke, R. M. (2001). Probability density decomposition for conditionally dependent random variables modeled by vines. *Annals of Mathematics and Artificial intelligence*, 32(1):245–268.

Bedford, T. and Cooke, R. M. (2002). Vines: A new graphical model for dependent random variables. *Annals of Statistics*, 30(4):1031–1068.

Bedford, T., Quigley, J., and Walls, L. (2006). Expert elicitation for reliable system design. *Statistical Science*, 21(4):428–450.

Benoumechiara, N. (2018). dep-impact: Uncertainty quantification under incomplete probability information with Python. https://github.com/nazben/dep-impact.

Benoumechiara, N. and Elie-Dit-Cosaque, K. (2019). Shapley effects for sensitivity analysis with dependent inputs: bootstrap and kriging-based algorithms. *ESAIM: Proceedings and Surveys*, 65:266–293.

Benoumechiara, N., Michel, B., Saint-Pierre, P., and Bousquet, N. (2018). Detecting and modeling worst-case dependence structures between random inputs of computational reliability models. *arXiv preprint arXiv:1804.10527*.

Biau, G., Devroye, L., and Lugosi, G. (2008). Consistency of random forests and other averaging classifiers. *Journal of Machine Learning Research*, 9(Sep):2015–2033.

Bobkov, S. and Ledoux, M. (2014). One-dimensional empirical measures, order statistics and kantorovich transport distances. https://perso.math.univ-toulouse.fr/ledoux/files/2016/12/MEMO.pdf.

Borgonovo, E. (2007). A new uncertainty importance measure. *Reliability Engineering & System Safety*, 92(6):771–784.

Borgonovo, E., Castaings, W., and Tarantola, S. (2011). Moment Independent Importance Measures: New Results and Analytical Test Cases. *Risk Analysis*, 31(3):p.404–428.

Breiman, L. (1996). Bagging predictors. *Machine learning*, 24(2):123–140.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Browne, T., Fort, J.-C., looss, B., and Le Gratiet, L. (2017). Estimate of quantile-oriented sensitivity indices. Preprint, https://hal.inria.fr/hal-01450891.

Caniou, Y. (2012). *Global sensitivity analysis for nested and multiscale modelling*. PhD thesis, Université Blaise Pascal-Clermont-Ferrand II.

Castro, J., Gómez, D., and Tejada, J. (2009). Polynomial calculation of the shapley value based on sampling. *Computers & Operations Research*, 36(5):1726–1730.

Chastaing, G., Gamboa, F., Prieur, C., et al. (2012). Generalized hoeffding-sobol decomposition for dependent variables-application to sensitivity analysis. *Electronic Journal of Statistics*, 6:2420–2448.

Chazal, F., Massart, P., and Michel, B. (2015). Rates of convergence for robust geometric inference. *to appear in Electronic Journal of Statistics, arXiv preprint arXiv:1505.07602*.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794. ACM.

Cherubini, U., Luciano, E., and Vecchiato, W. (2004). *Copula methods in finance*. John Wiley & Sons.

Commission, N. R. et al. (1975). Reactor safety study. an assessment of accident risks in us commercial nuclear power plants. appendices iii and iv. Technical report, Nuclear Regulatory Commission.

Council, N. R. (2012). *Assessing the Reliability of Complex Models: Mathematical and statistical foundations of Verification, Validation and Uncertainty Quantification*. Washington, DC: The National Academies Press.

Crestaux, T., Le Maitre, O., and Martinez, J.-M. (2009). Polynomial chaos expansion for sensitivity analysis. *Reliability Engineering & System Safety*, 94(7):1161–1172.

Csorgo, M. (1983). *Quantile processes with statistical applications*, volume 42. SIAM.

Cukier, R., Levine, H., and Shuler, K. (1978). Nonlinear sensitivity analysis of multiparameter model systems. *Journal of computational physics*, 26(1):1–42.

Czado, C. (2010). Pair-copula constructions of multivariate copulas. In *Copula theory and its applications*, pages 93–109. Springer.

Da Veiga, S. (2015). Global sensitivity analysis with dependence measures. *Journal of Statistical Computation and Simulation*, 85(7):1283–1305.

de Rocquigny, E., Devictor, N., and Tarantola, S. (2008). *Uncertainty in industrial practice: a guide to quantitative uncertainty management*. John Wiley & Sons.

Demarta, S. and McNeil, A. J. (2005). The t copula and related copulas. *International statistical review*, 73(1):111–129.

Denil, M., Matheson, D., and De Freitas, N. (2014). Narrowing the gap: Random forests in theory and in practice. In *International conference on machine learning*, pages 665–673.

Derennes, P., Morio, J., and Simatos, F. (2018). Estimation of moment independent importance measures using a copula and maximum entropy framework. In *2018 Winter Simulation Conference (WSC)*, pages 1623–1634. IEEE.

Derennes, P., Morio, J., and Simatos, F. (2019). A nonparametric importance sampling estimator for moment independent importance measures. *Reliability Engineering & System Safety*, 187:3–16.

Dissmann, J., Brechmann, E. C., Czado, C., and Kurowicka, D. (2013). Selecting and estimating regular vine copulae and application to financial returns. *Computational Statistics & Data Analysis*, 59:52–69.

Dvoretzky, A., Kiefer, J., and Wolfowitz, J. (1956). Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, pages 642–669.

D'Auria, F., Camargo, C., and Mazzantini, O. (2012). The best estimate plus uncertainty (bepu) approach in licensing of current nuclear reactors. *Nuclear Engineering and Design*, 248:317–328.

Efron, B. (1981). Nonparametric standard errors and confidence intervals. *canadian Journal of Statistics*, 9(2):139–158.

Embrechts, P., McNeil, A., and Straumann, D. (2002). Correlation and dependence in risk management: properties and pitfalls. *Risk management: value at risk and beyond*, pages 176–223.

Evans, L. C. and Gariepy, R. F. (2015). *Measure theory and fine properties of functions*. CRC press.

Fang, K.-T., Li, R., and Sudjianto, A. (2005). *Design and modeling for computer experiments*. CRC Press.

Fisher, R. (1925). *Statistical methods for research workers*. Oliver & Boyd, Edinburgh.

Fort, J.-C., Klein, T., and Rachdi, N. (2016). New sensitivity analysis subordinated to a contrast. *Communications in Statistics-Theory and Methods*, 45(15):4349–4364.

Frechet, M. (1951). Sur les tableaux de correlation dont les marges sont dollnes. *AIm. Unlv.*

Frees, E. W. and Valdez, E. A. (1998). Understanding relationships using copulas. *North American actuarial journal*, 2(1):1–25.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

Genuer, R., Poggi, J.-M., and Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognition Letters*, 31(14):2225–2236.

Goda, K. (2010). Statistical modeling of joint probability distribution using copula: Application to peak and permanent displacement seismic demands. *Structural Safety*, 32(2):112–123.

Gregorutti, B., Michel, B., and Saint-Pierre, P. (2015). Grouped variable importance with random forests and application to multiple functional data analysis. *Computational Statistics & Data Analysis*, 90:15–35.

Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. (2005). Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pages 63–77. Springer.

Grigoriu, M. and Turkstra, C. (1979). Safety of structural systems with correlated resistances. *Applied Mathematical Modelling*, 3(2):130–136.

Grömping, U. (2009). Variable importance assessment in regression: linear regression versus random forest. *The American Statistician*, 63(4):308–319.

Gruber, L. and Czado, C. (2015). Sequential bayesian model selection of regular vine copulas. *Bayesian Analysis*, 10(4):937–963.

Haff, I. H. (May 9-11, 2016, Oslo, Norway). How to select a good vine. *International FocuStat Workshop on Focused Information Criteria and Related Themes*.

Helton, J. (2011). Quantification of margins and uncertainties: conceptual and computational basis. *Reliability Engineering and System Safety*, 96:976–1013.

Hoeffding, W. (1940). Scale-invariant correlation theory. *Schriften des Mathematischen Instituts und des Instituts fur Angewandte Mathematik der Universit at Berlin*, 5(3):181–233.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics*, 19(3):293–325.

Homma, T. and Saltelli, A. (1996). Importance measures in global sensitivity analysis of nonlinear models. *Reliability Engineering & System Safety*, 52(1):1–17.

Hora, S. C. and Iman, R. L. (1986). Comparison of maximus/bounding and bayes/monte carlo for fault tree uncertainty analysis. Technical report, Hawaii Univ., Hilo (USA); Sandia National Labs., Albuquerque, NM (USA).

Iman, R. L. and Conover, W.-J. (1982). A distribution-free approach to inducing rank correlation among input variables. *Communications in Statistics-Simulation and Computation*, 11(3):311–334.

Iman, R. L. and Hora, S. C. (1990). A robust measure of uncertainty importance for use in fault tree system analysis. *Risk analysis*, 10(3):401–406.

Iooss, B. and Lemaître, P. (2015). A review on global sensitivity analysis methods. In *Uncertainty Management in Simulation-Optimization of Complex Systems*, pages 101–122. Springer.

Iooss, B. and Prieur, C. (2017). Shapley effects for sensitivity analysis with dependent inputs: comparisons with sobol'indices, numerical estimation and applications. *International Journal of uncertainty Quantification, In press*.

Ishigami, T. and Homma, T. (1990). An importance quantification technique in uncertainty analysis for computer models. In *Uncertainty Modeling and Analysis, 1990. Proceedings., First International Symposium on*, pages 398–403. IEEE.

Ishwaran, H. et al. (2007). Variable importance in binary regression trees and forests. *Electronic Journal of Statistics*, 1:519–537.

Ishwaran, H. and Kogalur, U. B. (2010). Consistency of random survival forests. *Statistics & probability letters*, 80(13-14):1056–1064.

Ishwaran, H. and Lu, M. (2018). Standard errors and confidence intervals for variable importance in random forest regression, classification, and survival. *Statistics in medicine*.

Janon, A., Klein, T., Lagnoux, A., Nodet, M., and Prieur, C. (2014). Asymptotic normality and efficiency of two sobol index estimators. *ESAIM: Probability and Statistics*, 18:342–364.

Jansen, M. J. (1999). Analysis of variance designs for model output. *Computer Physics Communications*, 117(1–2):35–43.

Jansen, M. J., Rossing, W. A., and Daamen, R. A. (1994). Monte carlo estimation of uncertainty contributions from several independent multivariate sources. In *Predictability and Nonlinear Modelling in Natural Sciences and Economics*, pages 334–343. Springer.

Jiang, C., Zhang, W., Han, X., Ni, B., and Song, L. (2015). A vine-copula-based reliability analysis method for structures with multidimensional correlation. *Journal of Mechanical Design*, 137(6):061405.

Joe, H. (1994). Multivariate extreme-value distributions with applications to environmental data. *Canadian Journal of Statistics*, 22(1):47–64.

Joe, H. (1996). Families of m-variate distributions with given margins and m (m-1)/2 bivariate dependence parameters. *Lecture Notes-Monograph Series*, pages 120–141.

Johns, M. V. (1988). Importance sampling for bootstrap confidence intervals. *Journal of the American Statistical Association*, 83(403):709–714.

Johnson, J. W. and LeBreton, J. M. (2004). History and use of relative importance indices in organizational research. *Organizational research methods*, 7(3):238–257.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

Krige, D. G. (1951). A statistical approach to some basic mine valuation problems on the witwatersrand. *Journal of the Southern African Institute of Mining and Metallurgy*, 52(6):119–139.

Kucherenko, S., Tarantola, S., and Annoni, P. (2012). Estimation of global sensitivity indices for models with dependent variables. *Computer Physics Communications*, 183(4):937–946.

Kurowicka, D. (2011). Optimal truncation of vines. In Kurowicka, D. and Joe, H., editors, *Dependence Modeling: Vine Copula Handbook*. World Scientific Publishing Co.

Kurowicka, D. and Cooke, R. M. (2006). *Uncertainty analysis with high dimensional dependence modelling*. John Wiley & Sons.

Le Gratiet, L., Cannamela, C., and Iooss, B. (2014). A bayesian approach for global sensitivity analysis of (multifidelity) computer codes. *SIAM/ASA Journal on Uncertainty Quantification*, 2(1):336–363.

Lemaire, M., Chateauneuf, A., and Mitteau, J.-C. (2005). *Fiabilité des structures: Couplage mécano-fiabiliste statique*. Hermès Science Publications.

Lemaire, M., Chateauneuf, A., and Mitteau, J.-C. (2010). *Structural reliability*. Wiley.

Li, G. and Rabitz, H. (2010). Global sensitivity analysis for systems with independent and/or correlated inputs. *Procedia - Social and Behavioral Sciences*, 2(6):7587–7589.

Li, G., Wang, S.-W., Rosenthal, C., and Rabitz, H. (2001). High dimensional model representations generated from low dimensional data samples. i. mp-cut-hdmr. *Journal of Mathematical Chemistry*, 30(1):1–30.

Louppe, G., Wehenkel, L., Sutera, A., and Geurts, P. (2013). Understanding variable importances in forests of randomized trees. In *Advances in neural information processing systems*, pages 431–439.

Lundberg, S. M., Erion, G. G., and Lee, S.-I. (2018). Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888*.

Lunetta, K. L., Hayward, L. B., Segal, J., and Van Eerdewegh, P. (2004). Screening large-scale association study data: exploiting interactions using random forests. *BMC genetics*, 5(1):32.

Malevergne, Y., Sornette, D., et al. (2003). Testing the gaussian copula hypothesis for financial assets dependences. *Quantitative Finance*, 3(4):231–250.

Mara, T. A. and Tarantola, S. (2012). Variance-based sensitivity indices for models with dependent inputs. *Reliability Engineering & System Safety*, 107:115–121.

Mara, T. A., Tarantola, S., and Annoni, P. (2015). Non-parametric methods for global sensitivity analysis of model output with dependent inputs. *Environmental Modelling & Software*, 72:173–183.

Marrel, A., Iooss, B., Laurent, B., and Roustant, O. (2009). Calculations of sobol indices for the gaussian process metamodel. *Reliability Engineering & System Safety*, 94(3):742–751.

Martin, J. D. and Simpson, T. W. (2004). On the use of kriging models to approximate deterministic computer models. In *ASME 2004 international design engineering technical conferences and computers and information in engineering conference*, pages 481–492. American Society of Mechanical Engineers.

Massart, P. (1990). The tight constant in the dvoretzky-kiefer-wolfowitz inequality. *The Annals of Probability*, pages 1269–1283.

Matheron, G. (1962). *Traité de géostatistique appliquée. 1 (1962)*, volume 1. Editions Technip.

Maume-Deschamps, V. and Niang, I. (2018). Estimation of quantile oriented sensitivity indices. *Statistics & Probability Letters*, 134:122–127.

McKay, M. D. (1997). Nonparametric variance-based methods of assessing uncertainty importance. *Reliability engineering & system safety*, 57(3):267–279.

McKay, M. D., Beckman, R. J., and Conover, W. J. (1979). Comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 21(2):239–245.

McNeil, A. J. and Nešlehová, J. (2009). Multivariate archimedean copulas, d-monotone functions and l-norm symmetric distributions. *The Annals of Statistics*, pages 3059–3097.

Mentch, L. and Hooker, G. (2016). Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *The Journal of Machine Learning Research*, 17(1):841–881.

Morales Nápoles, O. (2010). *Bayesian belief nets and vines in aviation safety and other applications*. PhD thesis, TU Delft, Delft University of Technology.

Morales-Nápoles, O. (2011). *Counting vines*. World Scientific.

Morris, M. D. (1991). Factorial sampling plans for preliminary computational experiments. *Technometrics*, 33(2):161–174.

Nelsen, R. B. (2007). *An introduction to copulas*. Springer Science & Business Media.

Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of econometrics*, 4:2111–2245.

Nilsen, T. and Aven, T. (2003). Models and model uncertainty in the context of risk analysis. *Reliability Engineering & System Safety*, 79(3):309–317.

Oakley, J. E. and O'Hagan, A. (2004). Probabilistic sensitivity analysis of complex models: a bayesian approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(3):751–769.

Osborne, M., Garnett, R., and Roberts, S. (2009). Gaussian processes for global optimization. *Proceedings of the 3rd International Conference on Learning and Intelligent Optimization*.

Owen, A. B. (2014). Sobol'indices and shapley value. *SIAM/ASA Journal on Uncertainty Quantification*, 2(1):245–251.

Owen, A. B. and Prieur, C. (2017). On shapley value for measuring importance of dependent inputs. *SIAM/ASA Journal on Uncertainty Quantification*, 5(1):986–1002.

Parzen, E. (1962). On estimation of a probability density function and mode. *Ann. Math. Statist.*, 33(3):1065–1076.

Pass, Brendan (2015). Multi-marginal optimal transport: Theory and applications. *ESAIM: M2AN*, 49(6):1771–1790.

Rahman, S. (2018). A polynomial chaos expansion in dependent random variables. *Journal of Mathematical Analysis and Applications*, 464(1):749–775.

Rall, L. B. (1981). Automatic differentiation: Techniques and applications.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407.

Rosenblatt, M. (1952). Remarks on a multivariate transformation. *The annals of mathematical statistics*, 23(3):470–472.

Sacks, J., Welch, W. J., Mitchell, T. J., and Wynn, H. P. (1989). Design and analysis of computer experiments. *Statistical science*, pages 409–423.

Saltelli, A. (2002). Making best use of model evaluations to compute sensitivity indices. *Computer Physics Communications*, 145(2):280–297.

Saltelli, A., Annoni, P., Azzini, I., Campolongo, F., Ratto, M., and Tarantola, S. (2010). Variance based sensitivity analysis of model output. design and estimator for the total sensitivity index. *Computer Physics Communications*, 181(2):259–270.

Saltelli, A., Chan, K., Scott, E. M., et al. (2000). *Sensitivity analysis*, volume 1. Wiley New York.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., and Tarantola, S. (2008). *Global sensitivity analysis: the primer*. John Wiley & Sons.

Saltelli, A. and Tarantola, S. (2002). On the relative importance of input factors in mathematical models: safety assessment for nuclear waste disposal. *Journal of the American Statistical Association*, 97(459):702–709.

Saltelli, A., Tarantola, S., Campolongo, F., and Ratto, M. (2004). *Sensitivity analysis in practice: a guide to assessing scientific models*. John Wiley & Sons.

Saltelli, A., Tarantola, S., and Chan, K.-S. (1999). A quantitative model-independent method for global sensitivity analysis of model output. *Technometrics*, 41(1):39–56.

Salvador, R., Pinol, J., Tarantola, S., and Pla, E. (2001). Global sensitivity analysis and scale effects of a fire propagation model used over mediterranean shrublands. *Ecological Modelling*, 136(2-3):175–189.

Santner, T. J., Williams, B. J., Notz, W., and Williams, B. J. (2003). *The design and analysis of computer experiments*, volume 1. Springer.

Scarf, H., Arrow, K., and Karlin, S. (1958). A min-max solution of an inventory problem. *Studies in the mathematical theory of inventory and production*, 10(2):201.

Schoelzel, C. and Friederichs, P. (2008). Multivariate non-normally distributed random variables in climate research–introduction to the copula approach. *Nonlin. Processes Geophys.*, 15(5):761–772.

Scornet, E., Biau, G., Vert, J.-P., et al. (2015). Consistency of random forests. *The Annals of Statistics*, 43(4):1716–1741.

Shapley, L. S. (1953). A value for n-person games. *Contributions to the Theory of Games*, 2(28):307–317.

Shapley, L. S. and Shubik, M. (1954). A method for evaluating the distribution of power in a committee system. *American political science review*, 48(3):787–792.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*, volume 26. CRC press.

Sklar, A. (1959). *Fonctions de répartition à n dimensions et leurs marges*, volume 8. ISUP.

Sklar, A. (1996). Random variables, distribution functions, and copulas: a personal look backward and forward. *Lecture notes-monograph series*, pages 1–14.

Sobol', I. (2007). Global sensitivity analysis indices for the investigation of nonlinear mathematical models. *Matematicheskoe Modelirovanie*, 19(11):23–24.

Sobol, I. M. (1993). Sensitivity estimates for nonlinear mathematical models. *Mathematical Modelling and Computational Experiments*, 1(4):407–414.

Sobol, I. M. (2001). Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates. *Mathematics and computers in simulation*, 55(1-3):271–280.

Song, E., Nelson, B. L., and Staum, J. (2016). Shapley effects for global sensitivity analysis: Theory and computation. *SIAM/ASA Journal on Uncertainty Quantification*, 4(1):1060–1083.

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., and Zeileis, A. (2008). Conditional variable importance for random forests. *BMC bioinformatics*, 9(1):307.

Strobl, C., Boulesteix, A.-L., Zeileis, A., and Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC bioinformatics*, 8(1):25.

Strobl, C., Malley, J., and Tutz, G. (2009). An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological methods*, 14(4):323.

Sudret, B. (2007). Uncertainty propagation and sensitivity analysis in mechanical models–contributions to structural reliability and stochastic spectral methods. *Habilitationa diriger des recherches, Université Blaise Pascal, Clermont-Ferrand, France*.

Székely, G. J. and Rizzo, M. L. (2013). Energy statistics: A class of statistics based on distances. *Journal of statistical planning and inference*, 143(8):1249–1272.

Tang, X.-S., Li, D.-Q., Rong, G., Phoon, K.-K., and Zhou, C.-B. (2013). Impact of copula selection on geotechnical reliability under incomplete probability information. *Computers and Geotechnics*, 49:264–278.

Tang, X.-S., Li, D.-Q., Zhou, C.-B., and Phoon, K.-K. (2015). Copula-based approaches for evaluating slope reliability under incomplete probability information. *Structural Safety*, 52:90–99.

Tarantola, S., Gatelli, D., Kucherenko, S., Mauntz, W., et al. (2007). Estimating the approximation error when fixing unessential factors in global sensitivity analysis. *Reliability Engineering & System Safety*, 92(7):957–960.

Tarantola, S., Gatelli, D., and Mara, T. A. (2006). Random balance designs for the estimation of first order global sensitivity indices. *Reliability Engineering & System Safety*, 91(6):717–727.

Thoft-Christensen, P. and Sørensen, J. D. (1982). Reliability of structural systems with correlated elements. *Applied Mathematical Modelling*, 6(3):171–178.

Thomas, D. R., Hughes, E., and Zumbo, B. D. (1998). On variable importance in linear regression. *Social Indicators Research*, 45(1-3):253–275.

Tissot, J.-Y. and Prieur, C. (2012). Bias correction for the estimation of sensitivity indices based on random balance designs. *Reliability Engineering & System Safety*, 107:205–213.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge University Press.

Villani, C. (2008). *Optimal transport: old and new*, volume 338. Springer Science & Business Media.

Wager, S., Hastie, T., and Efron, B. (2014). Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. *The Journal of Machine Learning Research*, 15(1):1625–1651.

Wei, P., Lu, Z., and Song, J. (2015). A comprehensive comparison of two variable importance analysis techniques in high dimensions: Application to an environmental multi-indicators system. *Environmental Modelling & Software*, 70:178–190.

Wei, P., Lu, Z., and Yuan, X. (2013). Monte carlo simulation for moment-independent sensitivity analysis. *Reliability Engineering & System Safety*, 110:60–67.

Wilson, G. E. (2013). Historical insights in the development of best estimate plus uncertainty safety analysis. *Annals of Nuclear Energy*, 52:2–9.

Winham, S. J., Colby, C. L., Freimuth, R. R., Wang, X., de Andrade, M., Huebner, M., and Biernacka, J. M. (2012). Snp interaction detection with random forests in high-dimensional genetic data. *BMC bioinformatics*, 13(1):164.

Zhang, L., Lu, Z., Cheng, L., and Fan, C. (2014). A new method for evaluating borgonovo moment-independent importance measure with its application in an aircraft structure. *Reliability Engineering & System Safety*, 132:163–175.

Zhu, R., Zeng, D., and Kosorok, M. R. (2015). Reinforcement learning trees. *Journal of the American Statistical Association*, 110(512):1770–1784.

Zinkle, S. J. and Was, G. (2013). Materials challenges in nuclear energy. *Acta Materialia*, 61(3):735–758.

Résumé

Les études de fiabilité des structures ont recours à des approches probabilistes permettant de quantifier le risque qu'un événement accidentel se produise. La dépendance entre les variables aléatoires d'entrée d'un modèle peut avoir un impact significatif sur les résultats de l'étude de sureté. Cette thèse apporte une contribution au traitement de la dépendance en fiabilité des structures. Les deux principaux thèmes traités dans ce document sont, d'une part, l'analyse de sensibilité pour variables dépendantes lorsque la dépendance est connue et, d'autre part, l'évaluation d'un risque de fiabilité lorsque la dépendance est inconnue. Dans un premier temps, nous proposons une extension des mesures d'importance par permutation de l'algorithme des forêts aléatoires au cas de données dépendantes. Nous adaptons aussi l'algorithme d'estimation des indices de Shapley, utilisés en théorie des jeux, afin de prendre compte l'erreur d'estimation des indices. Dans un second temps, lorsque la structure de dépendance est inconnue, nous proposons une estimation conservative du risque de fiabilité basée sur une modélisation de la dépendance qui permet de déterminer la structure de dépendance la plus pénalisante. La méthodologie proposée est appliquée à un exemple de fiabilité structurelle permettant d'obtenir une estimation conservative du risque.

Mots-clés : fiabilité des structures, copules, analyse de sensibilité, forêts aléatoires, estimation conservative

Abstract

Structural reliability studies use probabilistic approaches to quantify the risk of an accidental event occurring. The dependence between the random input variables of a model can have a significant impact on the results of the reliability study. This thesis contributes to the treatment of dependency in structural reliability studies. The two main topics covered in this document are the sensitivity analysis for dependent variables when the dependence is known and, as well as the assessment of a reliability risk when the dependence is unknown. First, we propose an extension of the permutation-based importance measures of the random forest algorithm towards the case of dependent data. We also adapt the Shapley index estimation algorithm, used in game theory, to take into account the index estimation error. Secondly, in the case of dependence structure being unknown, we propose a conservative estimate of the reliability risk based on dependency modelling to determine the most penalizing dependence structure. The proposed methodology is applied to an example of structural reliability to obtain a conservative estimate of the risk.

Key-words: structural reliability, copulas, sensitivity analysis, random forests, conservative estimate