Développement de mesures d'incertitudes pour le risque de modèle dans des contextes incluant de la dépendance stochastique

Kévin Elie-Dit-Cosaque

N° d'ordre NNT : 2020LYSE1204

THÈSE DE DOCTORAT DE L'UNIVERSITÉ DE LYON
opérée au sein de
l'Université Claude Bernard Lyon 1

École Doctorale ED 512
InfoMaths

Spécialité de doctorat : Mathématiques
Discipline : Mathématiques Appliquées

Soutenue publiquement le 13 Novembre 2020, par :

Kévin Elie-Dit-Cosaque

Développement de mesures d'incertitudes pour le risque de modèle dans des contextes incluant de la dépendance stochastique

---

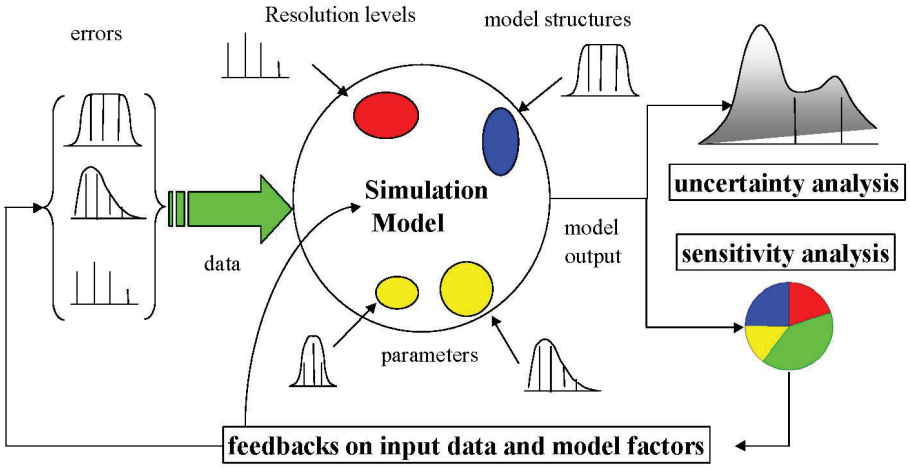
Devant le jury composé de :

Michel Bertrand, Professeur	École Centrale de Nantes	Rapporteur
Prieur Clémentine, Professeure	Université Grenoble Alpes	Rapporteure
Cuberos Andrés, Actuaire	SCOR	Examinateur
Fougères Anne-Laure, Professeure	Université Lyon 1	Examinatrice
Iooss Bertrand, Ingénieur de Recherche	EDF R&D	Examinateur
Maume-Deschamps Véronique, Professeure	Université Lyon 1	Directrice de thèse
Nisipasu Ecaterina, Actuaire	SCOR	Encadrante
Poulin Mathieu, Actuaire	SCOR	Encadrant

Institut Camille Jordan

Laboratoire de recherche en mathématiques Lyon/Saint-Étienne

Développement de mesures d'incertitudes pour le risque de modèle dans des contextes incluant de la dépendance stochastique



Kévin Elie-Dit-Cosaque

Thèse de doctorat

Université Claude Bernard Lyon 1

Université Claude Bernard – LYON 1

Administrateur provisoire de l'Université	M. Frédéric FLEURY
Président du Conseil Académique	M. Hamda BEN HADID
Vice-Président du Conseil d'Administration	M. Didier REVEL
Vice-Président du Conseil des Etudes et de la Vie Universitaire	M. Philippe CHEVALLIER
Vice-Président de la Commission de Recherche	M. Jean-François MORNEX
Directeur Général des Services	M. Pierre ROLLAND

COMPOSANTES SANTE

Département de Formation et Centre de Recherche en Biologie Humaine	Directrice : Mme Anne-Marie SCHOTT
Faculté d'Odontologie	Doyenne : Mme Dominique SEUX
Faculté de Médecine et Maïeutique Lyon Sud - Charles Mérieux	Doyenne : Mme Carole BURILLON
Faculté de Médecine Lyon-Est	Doyen : M. Gilles RODE
Institut des Sciences et Techniques de la Réadaptation (ISTR)	Directeur : M. Xavier PERROT
Institut des Sciences Pharmaceutiques et Biologiques (ISBP)	Directrice : Mme Christine VINCIGUERRA

COMPOSANTES & DEPARTEMENTS DE SCIENCES & TECHNOLOGIE

Département Génie Electrique et des Procédés (GEP)	Directrice : Mme Rosaria FERRIGNO
Département Informatique	Directeur : M. Behzad SHARIAT
Département Mécanique	Directeur M. Marc BUFFAT
Ecole Supérieure de Chimie, Physique, Electronique (CPE Lyon)	Directeur : Gérard PIGNAULT
Institut de Science Financière et d'Assurances (ISFA)	Directeur : M. Nicolas LEBOISNE
Institut National du Professorat et de l'Education	Administrateur Provisoire : M. Pierre CHAREYRON
Institut Universitaire de Technologie de Lyon 1	Directeur : M. Christophe VITON
Observatoire de Lyon	Directrice : Mme Isabelle DANIEL
Polytechnique Lyon	Directeur : Emmanuel PERRIN
UFR Biosciences	Administratrice provisoire : Mme Kathrin GIESELER
UFR des Sciences et Techniques des Activités Physiques et Sportives (STAPS)	Directeur : M. Yannick VANPOULLE
UFR Faculté des Sciences	Directeur : M. Bruno ANDRIOLETTI

Institut Camille Jordan (ICJ)
Université Claude Bernard Lyon 1
43 boulevard du 11 novembre 1918
F-69622 Villeurbanne Cedex

Group Actuarial Department - Actuarial Modelling
SCOR SE
5, avenue Kléber
75795 Paris Cedex 16
France

Remerciements

Je présente par avance mes excuses à celles et ceux que j'aurais oubliés dans ces remerciements.

Tout d'abord, je remercie chaleureusement Véronique Maume-Deschamps d'avoir accepté d'être ma directrice de thèse. Je tiens à lui témoigner toute ma reconnaissance pour son engagement, sa rigueur et son accompagnement sans faille tout au long de ces années de collaboration. Je tiens également à lui présenter toute ma gratitude pour sa grande disponibilité et pour avoir toujours su m'apporter l'aide escomptée ainsi que les réponses à mes questions pas toujours clairement formulées.

Cette thèse ayant été établie sous convention CIFRE entre l'Institut Camille Jordan (ICJ) et SCOR, je remercie SCOR et tout particulièrement Eric Lecoeur et Ecaterina Nisipasu de m'avoir donné l'opportunité de faire ma thèse au sein de la direction actuariat groupe. Merci, Kati, de m'avoir fait confiance et accompagné tout au long de ce projet. Nos conversations ont été très constructives et tes remarques et conseils ont contribué à la réalisation de ce travail. Enfin, je porte une attention particulière à la grande bienveillance de Mathieu Poulin sans qui cette aventure n'aurait pu aller à son terme et qui m'a permis d'achever ma thèse dans les meilleures conditions.

Je remercie particulièrement, Bertrand Michel et Clémentine Prieur, de m'avoir fait l'honneur d'être rapporteurs de ma thèse. Je vous adresse toute ma reconnaissance pour le temps consacré à la lecture de mon manuscrit ainsi que pour vos commentaires constructifs ayant contribué à améliorer sa qualité. Je remercie également Andrés Cuberos, Anne-Laure Fougères et Bertrand Iooss d'avoir accepté d'être membres de mon jury mais aussi de m'avoir aidé et accompagné durant ces trois années de thèse.

Je tiens à saluer amicalement les différentes personnes, qui m'ont aidé d'une façon ou d'une autre, de près ou de loin, sans forcément le savoir, au bon déroulement de ces trois années.

J'adresse mes remerciements à tous les membres de l'ICJ, les chercheurs, les enseignants et le personnel administratif. En particulier, je présente ma profonde reconnaissance à Roland Denis et Benoit Fabrèges pour leur support informatique, leur disponibilité et leur engagement au cours de ces trois années. J'ai apprécié travailler avec vous, spécialement quand il y avait des bugs numba sous Linux et non sous Win-

dows ! Mes pensées vont également à Laurent Azema et Vincent Farget pour leur aide technique, à Thibault Espinasse et Cécile Mercadier chez qui j'ai toujours trouvé porte ouverte pour répondre à mes questions. Merci également à Lydia Barlerin et Christine Lesueur pour leur aide administrative et leur bienveillance envers les doctorants.
Un grand merci à Esterina Masiello et Didier Rullière pour leur soutien ainsi qu'à Pierre Ribereau pour nos échanges sur Cuba !
Mes remerciements vont également à tous les doctorants de l'ICJ avec qui j'ai pu échanger et partager cette expérience et notamment Oskar, Mélanie, Marina, Gauthier, Thomas, Simon, Octave, Sam, Garry, Colin, Lola, Shmuel et les anciens Antoine, Ibrahima, Manaf, Abdul-Fatah et Khalil. Une mention spéciale à mes collègues des bureaux 118 et 226 : Alexis, Anatole, Benjamin, Dimitri et Uran.
Je leur souhaite à tous le meilleur pour la suite.

Je souhaite remercier sincèrement mes collègues de SCOR : Julie De., Julie Do., Imen, Héloïse, Philippe, Keunoo, Emmanuelle, Tze-Yang, Jérôme, François, Didier, Amy, Adrien, Virginie, Markus et Romain et, particulièrement, Dorothée et Przemek qui supportent mes innombrables questions et font preuve d'une patience admirable. Benjamin, rendez-vous pour une nouvelle course au Bois ! J'en profite pour exprimer une pensée chaleureuse à Bastien, Charles, Enzo, Erwan, Naöelle, Pauline et Rhiannon pour nos discussions et agréables moments.

Un énorme « Big Up » à mes amis pour leur bonne humeur et leur compréhension malgré mes absences régulières aux invitations. Une grosse pensée à mes amis du CEMRACS, Nazih, Houssam, Mohamed, Alexandre, Laurence, Bruno, Nordine et tous les autres, avec qui j'ai partagé une expérience enrichissante tant professionnelle que personnelle. Les hostilités peuvent reprendre !

Je souhaite adresser toute ma gratitude à l'ensemble de ma famille et de ma belle-famille pour leur soutien et leurs pensées. En particulier, je remercie du fond du cœur ma mère et mon père pour leur soutien indéfectible dans chacun de mes choix. Vous avez tout fait pour que j'arrive jusqu'ici et je vous en suis sincèrement reconnaissant.

Mes derniers mots vont à la personne qui a vécu cette thèse de l'intérieur et qui a accepté tous les sacrifices que cela représente, y compris de mettre de côté notre vie sociale. Merci infiniment pour ton amour et ton soutien inconditionnel tant pour ce projet, malgré ton allergie pour les mathématiques, que pour tout le reste !

Mèsi anpil !

If you can't fly then run,
if you can't run then walk,
if you can't walk then crawl,
but whatever you do,
you have to keep moving forward.

Martin Luther King Jr.

Table des matières

Chapitre 1

Introduction

Le travail présenté dans ce manuscrit a été réalisé dans le cadre d'une thèse CIFRE régie par une convention de recherche entre la société de réassurance SCOR et l'Institut Camille Jordan. Cette thèse a pour objectif de développer des mesures d'incertitudes dans des contextes pouvant inclure de la dépendance stochastique. Une motivation est l'étude du risque de modèle.

Dans cette introduction, nous énonçons le contexte et les considérations pratiques en lien avec le risque de modèle et l'analyse de sensibilité, avant de présenter les différentes contributions de cette thèse.

Sommaire

1.1 (Ré) assurance

La principale caractéristique qui différencie le fonctionnement de l'assurance des autres secteurs de l'économie est l'inversion de son cycle de production. En effet, l'acquisition des primes survient en amont de la survenance du risque couvert. Ce cycle oblige les compagnies à évaluer le moment de la matérialisation du risque ainsi que son montant. Cela s'effectue par le biais de différents modèles mathématiques plus ou moins complexes en fonction de la nature du risque.

L'article du CRO Forum [2017][1] définit un modèle comme une méthode, un système ou une approche quantitative qui applique des théories, techniques et hypothèses statistiques, économiques, financières ou mathématiques à des données d'entrée dans le but d'obtenir des estimations quantitatives (définition similaire à celle donnée par la FED[2] Reserve [2011]). Un modèle est un outil d'aide à la décision qui est construit dans un but précis. En conséquence, les valeurs retournées sont traduites en une information utilisable en fonction de l'objectif visé. La (ré)assurance utilise depuis longtemps différentes typologies de modèle pour évaluer les risques pour, par exemple, la tarification des contrats, le provisionnement ou encore pour calculer l'exigence de capital et la solvabilité.

Réglementation de l'assurance

Le secteur de l'assurance est fortement réglementé tout comme celui des banques du fait de leur importance dans le paysage économique. Depuis plusieurs années, il y a une tendance générale de convergence des réglementations vers une prise en compte plus approfondie de tous les risques qui sont supportés par les sociétés d'assurance ou de réassurance.

Par exemple, au niveau européen, la Directive *Solvabilité II*, entrée en vigueur au 1er janvier 2016, a été développée pour succéder à la Directive Solvabilité I qui était en place depuis 1973 mais présentait de nombreuses limitations. En effet, cette directive ne tenait pas compte du profil de risque des compagnies et reposait sur une formule trop simple pour évaluer tous les risques auxquels elles sont exposées tels que les risques économiques ou le risque opérationnel. Solvabilité II avait donc pour principal objectif d'harmoniser et de moderniser à l'échelle européenne les règles de solvabilité imposées aux (ré)assureurs tout en intégrant l'ensemble des risques qui pèsent de manière effective sur les compagnies.

La Directive Solvabilité II s'est inspirée des accords de Bâle II, établis dans le secteur bancaire, pour proposer une structure en trois piliers. Au travers de ceux-ci sont définis les exigences quantitatives et qualitatives imposées aux compagnies d'assurance ou de réassurance afin d'assurer leur solvabilité et la stabilité financière.

[1] CRO : Chief Risk Officer
[2] FED : Federal Reserve System

Tout d'abord, le Pilier I comprend principalement une exigence quantitative en fonds propres que doit détenir la compagnie pour être en mesure d'éviter toute faillite à horizon un an avec une probabilité de 99.5%.

Le Pilier II décrit principalement les normes qualitatives en termes de gouvernance et de gestion des risques pour les compagnies d'assurance. L'objectif de ce pilier repose essentiellement sur la volonté d'établir une organisation efficace garantissant une gestion prudente, sûre et pérenne de l'activité. Les mesures à appliquer consistent notamment en la mise en place de quatre fonctions clés (Actuariat, Audit interne, Contrôle interne et Gestion des risques) ainsi qu'en une évaluation interne des risques et de la solvabilité notamment au travers du dispositif ORSA(Own Risk and Solvency Assessment).

Enfin, le Pilier III repose, quant à lui, sur la mise en place d'une communication externe et d'une transparence de l'information (exigences adaptées à la taille de l'entreprise). La compagnie se doit de rendre publique des informations relatives à sa situation comptable et financière.

Capital de solvabilité

Comme énoncé précédemment, Solvabilité II exige que les compagnies disposent d'un capital suffisant tel que la probabilité de ruine économique à un an soit inférieure à 0.5%. Ce capital est dénommé dans la Directive comme *capital de solvabilité requis*, ou encore SCR pour *Solvency Capital Requirement*. Les compagnies seront dès lors considérées comme solvables si leur montant de fonds propres est supérieur au SCR. La détermination du SCR peut se faire soit par l'utilisation de la Formule Standard fournie par le régulateur, soit par l'utilisation d'un modèle interne partiel ou complet développé par la compagnie et approuvé par les autorités de contrôle.

La Formule Standard se base sur une structure de calcul reposant sur des principes et des hypothèses fixés par la Directive. Il est important de noter que le calibrage de la Formule Standard a été effectué afin de refléter le profil de risque de la plupart des entreprises d'assurance et de réassurance européennes. Par conséquent, il peut arriver que cette approche standardisée ne traduise pas correctement le profil de risque très particulier d'une entreprise.
Les compagnies ont aussi la possibilité de concevoir un modèle interne qui couvrira tous les risques importants auxquels elles sont exposées. Contrairement à la Formule Standard, le modèle interne permet aux compagnies de choisir les méthodes de modélisation qu'elles souhaitent adopter pour leurs risques individuels. Toutefois, indépendamment des méthodes de calcul retenues, la Directive impose que ce modèle intègre à minima les risques suivants répertoriés au paragraphe 4 de l'Article 101 : risque de souscription en non-vie, risque de souscription en vie, risque de souscription en santé, risque de marché, risque de crédit et le risque opérationnel.

Ainsi, bien que le développement d'un modèle interne soit extrêmement contrai-

gnant et coûteux, cela constitue un véritable atout pour la compagnie car il reflète tous ses risques aussi bien à l'actif qu'au passif de façon plus précise et pertinente que la Formule Standard. Il joue un rôle important dans le système de gestion des risques, permet d'orienter les prises de décision et surtout d'optimiser l'allocation du capital nécessaire. En effet, la compagnie se doit d'avoir un modèle qui valorise au mieux ses fonds propres requis afin de ne pas risquer de les sous-estimer ou surestimer. Une sous-estimation des fonds propres requis pourrait rendre l'assureur ou le réassureur incapable de faire face à ses engagements et une surestimation pourrait engendrer le blocage de plus de capital que nécessaire, ce qui rendrait à contrario l'assureur ou le réassureur moins compétitif sur le marché.

A la lumière de l'exemple du capital de solvabilité requis et des enjeux économiques sous-jacents, il apparaît donc crucial que les méthodes de modélisation utilisées reflètent au mieux les risques couverts par les contrats d'assurance ou de réassurance. Cela a été rendu possible grâce à l'innovation informatique qui a mené à une augmentation significative de la capacité de calcul et permis l'élaboration de méthodes plus performantes. L'évolution de la théorie mathématique ainsi que la disponibilité de données de plus en plus détaillées ont également donné lieu à une modélisation plus fidèle de certains phénomènes réels. Cette précision vient au prix d'une plus grande complexité des modèles qu'il est nécessaire de contrôler. Enfin, malgré toutes ces avancées et en raison de l'évolution de l'environnement du risque, les compagnies d'assurance et de réassurance sont aujourd'hui confrontées à des difficultés de modélisation des risques émergents (épidémies, terrorisme, cybercriminalité, certaines catastrophes naturelles, etc.). Ces derniers semblent particulièrement difficiles à modéliser, notamment en raison du manque de données lié à l'absence de précédent.

1.2 Risque de modèle

En tant que simplification ou approximation de la réalité et non la réalité en elle-même, les modèles sont, par définition, soumis à de potentielles erreurs statistiques. En outre, une utilisation erronée des résultats d'un modèle peut mener à des résultats non conformes à la réalité. Par conséquent, il existe un risque inhérent au modèle mathématique en lui-même mais également à l'usage du modèle et de ses résultats. Ces deux grandes classes de risque d'erreur sont généralement regroupées dans la littérature sous le terme de risque de modèle. Ces deux classes sont nommées dans l'article du CRO Forum [2017] comme étant : le *risque structurel* et le *risque opérationnel*. Le risque structurel est généré par la nature même des modèles mathématiques qui sont une représentation simplifiée du monde réel mais également par les données utilisées, l'estimation des paramètres, le choix des modèles, etc. Le risque opérationnel découle, pour sa part, de l'erreur humaine telle que l'erreur d'implémentation ou la mauvaise application du modèle.

En raison de la place prépondérante des modèles dans les compagnies d'assurance et de réassurance, le risque de modèle requiert une attention particulière. La quantifi-

cation précise d'un tel risque reste aujourd'hui un sujet majeur qui suscite un intérêt croissant à la fois de la part de l'industrie et dans le milieu académique. Les articles de Sibbertsen et al. [2008]; Planchet and Therond [2012]; Glasserman and Xu [2014]; Barrieu and Scandolo [2015] constituent un aperçu des travaux les plus représentatifs sur le sujet. De même, plusieurs travaux résumés dans deux mémoires actuariels [Lallement, 2014; Davesne, 2015] ont également été menés au sein de SCOR ces dernières années afin de déterminer une approche quantitative du risque de modèle pouvant être appliquée aux modèles évaluant les risques assurantiels.

En pratique, le risque de modèle ne peut être complètement éliminé mais il existe divers mécanismes permettant de l'atténuer. Cela s'effectue, par exemple, par une identification et communication adéquates des limites et hypothèses du modèle, l'implémentation de systèmes appropriés de gouvernance, de contrôle des processus, etc. Un moyen additionnel d'atténuation du risque de modèle fortement recommandé par l'article du CRO Forum [2017] est la mise en place d'une revue/validation indépendante du modèle permettant de porter un second jugement tant sur la partie opérationnelle que structurelle.

Globalement, le rôle de la validation est de vérifier la qualité des données d'entrée, l'approche de modélisation, la mise en œuvre du modèle et l'utilisation de ses résultats, contribuant ainsi à une amélioration continue de sa validité dans le temps.

La validation d'un modèle s'effectue à l'aide d'un grand éventail d'outils quantitatifs ou qualitatifs permettant d'évaluer son erreur structurelle. Parmi ces méthodes, nous pouvons citer le backtesting qui consiste à tester la pertinence de la modélisation en s'appuyant sur un jeu de données historique, les tests de résistance qui permettent de valider le modèle dans des environnements économiques défavorables ou encore des tests de sensibilité aux hypothèses, paramètres ou données. Ces derniers sont réalisés en comparant, par exemple, les résultats du modèle obtenus avec les valeurs extrêmes des paramètres ou, à partir de données différentes. Cette méthodologie est facile à implémenter et largement utilisée mais limitée car elle ne fournit pas une mesure statistique de l'impact des paramètres ou de la donnée.

Dans cette thèse, nous nous sommes intéressés au risque de modèle structurel et plus particulièrement aux outils d'analyse de sensibilité qui permettent d'effectuer une validation pertinente des différents paramètres utilisés dans un modèle. En effet, comme énoncé précédemment, la gestion des risques supportés par les compagnies d'assurance ou de réassurance repose sur la mise en œuvre de modèles dont les paramètres doivent être préalablement estimés. Cependant, en pratique, l'estimation de ces paramètres contient une part inhérente d'incertitude qu'il est parfois possible de quantifier à l'aide de méthodes statistiques. Dans ce contexte, comment distinguer les paramètres dont l'estimation doit être très précise, de ceux pour lesquels une estimation grossière suffit ?

Les outils développés dans le domaine de l'analyse de sensibilité permettent de répondre à cette question en étudiant comment la réponse du modèle réagit aux variations de ses variables d'entrée. Appliquées initialement dans le domaine industriel

lors de la construction et de l'utilisation d'un modèle numérique de simulation, les méthodes d'analyse de sensibilité sont des outils précieux, dont l'applicabilité dans le domaine de l'assurance et de la réassurance a commencé à être étudiée ces dernières années. Comme souligné par Jacques [2005], ces outils permettent de vérifier si le modèle reflète correctement le phénomène modélisé. En effet, si une variable communément connue comme non influente est déterminée comme fortement influente, l'adéquation du modèle au phénomène sous-jacent ou la compréhension de l'impact des entrées sur la réponse du modèle devra être remise en cause. On peut également déterminer les variables d'entrée dont l'incertitude contribue le plus à l'incertitude de la sortie du modèle, celles qui n'ont pas d'influence ainsi que celles qui interagissent au sein du modèle.

1.3 Analyse de sensibilité

Nous considérons dans le reste du manuscrit η comme un modèle d'évaluation d'un risque assurantiel fonction de plusieurs paramètres $\mathbf{x} = (x_1, \ldots, x_d)$ de sorte que la sortie du modèle y s'exprime de la manière suivante

$$y = \eta(\mathbf{x}), \quad \mathbf{x} \in \mathbb{R}^d, \ y \in \mathbb{R}. \tag{1.1}$$

Nous étudions l'impact des entrées $x_1, \ldots, x_d$ sur la sortie y avec les méthodes de l'analyse de sensibilité. Ces méthodes sont présentées ci-dessous en adoptant la classification suivante proposée par Saltelli et al. [2000] : les méthodes de criblage, les méthodes locales et les méthodes globales.

Les méthodes de criblage ou *screening* analysent qualitativement l'importance des variables d'entrée sur la variabilité de la réponse du modèle. Elles ont pour objectif d'identifier les paramètres ayant un effet sur la sortie, sans en quantifier l'importance, afin de réduire le nombre de paramètres à analyser ultérieurement avec des méthodes plus précises et/ou plus coûteuses. Ces méthodes, n'exigeant que quelques simulations, sont souvent appliquées pour des modèles contenant un nombre considérable de paramètres d'entrée. Toutefois, leur utilisation est limitée car elles ne dispensent qu'une information qualitative sur l'effet des paramètres d'entrée et ne prennent pas compte les effets d'interactions qui peuvent avoir un impact significatif sur la réponse du modèle.

Les méthodes locales d'analyse de sensibilité évaluent quantitativement l'impact d'une petite variation autour d'une valeur nominale $\mathbf{x}^*$ des entrées. Mathématiquement, cela se traduit notamment par le calcul des dérivées partielles $\frac{\partial \eta}{\partial x_j}(\mathbf{x}^*)$, $j = 1, \ldots, d$, de la réponse du modèle par rapport aux différents paramètres d'entrée. Une fois ces valeurs obtenues, nous les comparons entre elles afin de conclure sur l'impact de chaque paramètre localement sur la sortie. Une telle analyse est cependant incomplète, car elle ne tient compte d'aucune information relative à la plage de variation des paramètres, à l'exception d'une valeur nominale. De plus, Saltelli [2006] fait remarquer

que cette méthode est non efficiente si les entrées du modèle sont incertaines ou si la linéarité du modèle en ses paramètres est inconnue.

Enfin, les méthodes d'analyse de sensibilité globale cherchent à combler les défauts des méthodes locales en se détachant de la valeur nominale initiale et en mesurant l'impact d'une variable sur l'ensemble de son intervalle de variation. Dans ce contexte, les entrées ne sont plus considérées comme déterministes mais aléatoires. La distribution respective de chacune est représentative de l'incertitude associée. In fine, la réponse du modèle sera également aléatoire. Nous notons donc $\mathbf{X} = (X_1, \ldots, X_d) \in \mathbb{R}^d$, le vecteur aléatoire représentant les paramètres incertains du modèle et $Y = \eta(\mathbf{X}) \in \mathbb{R}$, la variable aléatoire de sortie résultante. Par conséquent, en explorant simultanément le domaine de variation de chacun des paramètres, ces méthodes permettent d'analyser comment le modèle réagit aux variations jointes des entrées ainsi que la prise en compte d'une potentielle non-linéarité de ce dernier.

Ainsi, utiliser cette dernière classe de méthodes d'analyse de sensibilité amène à porter une analyse critique sur les modèles, les variables d'entrée et leur influence sur le résultat du calcul. En ce sens, nous nous concentrons dans ce manuscrit sur la méthodologie d'analyse de sensibilité globale qui permet d'alimenter une discussion sur la pertinence et la validité des modèles théoriques utilisés pour modéliser les risques assurantiels.

Les méthodes globales ont connu un essor important ces dernières années amenant à les distinguer en plusieurs sous-classes en fonction de la quantité d'intérêt. Nous entendons par quantité d'intérêt une caractéristique de la sortie Y sur laquelle on veut évaluer l'effet des entrées. Les méthodes les plus populaires sont basées sur l'analyse de la variance qui évalue l'impact des entrées sur la valeur moyenne de la sortie. Il est également possible de quantifier l'impact des paramètres sur la distribution complète de la sortie [Borgonovo, 2007] ou sur d'autres quantités d'intérêt telles qu'un quantile de la sortie ou une probabilité de dépassement de seuil comme établi par Fort et al. [2016]. Il est donc primordial, de définir au préalable, la quantité d'intérêt de la sortie sur laquelle on veut quantifier l'impact des entrées afin de ne pas utiliser une méthode inappropriée avec l'objectif de l'étude. Dans cette thèse, nous nous intéressons en particulier aux méthodes basées sur la **variance** qui ont pour but de mesurer comment la variabilité des entrées influe sur celle de la sortie ainsi qu'aux méthodes évaluant l'impact des entrées sur un **quantile** de la sortie.

Les méthodes basées sur la variance les plus populaires sont les indices de Sobol [1993]. Ils sont issus de l'analyse fonctionnelle de la variance (ANOVA) de la sortie Y qui décompose la variance globale de cette dernière en une somme de variances partielles représentant chacune la contribution de chaque sous-ensemble d'entrée. Les indices de Sobol [1993] sont ensuite obtenus en normalisant cette décomposition par la variance totale de la sortie et représentent chacun un pourcentage d'explication de la variance globale. Néanmoins, on a l'unicité de cette décomposition si les entrées sont indépendantes. Lorsque l'hypothèse d'indépendance entre les entrées n'est plus vérifiée, on peut toujours calculer les indices de Sobol, mais leur interprétation peut donner lieu à une analyse erronée de la contribution des entrées à la variabilité de la

sortie.
En raison de la dépendance inhérente de certains risques modélisés en assurance, il était nécessaire d'avoir des indices quantifiant rigoureusement l'impact des paramètres tout en tenant compte de la structure de dépendance. Les indices de Shapley récemment introduits dans le domaine de l'analyse de sensibilité par Owen [2014] répondent à cet objectif comme soulignées par de nombreuses études telles que Song et al. [2016]; Owen and Prieur [2017]; Iooss and Prieur [2019]. Cependant, les estimateurs des indices de Shapley exigent un nombre élevé d'évaluations du modèle η, ce qui les rend difficilement utilisables lorsque celui-ci est coûteux. L'utilisation de modèle de substitution tel que le krigeage[Sacks et al., 1989] permet de réduire ce coût, mais il est alors nécessaire de contrôler l'erreur induite par le métamodèle. L'une de nos **contributions** est d'étendre les méthodes d'estimations d'indices de sensibilité dans le cas d'entrées dépendantes à partir de métamodèle.

Les méthodes évaluant l'impact des entrées sur un quantile de la sortie ont été initialement définies dans Fort et al. [2016] pour des entrées indépendantes. Ces méthodes appelées indices QOSA (Quantile Oriented Sensitivity Analysis) ont ensuite été étudiées par Browne et al. [2017]; Maume-Deschamps and Niang [2018] qui ont chacun proposé un estimateur. Toutefois, ceux-ci restent lourds à manipuler en pratique. Une **contribution** supplémentaire de ce manuscrit sera de proposer un estimateur efficace de ces indices basé sur la méthode des forêts aléatoires [Breiman, 2001]. De plus, à notre connaissance, il n'existe pas à ce jour d'étude du comportement de ces indices dans le cas d'entrées dépendantes. Afin de remédier à cela, nous effectuons un travail préliminaire qui révèle que l'interprétation des indices QOSA en présence de dépendance stochastique entre les entrées n'est pas aisée. Cela nous amène à définir de nouveaux indices que nous nommons ***Indices de Shapley orientés quantile*** qui semblent prometteurs car ils permettent une répartition équitable de l'influence des variables sur un quantile de la sortie en prenant en compte les effets marginaux, d'interactions et de dépendance.

1.4 Organisation du manuscrit et présentation des contributions

Nous avons choisi de structurer notre travail en 4 chapitres qui peuvent être lus indépendamment les uns des autres. Le Chapitre 2 est une revue de la littérature sur les méthodes d'analyse de sensibilité globale, plus exactement les mesures basées sur la variance ainsi que de nouvelles adaptées à la quantité d'intérêt de l'étude. Le Chapitre 3 de cette thèse est constitué de travaux initiés lors du CEMRACS[3] 2017, qui ont mené à la publication de l'article Benoumechiara and Elie-Dit-Cosaque [2019] dans la revue *Esaim : Proceedings and Surveys*. Dans ce chapitre, nous étudions l'impact de la dépendance entre les entrées du modèle sur les indices de Shapley et proposons également une méthode d'estimation de ces derniers basée sur un métamodèle de kri-

[3]CEMRACS : Centre d'Eté Mathématique de Recherche Avancée en Calcul Scientifique

geage. Le Chapitre 4, reprend un article coécrit avec Véronique Maume-Deschamps et, a pour objet la consistance des forêts aléatoires dans le cadre de l'estimation des fonctions de répartition conditionnelles et des quantiles conditionnels et a été soumis pour publication. Enfin, le Chapitre 5 utilise la méthode des forêts aléatoires afin de proposer de nouveaux estimateurs des indices QOSA et sera également soumis pour publication sous peu.
A des fins d'exhaustivité, nous mentionnons que les indices de Sobol et QOSA d'ordre un ont été mis en œuvre sur le modèle de risque opérationnel de SCOR, mais les résultats issus de ces travaux ne seront pas présentés dans ce manuscrit par souci de confidentialité.

1.4.1 Chapitre 2 : État de l'art

Ce chapitre propose une revue de la littérature des outils d'analyse de sensibilité globale dont certains sont étudiés de façon plus détaillée dans les chapitres suivants. Sont tout d'abord introduits les indices de sensibilité basés sur la variance dans le cadre d'entrées indépendantes. Puis, diverses extensions étudiant le cas d'entrées dépendantes sont exposées, avec un intérêt particulier pour les indices fondés sur les valeurs de Shapley [Shapley, 1953]. Ensuite, de nouveaux indices permettant de quantifier l'impact des entrées sur une quantité d'intérêt bien spécifique de la sortie telle que la moyenne, une probabilité de dépassement de seuil sont présentés. Nous nous concentrons notamment sur les indices mesurant l'influence des entrées sur un quantile de la sortie et énonçons plusieurs propriétés pour ces derniers. Il s'avère que leur interprétation peut être délicate en dehors des modèles additifs dans le cas d'entrées indépendantes et pour tout type de modèle en présence de dépendance stochastique entre les entrées. Afin de surmonter ces limitations, nous introduisons des *indices de Shapley subordonnés à une fonction de contraste* qui donnent une information condensée (effets principaux et d'interaction avec les autres variables) de l'influence d'une entrée sur une caractéristique spécifique de la sortie. Nous nous intéressons en particulier aux *indices de Shapley orientés quantile* qui permettent une interprétation claire de l'impact de chaque entrée pour tout type de modèle à la fois dans le cas d'entrées indépendantes et dépendantes. Ces nouveaux indices évaluent l'influence des entrées sur une caractéristique particulière de la sortie et complètent ainsi les indices de Shapley reposant sur la mesure de Borgonovo définis par Derennes [2019] qui quantifient l'influence de chacune des entrées - avec prise en compte de tous les effets d'interaction - sur la distribution globale de la sortie.

1.4.2 Chapitre 3 : Indices de Shapley calculés avec un métamodèle de krigeage et quantification des erreurs Monte Carlo et métamodèle

Les méthodes quantitatives standards de l'analyse de sensibilité globale d'un modèle numérique, $Y = \eta(\mathbf{X})$ avec d variables aléatoires $\mathbf{X} = (X_1, ..., X_d)$, consistent à quantifier les contributions de chacun de ses paramètres d'entrée dans la variabilité de sa

sortie Y, i.e. identifier les variables contribuant le plus à la variabilité de la sortie, les variables les moins influentes et celles qui interagissent entre elles.

Dans le cas d'entrées indépendantes, l'analyse fonctionnelle de la variance résultant des travaux de Hoeffding [1948] permet de décomposer la variance globale de la sortie Y en une somme de variances partielles. Les indices de Sobol [1993] sont obtenus en normalisant cette décomposition par $\mathrm{Var}(Y)$ et représentent chacun un pourcentage d'explication de la variance globale :

$$S_i = \frac{\mathrm{Var}\left(\mathbb{E}\left[Y \middle| X_i\right]\right)}{\mathrm{Var}(Y)}, \qquad S_{ij} = \frac{\mathrm{Var}\left(\mathbb{E}\left[Y \middle| X_i, X_j\right]\right)}{\mathrm{Var}(Y)} - S_i - S_j, \qquad \dots$$

L'indice de sensibilité du premier ordre S_i mesure la part de la variance du modèle qui est due à la variable X_i, l'indice de sensibilité du second ordre S_{ij} mesure la part de la variance du modèle qui est due à l'interaction entre X_i et X_j et ainsi de suite pour les ordres supérieurs. Des indices de sensibilité totaux ont également été définis par Homma and Saltelli [1996] afin d'exprimer la sensibilité totale de la variance de Y due à une variable X_i :

$$ST_i = S_i + \sum_{i \neq j} S_{ij} + \dots + S_{1,\dots,d} = \frac{\mathbb{E}_{\mathbf{X}_{-i}}\left[\mathrm{Var}_{X_i}\left(Y \middle| \mathbf{X}_{-i}\right)\right]}{\mathrm{Var}(Y)} \ ,$$

où $\mathbf{X}_{-i}$ est le vecteur $\mathbf{X} = (X_1, \dots, X_d)$ privé de X_i.

Cependant, cette décomposition n'est plus valable avec des entrées dépendantes et l'interprétation des indices de Sobol classiques est erronée dans ce cas. Plusieurs travaux ont été effectués pour traiter cette difficulté et étendre les indices de Sobol au cas de dépendance stochastique entre les entrées. Parmi ces travaux, nous pouvons citer Li et al. [2010]; Kucherenko et al. [2012]; Chastaing et al. [2012]; Mara et al. [2015]. Toutefois, l'estimation de ces indices ainsi que leur interprétation restent assez difficiles en pratique. Par exemple, les indices développés dans Chastaing et al. [2012] peuvent être négatifs [Chastaing et al., 2015]. Cela est conceptuellement problématique car, comme déclaré dans Owen and Prieur [2017], une variable dont la fonction ne dépend pas du tout aura une importance nulle et sera donc plus importante qu'une variable dont la fonction dépend vraiment d'une façon qui lui attribue une importance négative.

Les valeurs de Shapley [1953] issues de la théorie des jeux ont récemment été proposées par Owen [2014] en analyse de sensibilité pour quantifier la contribution de la i-ème entrée à la variance de la sortie via l'expression

$$Sh^i = \sum_{\mathcal{J} \subseteq \mathcal{D} \setminus \{i\}} \frac{(d - |\mathcal{J}| - 1)!|\mathcal{J}|!}{d!} \left(c\left(\mathcal{J} \cup \{i\}\right) - c\left(\mathcal{J}\right)\right) \ , \tag{1.2}$$

où $c(\mathcal{J}) = \mathrm{Var}\left(\mathbb{E}\left[Y \middle| \mathbf{X}_{\mathcal{J}}\right]\right) / \mathrm{Var}(Y)$ est une fonction de coût mesurant la variance de Y causée par l'incertitude des entrées dans $\mathcal{J}$, $\mathcal{J} \subseteq \mathcal{D}$. Song et al. [2016] ont montré qu'utiliser la fonction de coût suivante $c(\mathcal{J}) = \mathbb{E}\left[\mathrm{Var}\left(Y \middle| \mathbf{X}_{-\mathcal{J}}\right)\right] / \mathrm{Var}(Y)$ donnait des indices de Shapley équivalents. Certaines études telles que Owen and Prieur [2017];

Iooss and Prieur [2019] ont mis en évidence le potentiel de ces indices dans le cas d'entrées dépendantes. Dans ce dernier cas, les indices de Shapley constituent une bonne alternative aux extensions existantes des indices de Sobol mentionnés ci-dessus car 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.

Une procédure d'estimation des indices de Shapley basée sur une méthode d'échantillonnage Monte-Carlo a initialement été proposée par Castro et al. [2009], puis améliorée par Song et al. [2016]. Néanmoins, celle-ci requiert des échantillons de grande taille afin d'obtenir une faible erreur d'estimation. Mais quand il s'agit de modèles coûteux, une estimation précise de ces indices peut être difficile à réaliser, voire irréalisable. Pour pallier cette difficulté, nous proposons d'utiliser un métamodèle. Il s'agit d'un modèle mathématique approximant le modèle numérique sous-étude construit sur une base d'apprentissage. Son principal avantage est d'être plus rapide à calculer que l'original. Nous utilisons le métamodèle de krigeage qui a démontré, dans de nombreuses situations pratiques, de bonnes capacités de prédiction (cf. Le Gratiet et al. 2014 par exemple). Ainsi, nous substituons la fonction exacte $Y = \eta(\mathbf{X})$ par le processus Gaussien noté $H_n(\mathbf{X})$ dans la fonction de coût de (1.2), ce qui nous donne une nouvelle quantité dénotée par Sh_n^i.

L'estimateur $\widehat{Sh}_n^i$ de ce nouveau terme présente donc deux sources d'incertitude : la première provenant de l'approximation par métamodèle et la seconde issue de l'échantillonnage Monte Carlo. Inspirés par l'idée utilisée dans Le Gratiet et al. [2014] pour les indices de Sobol et l'échantillonnage Bootstrap que nous avons intégré dans l'algorithme de Song et al. [2016], nous avons développé une méthode dans Benoumechiara and Elie-Dit-Cosaque [2019] permettant de quantifier ces deux types d'incertitude en décomposant la variance de $\widehat{Sh}_n^i$ comme suit

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

où $\mathrm{Var}_H\left(\mathbb{E}_X\left[\widehat{Sh}_n^i | H_n(x)\right]\right)$ est l'incertitude résultant de l'approximation par le métamodèle et $\mathrm{Var}_X\left(\mathbb{E}_H\left[\widehat{Sh}_n^i | \left(\mathbf{X}_{\kappa_l}\right)_{l=1,\dots,B}\right]\right)$ est celle liée à l'échantillonnage Monte-Carlo. Il est à noter qu'en décomposant la variance de $\widehat{Sh}_n^i$ de cette façon, nous supposons implicitement qu'il n'y a pas d'effet d'interaction entre le métamodèle $H_n(x)$ et l'échantillon Monte Carlo $(\mathbf{X}_{\kappa_l})_{l=1,\dots,B}$.
Une implémentation des estimateurs des indices de Shapley ainsi que de l'algorithme quantifiant les deux sources d'incertitude évoquées précédemment est mise à disposition au sein d'un package `python` nommé `shapley-effects` [Benoumechiara and Elie-Dit-Cosaque, 2018].

1.4.3 Chapitre 4 : Estimation de fonction de répartition conditionnelle et de quantile conditionnel basée sur les forêts aléatoires

L'estimation de fonction de répartition conditionnelle et de quantile conditionnel relève d'une importance majeure dans plusieurs domaines dont l'assurance, l'industrie mais aussi pour les indices QOSA. Les méthodes classiques paramétriques (e.g. régression quantile) ou non-paramétriques (e.g. méthodes à noyau) dédiées à cela présentent certaines limites dans la pratique. En effet, la performance des méthodes à noyau dépend fortement du choix de la fenêtre et se dégrade rapidement quand le nombre de covariables augmente. D'autre part, la régression quantile introduite par Koenker and Bassett Jr [1978] s'avère inadaptée dans un cadre non Gaussien puisque le vrai quantile conditionnel ne s'exprime pas forcément comme une combinaison linéaire des variables d'entrée. Dans le Chapitre 4, nous proposons d'explorer la méthode des forêts aléatoires.

L'algorithme des forêts aléatoires, introduit par Breiman [2001], a montré de très bonnes performances en pratique pour estimer la fonction de régression $\mathbb{E}[Y|\mathbf{X}=\mathbf{x}]$ dans le cadre de problèmes complexes (relations non linéaires, interactions, grande dimension, etc.). La méthode met en œuvre k arbres CART [Breiman et al., 1984] construits à partir d'échantillons bootstrap de la base d'apprentissage $\mathcal{D}_n = (\mathbf{X}^j, Y^j)_{j=1,\dots,n}$ qui sont ensuite agrégés pour améliorer la performance des arbres individuels.

En notant que la fonction de répartition conditionnelle est un cas particulier de fonction de régression en raison de l'égalité suivante

$$F(y|\mathbf{X}=\mathbf{x}) = \mathbb{P}(Y \leqslant y|\mathbf{X}=\mathbf{x}) = \mathbb{E}\left[\mathbb{1}_{\{Y\leqslant y\}}\middle|\mathbf{X}=\mathbf{x}\right] ,$$

nous proposons un premier estimateur basé sur les échantillons bootstrap de chaque arbre comme suit

$$F_{k,n}^{b}(y|\mathbf{X}=\mathbf{x};\Theta_1,\dots,\Theta_k,\mathcal{D}_n) = \sum_{j=1}^{n} w_{n,j}^{b}(\mathbf{x};\Theta_1,\dots,\Theta_k,\mathcal{D}_n)\,\mathbb{1}_{\{Y^j\leqslant y\}} . \tag{1.3}$$

Les poids sont ici définis par

$$w_{n,j}^{b}(\mathbf{x};\Theta_1,\dots,\Theta_k,\mathcal{D}_n) = \frac{1}{k}\sum_{\ell=1}^{k}\frac{B_j(\Theta_\ell^1,\mathcal{D}_n)\,\mathbb{1}_{\{\mathbf{X}^j\in A_n(\mathbf{x};\Theta_\ell,\mathcal{D}_n)\}}}{N_n^b(\mathbf{x};\Theta_\ell,\mathcal{D}_n)} ,$$

où $\Theta_\ell, \ell = 1,\dots,k$ sont des variables aléatoires i.i.d distribuées de la même façon qu'une variable aléatoire générique $\Theta = (\Theta_1,\Theta_2)$ indépendante de $\mathcal{D}_n$. Cette variable permet de tirer un échantillon bootstrap de $\mathcal{D}_n$ avant de construire chaque arbre et de choisir les variables candidates pour les splits au sein de chaque nœud. Pour le $\ell^{\text{ème}}$ arbre, $A_n(\mathbf{x};\Theta_\ell,\mathcal{D}_n)$ est la cellule de l'espace d'entrée contenant $\mathbf{x}$, $B_j(\Theta_\ell^1,\mathcal{D}_n)$, le nombre de fois que l'observation $(\mathbf{X}^j,Y^j)$ a été tirée de $\mathcal{D}_n$ et $N_n^b(\mathbf{x},\Theta_\ell,\mathcal{D}_n)$, le nombre d'éléments de l'échantillon bootstrap qui sont dans $A_n(\mathbf{x};\Theta_\ell,\mathcal{D}_n)$.

Une autre approche proposée par Meinshausen [2006] est d'engendrer les arbres avec les échantillons bootstrap et d'utiliser uniquement l'échantillon original $\mathcal{D}_n$ lors

de la prédiction. Cela nous donne l'estimateur suivant

$$F^o_{k,n}\left(y|\,\mathbf{X}=\mathbf{x};\Theta_1,\ldots,\Theta_k,\mathcal{D}_n\right)=\sum_{j=1}^{n} w^o_{n,j}\left(\mathbf{x};\Theta_1,\ldots,\Theta_k,\mathcal{D}_n\right)\mathbb{1}_{\{Y^j\leqslant y\}}\ , \tag{1.4}$$

avec

$$w^o_{n,j}\left(\mathbf{x};\Theta_1,\ldots,\Theta_k,\mathcal{D}_n\right)=\frac{1}{k}\sum_{\ell=1}^{k}\frac{\mathbb{1}_{\{\mathbf{X}^j\in A_n(\mathbf{x};\Theta_\ell,\mathcal{D}_n)\}}}{N^o_n\left(\mathbf{x};\Theta_\ell,\mathcal{D}_n\right)}\ . \tag{1.5}$$

$N^o_n(\mathbf{x},\Theta_\ell,\mathcal{D}_n)$ réfère maintenant au nombre d'éléments de l'échantillon d'apprentissage $\mathcal{D}_n$ qui sont dans la feuille $A_n(\mathbf{x};\Theta_\ell,\mathcal{D}_n)$.

Nous énonçons, dans le Chapitre 4, deux théorèmes montrant la consistance de ces estimateurs sous plusieurs hypothèses détaillées et justifiées dans celui-ci. Toutefois, les éléments essentiels à retenir sont qu'un bon contrôle sur la variation de la fonction de répartition conditionnelle au sein de chaque cellule ainsi qu'un choix approprié du nombre d'individus dans les feuilles sont suffisants pour assurer la consistance des estimateurs. Une hypothèse supplémentaire stipulant que la fonction de répartition conditionnelle doit avoir une certaine régularité permet d'obtenir la convergence uniforme presque sûre des estimateurs précédents comme déclarés dans les Théorèmes 4.4.1 et 4.4.2 reproduits ci-dessous.

Théorème 4.4.1.
Supposons que les Hypothèses 4.4.1 à 4.4.3 sont satisfaites, alors

$$\forall\mathbf{x}\in\mathcal{X},\quad \sup_{y\in\mathbb{R}}\left|F^b_{k,n}\left(y|\,\mathbf{X}=\mathbf{x}\right)-F\left(y|\,\mathbf{X}=\mathbf{x}\right)\right|\xrightarrow[n\to\infty]{a.s.}0$$

Théorème 4.4.2.
Supposons que les Hypothèses 4.4.1 à 4.4.3 sont satisfaites, alors

$$\forall\mathbf{x}\in\mathcal{X},\quad \sup_{y\in\mathbb{R}}\left|F^o_{k,n}\left(y|\,\mathbf{X}=\mathbf{x}\right)-F\left(y|\,\mathbf{X}=\mathbf{x}\right)\right|\xrightarrow[n\to\infty]{a.s.}0$$

Il doit être noté que le résultat du Théorème 4.4.1 est à notre connaissance le premier intégrant la composante bootstrap de l'algorithme original de Breiman [2001]. Il est également important de mentionner que Meinshausen [2006] a montré la consistance de l'estimateur (1.4) en utilisant un modèle simplifié de forêts aléatoires. En effet, il considère dans sa preuve que les poids (1.5) sont non aléatoires. Nous allons plus loin en montrant la consistance du vrai estimateur, i.e. avec les poids aléatoires.

Ainsi, nous pouvons estimer les quantiles conditonnels en pluggant les estimateurs (1.3) ou (1.4) au lieu de $F\left(y|\,\mathbf{X}=\mathbf{x}\right)$. La consistance de ceux-ci peut être facilement établie par des arguments standards à partir de la consistance des estimateurs de la fonction de répartition conditionnelle.

Une implémentation des estimateurs proposés est disponible au sein d'un package `Julia` nommé `ConditionalDistributionForest` [Fabrège and Maume-Deschamps, 2020] ainsi que dans un package `python` appelé `qosa-indices` [Elie-Dit-Cosaque, 2020]. Elle permet d'investiguer la performance de ces derniers notamment à travers une application numérique développée dans le Chapitre 4.

1.4.4 Chapitre 5 : Estimation des indices QOSA basée sur les forêts aléatoires

Les indices étudiés dans le Chapitre 3 constituent des outils intéressants d'analyse de sensibilité si on s'intéresse à une caractéristique particulière de la distribution de la sortie Y : la moyenne $\mathbb{E}[Y]$ du modèle numérique. En effet, ils permettent de quantifier les variables qui influent le plus sur la moyenne en utilisant la variance comme mesure de distance. En revanche, si on considère une autre caractéristique de la distribution Y comme quantité d'intérêt, par exemple, un quantile d'ordre α, il semble intuitif qu'un quantile extrême puisse être sensible à des variables différentes de celles qui influent sur la moyenne.

De ce fait, lorsque la quantité d'intérêt est le quantile d'ordre $\alpha \in]0,1[$ de la distribution Y, des indices adaptés appelés QOSA (Quantile Oriented Sensitivity Analysis) basés sur une fonction de contraste spécifique ont été développés dans Fort et al. [2016] afin de déterminer les variables d'entrée les plus influentes :

$$S_i^\alpha = 1 - \frac{\mathbb{E}\left[\min_{\theta\in\mathbb{R}} \mathbb{E}\left[\psi_\alpha\left(Y,\theta\right)|\, X_i\right]\right]}{\min_{\theta\in\mathbb{R}} \mathbb{E}\left[\psi_\alpha\left(Y,\theta\right)\right]} = 1 - \frac{\mathbb{E}\left[\psi_\alpha\left(Y, q^\alpha\left(Y|\, X_i\right)\right)\right]}{\mathbb{E}\left[\psi_\alpha\left(Y, q^\alpha\left(Y\right)\right)\right]}\,, \quad i = 1,\dots,d, \tag{1.6}$$

avec $\psi_\alpha : (y,\theta) \mapsto (y-\theta)\left(\alpha - \mathbb{1}_{\{y\leqslant\theta\}}\right)$.

L'indice S_i^α compare donc la distance moyenne entre Y et son quantile conditionnel à la distance moyenne entre Y et son quantile, où la distance considérée est la fonction de contraste ψ_α. Plusieurs propriétés de ces indices sont présentées dans le Chapitre 2, dans lequel l'impact de la dépendance stochastique entre entrées est également investigué.

Des estimateurs de l'indice QOSA d'ordre un basés sur des méthodes à noyau ont été proposés dans Browne et al. [2017]; Maume-Deschamps and Niang [2018] mais ceux-ci restent lourds à manipuler. En pratique, on est confronté à la détermination de la fenêtre optimale lors de l'utilisation de méthodes à noyau. De plus, deux jeux de données $\mathcal{D}_n = \left(\mathbf{X}^j, Y^j\right)_{j=1,\dots,n}$ et $\mathcal{D}_n^\diamond = \left(\mathbf{X}^{\diamond j}, Y^{\diamond j}\right)_{j=1,\dots,n}$ sont nécessaires pour utiliser l'estimateur de Maume-Deschamps and Niang [2018]. Cela peut s'avérer problématique si on traite avec des modèles coûteux. L'estimateur de Browne et al. [2017] exige, quant à lui, un jeu de données complet $\mathcal{D}_n = \left(\mathbf{X}^j, Y^j\right)_{j=1,\dots,n}$ et un partiel $\left(\mathbf{X}^{\diamond j}\right)_{j=1,\dots,n}$. Toutefois, son estimateur fait intervenir la densité de la $i^{\text{ème}}$ variable d'entrée qui n'est malheureusement pas toujours connue. La principale contribution du Chapitre 5 est donc de proposer différents estimateurs de l'indice QOSA, basés sur la méthode des forêts aléatoires, afin de s'affranchir des inconvénients évoqués précédemment. Ceux-ci sont implémentés au sein d'un package `python` intitulé `qosa-indices` [Elie-Dit-Cosaque, 2020].

Parmi les différents estimateurs expérimentés, un nous apparaît particulièrement intéressant car il présente de bonnes performances tout en exigeant uniquement un

seul jeu de données $\mathcal{D}_n = (\mathbf{X}^j, Y^j)_{j=1,\dots,n}$. Celui-ci est exposé ci-après.

La quantité $\mathbb{E}\left[\psi_\alpha\left(Y, q^\alpha\left(Y\right)\right)\right]$ dans (1.6) est estimée comme suit

$$\widehat{P}_1 = \frac{1}{n}\sum_{j=1}^{n} \psi_\alpha\left(Y^{\diamond j}, \widehat{q}^\alpha(Y)\right) \ ,$$

où $\widehat{q}^\alpha(Y)$ est l'estimateur empirique classique $q^\alpha\left(Y\right)$ calculé avec $\mathcal{D}_n$.

Ensuite, nous tirons parti de la structure des arbres de la forêt aléatoire afin d'estimer le terme $\mathbb{E}\left[\min_{\theta\in\mathbb{R}} \mathbb{E}\left[\psi_\alpha\left(Y,\theta\right)|\, X_i\right]\right]$. Considérons qu'une forêt aléatoire est construite avec les observations $\mathcal{D}_n^i = \left(X_i^j, Y^j\right)_{j=1,\dots,n}$ issues de $\mathcal{D}_n$, i.e. en expliquant Y avec uniquement la composante X_i. Il apparaît que pour le $\ell^{\text{ème}}$ arbre de la forêt, les observations se trouvant dans sa $m^{\text{ème}}$ feuille approximent la distribution conditionnelle de Y étant donné un certain point $X_i = x_i$, ce qui permet d'estimer le minimum de l'espérance conditionnelle $\min_{\theta\in\mathbb{R}} \mathbb{E}\left[\psi_\alpha\left(Y,\theta\right)|\, X_i = x_i\right]$. De plus, en notant que les feuilles sont distribuées selon la distribution de X_i, on peut donc faire la moyenne sur celles-ci pour traiter l'espérance portant sur X_i. Ainsi, en notant N_{leaves}^ℓ comme le nombre de feuilles du $\ell^{\text{ème}}$ arbre et $\mathcal{L}_{\ell,m}^o$ comme l'ensemble des observations de la base d'apprentissage $\mathcal{D}_n^i$ tombant dans la $m^{\text{ème}}$ feuille de celui-ci, nous pouvons alors définir l'estimation associée au $\ell^{\text{ème}}$ arbre du terme $\mathbb{E}\left[\min_{\theta\in\mathbb{R}} \mathbb{E}\left[\psi_\alpha\left(Y,\theta\right)|\, X_i\right]\right]$:

$$\frac{1}{N_{leaves}^\ell}\sum_{m=1}^{N_{leaves}^\ell}\left(\min_{p\in\mathcal{L}_{\ell,m}^o}\sum_{j\in\mathcal{L}_{\ell,m}^o}\frac{\psi_\alpha\left(Y^j, Y^p\right)}{|\mathcal{L}_{\ell,m}^o|}\right) \ .$$

Les estimateurs issus des k arbres sont alors agrégés pour former l'estimateur forêt aléatoire noté par $\widehat{Q}_i^{2,o}$:

$$\widehat{Q}_i^{2,o} = \frac{1}{k}\sum_{\ell=1}^{k}\left[\frac{1}{N_{leaves}^\ell}\sum_{m=1}^{N_{leaves}^\ell}\left(\min_{p\in\mathcal{L}_{\ell,m}^o}\sum_{j\in\mathcal{L}_{\ell,m}^o}\frac{\psi_\alpha\left(Y^j, Y^p\right)}{|\mathcal{L}_{\ell,m}^o|}\right)\right] \ .$$

Finalement, nous obtenons l'estimateur suivant pour calculer l'indice QOSA

$$\widehat{S}_i^\alpha = 1 - \frac{\widehat{Q}_i^{2,o}}{\widehat{P}_1} \ , \quad i = 1,\dots,d \ . \tag{1.7}$$

Il permet d'avoir une valeur approchée des d indices $S_1^\alpha, \dots, S_d^\alpha$ à l'aide de n évaluations du modèle.

La qualité de l'estimateur (1.7) s'est avérée être hautement dépendante de certains hyperparamètres de la forêt aléatoire. Différentes procédures ont été proposées dans le Chapitre 5 afin de les ajuster correctement, ce qui permet in fine d'améliorer significativement la performance de notre estimateur.

Afin de tester la qualité de l'estimateur (1.7) et de le comparer avec l'estimateur de Browne et al. [2017] (resp. Maume-Deschamps and Niang [2018]) noté $\widetilde{S}_i^\alpha$ (resp. $\check{S}_i^\alpha$), nous illustrons nos résultats sur un exemple de l'article Fort et al. [2016]. Dans l'exemple considéré, la sortie du modèle est donnée par $Y = X_1 - X_2$ où $X_1, X_2 \sim \mathcal{E}(1)$ avec X_1 et X_2 indépendantes. Les auteurs donnent les valeurs explicites des indices QOSA associés aux entrées X_1 et X_2. Nous avons représenté dans le Tableau 1.1, le RMSE (Root Mean Square Error) entre $\widehat{S}_i^\alpha$ et S_i^α, pour différents niveaux de confiance α. Le nombre de simulations n est fixé à 10000 et l'estimateur $\widehat{S}_i^\alpha$ a été calculé avec une forêt contenant 100 arbres. Nous rappelons que le RMSE des variables $X_i, i = 1, 2$ est défini comme la quantité suivante :

$$RMSE_i^\alpha = \sqrt{\frac{1}{s}\sum_{j=1}^{s}\left(\widehat{S}_i^{\alpha,j} - S_i^\alpha\right)^2}.$$

	$\widehat{S}_i^\alpha$ avec $\widehat{Q}_i^{2,o}$		$\widetilde{S}_i^\alpha$		$\check{S}_i^\alpha$	
	X_1	X_2	X_1	X_2	X_1	X_2
$\alpha = 0.1$	0.009	0.006	0.061	0.006	0.020	0.044
$\alpha = 0.25$	0.009	0.006	0.042	0.012	0.013	0.036
$\alpha = 0.5$	0.008	0.007	0.027	0.025	0.019	0.021
$\alpha = 0.75$	0.008	0.008	0.014	0.042	0.035	0.012
$\alpha = 0.99$	0.006	0.018	0.013	0.11	0.084	0.071

Table 1.1 *RMSE de l'estimateur basé sur la forêt aléatoire : $\widehat{S}_i^\alpha$ et ceux utilisant les méthodes à noyau : $\widetilde{S}_i^\alpha$ et $\check{S}_i^\alpha$.*

Comme nous pouvons l'observer, nous obtenons des RMSE plus faibles pour notre estimateur que ceux basés sur les méthodes à noyau tout en exigeant moins de données. Cela met donc en exergue ses qualités d'un point de vue pratique. Cependant, il serait intéressant dans un travail ultérieur d'obtenir des garanties théoriques telles que la consistance ou la normalité asymptotique pour cet estimateur.

Chapter 2

State of the Art

Abstract
This chapter discusses global sensitivity measures. The most common variance-based sensitivity indices for independent random input variables are first introduced. Some extensions dealing with the case of dependent random input variables are subsequently described, including the Shapley effects which have good properties in this framework. Then, new indices dealing with the sensitivity of features other than the variance are presented. A focus is made on those quantifying the impact of inputs on a quantile of the output distribution. Their intrepretation turns out to be tricky for non-additive models with independent inputs and for any type of model when using dependent inputs. We therefore propose a new index named *Quantile oriented Shapley effect* overcoming these issues and that fits both independent and dependent inputs.

Contents

2.1 Introduction

Consider a model $Y = \eta(\mathbf{X})$ with d random inputs denoted by $\mathbf{X} = (X_1, X_2, \ldots, X_d)$. Let $\mathbf{X}_{\mathcal{J}}$ indicate the vector of inputs corresponding to the index set $\mathcal{J} \subseteq \mathcal{D}$ where $\mathcal{D} = \{1, 2, \ldots, d\}$. The term "input" stands for any quantity that can be modified in the model prior to execution (variable, parameter, initial condition, etc.) and $\eta : \mathbb{R}^d \to \mathbb{R}$ is a deterministic squared integrable function and $Y \in \mathbb{R}$ the model output random variable.

Sensitivity Analysis (SA) is defined by Saltelli et al. [2004] as "the study of how the uncertainty in the output of a model can be apportioned to different sources of uncertainty in the model input". Various tools exist today to perform a SA (see e.g. Iooss and Lemaître [2015] for a review of SA methods) but it is necessary to distinctly specify the goal of such an analysis. "Do we, for instance, wish to understand whether model inputs are involved in interactions or whether a certain model input is a key uncertainty driver?" as stated by Borgonovo et al. [2017]. These are two distinct questions that may require different approaches to address them depending on the problem. A poor formulation of the sensitivity question may lead the practitioner to use an inappropriate method and, thereby, obtain only a partially informative (if not wrong) answer to the problem at hand. Hence, SA should be performed according to a range of general conceptual objectives, called SA settings, properly defined in various references such as Saltelli et al. [2004]; Borgonovo et al. [2017]. The objectives set by Saltelli et al. [2004] are recalled below:

- **Factor Prioritization** (FP) setting: the goal is to identify the key input drivers of the model behavior. Thus, a possible reduction of the uncertainty affecting these inputs may lead to the largest reduction of the output uncertainty;
- **Factor Fixing** (FF) setting: the aim is to identify the noninfluential inputs which could be fixed at some given values without any loss of information about the output;
- **Variance Cutting** (VC) setting: the aim is to identify which inputs should be fixed so as to reach a target value on the output variance;
- **Factor Mapping** (FM) setting: the aim is to identify the key inputs responsible for producing values of the output in a given region of interest.

Borgonovo et al. [2017] suggests adding to those:

- **Model Structure** (MS) setting: the aim is to analyze the possible interactions between inputs;
- **Sign of Change** (SC) setting: the aim is to identify whether an increase in the inputs gives rise to an increase or a decrease in the model's output;
- **Stability** (S) setting : the aim is to analyze whether perturbations in the inputs may cause the preferred alternative to change.

Methods of SA, allowing to address one or more of these settings simultaneously,

can be broadly categorized into three major classes: *screening methods*, *local sensitivity methods* and *global sensitivity methods*. *Screening methods* focus on the qualitative aspects of sensitivity and aim at reducing the input dimensionality by identifying the non-influential inputs with a low computational cost in terms of model evaluation. The method of Morris [1991] is one of the most commonly used screening methods. *Local sensitivity methods* are quantitative techniques, and consist in studying the impact of a local perturbation around the nominal value of an input. This involves looking at the partial derivatives of the model output with respect to the different parameters and then comparing them to understand how these variations affect the output. However, such an analysis is limited because it does not take into account any knowledge about the parameters, except for a nominal value. *Global sensitivity methods* go further by studying the effects of simultaneous variation of the inputs on the model output in their entire domain. Thus, one can look closer at both output variations induced by individual inputs and/or interactions between several of them (i.e., groups of input variables).

We are specifically interested in Global Sensitivity Analysis (GSA) in this manuscript. For a detailed description of all sensitivity analysis methods, the interested reader can refer to the various survey papers dedicated to this topic [Saltelli et al., 2004, 2008; Faivre et al., 2016; Borgonovo and Plischke, 2016; Borgonovo et al., 2017].
In the sequel, the two first sections present variance-based methods dealing with both indepedent and dependent inputs. Then, Section 2.4 introduces Goal Oriented Sensitivity Analysis (GOSA) which allows to quantify the sensitivity over a feature other than the expectation. A special attention is given to indices quantifying the impact of the inputs on the α-quantile of the output in Section 2.5. We also provide some properties of these indices. A preliminary work is carried out in order to understand the impact of the statistical dependence between the inputs over these indices. Facing some interpretation issues, we finally propose in Section 2.6 new indices based on the Shapley values [Shapley, 1953] which seem to be a promising alternative.

2.2 Regression-Based Methods

We introduce in this section GSA methods for linear or monotonic models. These indices rest on the behaviour of $Y = \eta(\mathbf{X})$ given the values of $\mathbf{X}$ assumed with independent marginals herein.

2.2.1 Linear model

Suppose that the model input-output mapping can be accurately fitted by a linear model as follows

$$Y = \beta_0 + \sum_{i=1}^{d} \beta_i X_i \ .$$

As the model inputs are independently distributed, then the variance of the model output Y is given by

$$\mathrm{Var}\left(Y\right) = \sum_{i=1}^{d} \beta_i^2 \mathrm{Var}\left(X_i\right) \ ,$$

where $\beta_i^2 \mathrm{Var}\left(X_i\right)$ is the part of variance due to the variable X_i. The sensitivity of Y to X_i can therefore simply be quantified by the ratio of the share of variance due to X_i to the total variance [Helton, 1993; Kleijnen and Helton, 1999]. That defines the following Standardized Regression Coefficient (SRC) index

$$SRC_i = \frac{\beta_i^2 \mathrm{Var}\left(X_i\right)}{\mathrm{Var}\left(Y\right)} \ .$$

Representing a share of the variance, the index SRC_i is always positive ($SRC_i \in [0, 1]$). Hence, the closer the index SRC_i is to 1, the more influential the variable is.

2.2.2 Monotonic model

When the model η is no longer linear but still monotonic, the previously described sensivity indices cannot be used directly, but they can be modified so that they fit the model structure. First, we carry out a rank transformation [Iman and Conover, 1979] substituting the raw data with the corresponding ranks. This means that for a given sample, the variables are ordered with their values. For the output Y for instance, we attribute 1 to the highest value and n to the lowest value. This transformation is done for each of the X_i. We then obtain a matrix of model-input ranks denoted by $R_{\mathbf{X}}$ and a vector of model-output ranks named R_Y. The non-parametric sensitivity indices are then derived by using the following linear regression model

$$R_Y = \beta_0^R + \sum_{i=1}^{d} \beta_i^R R_{X_i} \ .$$

One can then define the Standardized Rank Regression Coefficient (SRRC) as

$$SSRC_i = \frac{\left(\beta_i^R\right)^2 \mathrm{Var}\left(R_{X_i}\right)}{\mathrm{Var}\left(R_Y\right)} \ .$$

Thereafter, we will see indices that are suitable sensitivity measures in the case where Y is a function that is neither linear nor monotonous of the inputs.

2.3 Variance-based sensitivity indices

Variance-based methods are common tools in the analysis of complex physical phenomenons. Most of them rest on an ANalysis Of VAriance (ANOVA) of the model

output and were originally defined for independent inputs as stated hereafter. But, this assumption is rarely verified in pratice. Recent approaches overcoming this issue were developed and are also presented in the sequel.

2.3.1 Sobol' indices with independent inputs

We suppose in this section that the inputs are independently and uniformly distributed within the unit hypercube, i.e. $\mathbf{X} \sim \mathcal{U}([0,1]^d)$. However, the following results can be extended to any marginal distributions.

Sobol' sensitivity indices stem from the works of Fisher and Mackenzie [1923] and Hoeffding [1948] on the U-statistics taken up by various authors over time such as Efron and Stein [1981]. Those ultimately lead to a functional ANOVA expansion of the model output η:

$$\eta(\mathbf{X}) = \eta_0 + \sum_{i=1}^{d} \eta_i(X_i) + \sum_{1 \leqslant i < j \leqslant d} \eta_{ij}(X_i, X_j) + \cdots + \eta_{1,\ldots,d}(\mathbf{X}) \ . \tag{2.1}$$

Thus, η can be decomposed into 2^d functions of increasing dimensionality. But this expansion also called High Dimensional Model Representations [Li et al., 2001] is not unique due to the infinite possible choices for these terms. To ensure the uniqueness of (2.1), the following orthogonality constraint over the components is necessary [Sobol, 1993]

$$\int_0^1 \eta_u\left(\mathbf{X_u}\right) \mathrm{d}X_i = 0 \quad \forall i \in u, \ \forall u \subseteq \mathcal{D} \ . \tag{2.2}$$

A consequence of the condition (2.2) is that the terms of (2.1) are orthogonal with each other, i.e.

$$\int_{[0,1]^d} f_u\left(\mathbf{X_u}\right) f_v\left(\mathbf{X_v}\right) \mathrm{d}\mathbf{X} = 0 \quad \forall u \neq v \subseteq \mathcal{D} \ , \tag{2.3}$$

and that the components of the functional decomposition can be univocally defined in terms of conditional expected values. In particular,

$$\begin{aligned} \eta_0 &= \mathbb{E}\left[Y\right] \ , \\ \eta_i(X_i) &= \mathbb{E}\left[Y|X_i\right] - \eta_0 \ , \\ \eta_{ij}(X_i, X_j) &= \mathbb{E}\left[Y|X_i, X_j\right] - \eta_i - \eta_j - \eta_0 \ , \end{aligned}$$

where η_0 is a constant representing the mean response of $\eta(\mathbf{X})$, the first-order component functions η_i, $i \in [\![1, d]\!]$ give the independent contribution to $\eta(\mathbf{X})$ by the i-th input variable acting alone, the functions η_{ij}, $i < j \in [\![1, d]\!]$ describe the pair correlated contribution of the input variables X_i and X_j upon the output $\eta(\mathbf{X})$ and so on.

Thanks to (2.2), representation (2.1) leads to the functional ANalysis Of VAriance which decomposes the global variance of the model output into a sum of partial

variances such as

$$\mathrm{Var}(Y) = \sum_{i=1}^{d} V_i + \sum_{1 \leqslant i < j \leqslant d} V_{ij} + \cdots + V_{1,\ldots,d} \ , \tag{2.4}$$

where

$$\begin{aligned}
V_i &= \mathrm{Var}\left(\eta_i\left(X_i\right)\right) = \mathrm{Var}\left(\mathbb{E}\left[Y \middle| X_i\right]\right) \ , \\
V_{ij} &= \mathrm{Var}\left(\eta_{ij}\left(X_i, X_j\right)\right) = \mathrm{Var}\left(\mathbb{E}\left[Y \middle| X_i, X_j\right]\right) - V_i - V_j \ , \\
V_{ijk} &= \mathrm{Var}\left(\eta_{ijk}\left(X_i, X_j, X_k\right)\right) = \mathrm{Var}\left(\mathbb{E}\left[Y \middle| X_i, X_j, X_k\right]\right) - V_{ij} - V_{ik} - V_{jk} - V_i - V_j - V_k \ , \\
&\vdots \\
V_{1,\ldots,d} &= \mathrm{Var}\left(\eta_{1,\ldots,d}(\mathbf{X})\right) = \mathrm{Var}\left(Y\right) - \sum_{i=1}^{d} V_i - \sum_{1 \leqslant i < j \leqslant d} V_{ij} - \ldots - \sum_{1 \leqslant i_1 < \ldots < i_{d-1} \leqslant d} V_{i_1 \ldots i_{d-1}} \ .
\end{aligned}$$

The so-called Sobol indices given in Sobol [1993] can be derived from (2.4) by dividing both sides with $\mathrm{Var}(Y)$. This operation results in the following property:

$$\sum_{i=1}^{d} S_i + \sum_{1 \leqslant i < j \leqslant d} S_{ij} + \cdots + S_{1,\ldots,d} = 1 \ . \tag{2.5}$$

We then obtain the following $2^d - 1$ sensitivity indices,

$$S_i = \frac{V_i}{\mathrm{Var}(Y)}, \qquad S_{ij} = \frac{V_{ij}}{\mathrm{Var}(Y)}, \qquad \ldots \tag{2.6}$$

where the first-order index S_i (also called main effect) measures the part of variance of the model output that stems from the variability in X_i, the second-order index S_{ij} measures the part of variance of the model output due to the interaction between X_i and X_j and so on for higher interaction orders.

Using the S_i, S_{ij} and higher order indices given above, one can build a picture of the importance of each variable in the output variance. However, when the number of variables is large, this requires the evaluation of $2^d - 1$ indices, which can be too computationally demanding and whose interpretation becomes difficult. For this reason, another popular variance based coefficient called Total-order index by Homma and Saltelli [1996] is used. It measures the contribution to the output variance of X_i, including its main effect as well as all its interaction effects, of any order, with any other input variables. This index is defined by

$$\begin{aligned}
ST_i &= S_i + \sum_{1 \leqslant i < j \leqslant d} S_{ij} + \cdots + S_{1,\ldots,d} \\
&= 1 - \frac{\mathrm{Var}_{\mathbf{X}_{-i}}\left(\mathbb{E}_{X_i}\left[Y \middle| \mathbf{X}_{-i}\right]\right)}{\mathrm{Var}(Y)} = \frac{\mathbb{E}_{\mathbf{X}_{-i}}\left[\mathrm{Var}_{X_i}\left(Y \middle| \mathbf{X}_{-i}\right)\right]}{\mathrm{Var}(Y)} \ ,
\end{aligned} \tag{2.7}$$

where the notation $\mathbf{X}_{-i}$ indicates the set of all variables except X_i. Note that the following property can easily be deduced :

$$0 \leqslant S_i \leqslant ST_i \leqslant 1 \ .$$

Hence, the closer the index ST_i is to 1, the more influential the variable is. It should be noted that the case of equality $S_i = ST_i$ occurs if we have a purely additive model.

It is also possible to define Sobol indices for a group of variables $\mathbf{X}_u, u \subseteq \mathcal{D}$. Given a subset of inputs $u \subseteq \mathcal{D}$, the Sobol index "closed" [Janon, 2012] is established as follows

$$S_u^{Cl} = \frac{\mathrm{Var}\left(\mathbb{E}\left[Y \middle| X_i, i \in u\right]\right)}{\mathrm{Var}\left(Y\right)} \ .$$

The latter quantifies the influence of the variation of the parameters indexed by u taken together. If we want to integrate the effect of the interaction of the inputs u with the other inputs, it suffices to consider the total index associated with u, defined by

$$ST_u = 1 - S^{Cl}_{\mathcal{D}\setminus\{u\}} \ .$$

As established in this section, Sobol indices are well-defined with independent inputs. Indeed, the functional decomposition of the output variance given in (2.4) is unique in this context. This makes possible to clearly identify the contribution of each input or group of inputs to the variance output.

However, in many applications, it is common for inputs to have a statistical dependence structure imposed by a probabilistic dependence function such as a copula function for instance. In such a case, the uniqueness of the functional ANOVA decomposition is no longer guaranteed. The classical Sobol indices can still be calculated but their interpretation becomes difficult. Indeed, as mentioned in Song et al. [2016], the sum of first-order effects may exceed the total variance of the output or the sum of the total effects may be lower than the total variance of the output.

Several works have been carried out to overcome this limitation and extend Sobol indices to the case of stochastic dependence such as Caniou [2012]; Kucherenko et al. [2012]; Mara and Tarantola [2012]. However, none of these works has given an univocal definition of the functional ANOVA decomposition for dependent inputs as the one provided by Sobol [1993] when inputs are independent. A new variable importance measure has been defined in Chastaing et al. [2012] through a generalization of ANOVA when inputs are dependent [Stone, 1994]. But this measure comes with two conceptual problems: it requires some restrictive conditions on the joint probability distribution of the inputs as underlined in Owen and Prieur [2017] and its interpretation remains difficult because it can be negative.

Another approach based on the estimation of four sensitivity indices by input was proposed in Mara et al. [2015]. The combined interpretation of these allows to determine if the importance of an input is due to its marginal contribution or to its dependence with other variables. This strategy is detailed in the following section.

2.3.2 Sobol' indices with dependent inputs

We assume in this section that $\mathbf{X}$ is a set of continuous dependent random variables, with joint probability density function $p_{\mathbf{X}}$. The strategy, proposed in Mara et al. [2015], rests on the Rosenblatt Transformation [Rosenblatt, 1952] defined below that transforms $\mathbf{X}$ into a random vector $\mathbf{U} \sim \mathcal{U}\left([0,1]^d\right)$ with independent and uniformly distributed components.

Definition 2.3.1.
Let $F_{\mathbf{X}}$ be the continuous d-dimensional distribution function of $\mathbf{X} = (X_1, \ldots, X_k, \ldots, X_d)$. The Rosenblatt transformation, $T_R : \mathbb{R}^d \to [0,1]^d$, associated with $F_{\mathbf{X}}$ is defined by

$$T_R : \left| \begin{array}{ccc} \mathbb{R}^d & \to & [0,1]^d \\ \mathbf{x} & \mapsto & \mathbf{u} = \begin{pmatrix} F_1(x_1) \\ \vdots \\ F_{k|1,\ldots,k-1}(x_k|x_1,\ldots,x_{k-1}) \\ \vdots \\ F_{d|1,\ldots,d-1}(x_d|x_1,\ldots,x_{d-1}) \end{pmatrix} \end{array} \right.$$

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

Thus, when applying the Rosenblatt Transformation for the following ordering of the components of $\mathbf{X} = (X_1, \ldots, X_k, \ldots, X_d)$, we obtain the random vector $\mathbf{U} = T_R(\mathbf{X})$ that is uniformly distributed over $[0,1]^d$.

However, in the case of dependent variables, the Rosenblatt Transformation is not unique. There are $d!$ possibilities due to the $d!$ different permutations of the components of $\mathbf{X}$. Note that in the procedure established by Mara et al. [2015], only the d Rosenblatt Transformations obtained after circularly reordering the elements of $\mathbf{X}$ are considered. We denote as $\mathbf{U}^i = (U_1^i, \ldots, U_d^i) = T_{R,i}(\mathbf{X})$ the random vector obtained from the Rosenblatt Transformation of the set $(X_i, X_{i+1}, \ldots, X_d, X_1, \ldots, X_{i-1})$ such as

$$(X_i, X_{i+1}, \ldots, X_d, X_1, \ldots, X_{i-1}) \sim p_{\mathbf{X}} \xrightarrow{T_{R,i}} (U_1^i, \ldots, U_d^i) \sim \mathcal{U}\left([0,1]^d\right) \ . \tag{2.8}$$

It is important to note that this Rosenblatt Transformation corresponds to a particular i-th ordering. Changing this order results in another one.

Such a mapping is bijective and we can consider a function g_i such as $Y = \eta(\mathbf{X}) = g_i(\mathbf{U}^i)$ with $g_i = f \circ T_{R,i}^{-1}$. Because the elements of $\mathbf{U}^i$ are independent, the functional ANOVA decomposition is unique and variance-based sensitivity indices can be computed. Thus, we can write

$$g_i(\mathbf{U}^i) = g_0 + \sum_{k=1}^{d} g_k(U_k^i) + \sum_{k=1}^{d}\sum_{k<l}^{d} g_{k,l}(U_k^i, U_l^i) + \cdots + g_{1,\ldots,d}(U_1^i, \ldots, U_d^i) \ , \tag{2.9}$$

where $g_0 = \mathbb{E}[g_i(\mathbf{U}^i)]$. Because the summands in (2.9) are orthogonal, the variance based decomposition can be derived, such that

$$\mathrm{Var}(Y) = \mathrm{Var}\left(g_i(\mathbf{U}^i)\right) = \sum_{k=1}^{d} V_k + \sum_{k=1}^{d}\sum_{k<l}^{d} V_{k,l} + \cdots + V_{1,\ldots,d} \ , \tag{2.10}$$

where $V_k = \mathrm{Var}\left(\mathbb{E}\left[g_i(\mathbf{U}^i)|U_k^i\right]\right)$, $V_{k,l} = \mathrm{Var}\left(\mathbb{E}\left[g_i(\mathbf{U}^i)|U_k^i, U_l^i\right]\right) - V_k - V_l$ and so on for higher orders. The Sobol' indices are then defined by dividing (2.10) with the total variance. We therefore obtain the first-order Sobol' index

$$S_k^i = \frac{\mathrm{Var}\left(\mathbb{E}\left[g_i(\mathbf{U}^i)|\, U_k^i\right]\right)}{\mathrm{Var}\left(g_i(\mathbf{U}^i)\right)} \ . \tag{2.11}$$

The total Sobol' index can also be written as

$$ST_k^i = \frac{\mathbb{E}\left[\mathrm{Var}\left(g_i(\mathbf{U}^i)|\, \mathbf{U}_{-k}^i\right)\right]}{\mathrm{Var}\left(g_i(\mathbf{U}^i)\right)} \ , \tag{2.12}$$

where $\mathbf{U}_{-k}^i = \mathbf{U}_{\mathcal{D}\setminus\{k\}}^i$. We remind that (2.11) and (2.12) are derived from the Rosenblatt Transformation associated to the ordered set $(X_i, X_{i+1}, \ldots, X_d, X_1, \ldots, X_{i-1})$. The Rosenblatt Transformation used in equation (2.8) determines the following mapping between $\mathbf{X}$ and $\mathbf{U}^i$,

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

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, Mara et al. [2015] proposed to consider only the variables U_1^i and U_d^i due to their 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 are 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

$$S_i^{full} = \frac{\mathrm{Var}\left(\mathbb{E}\left[g_i(\mathbf{U}^i)|\, U_1^i\right]\right)}{\mathrm{Var}\left(g_i(\mathbf{U}^i)\right)} = \frac{\mathrm{Var}\left(\mathbb{E}\left[\eta(\mathbf{X})|\, X_i\right]\right)}{\mathrm{Var}\left(\eta(\mathbf{X})\right)} \ , \tag{2.13}$$

$$ST_i^{full} = \frac{\mathbb{E}\left[\mathrm{Var}\left(g_i(\mathbf{U}^i)|\mathbf{U}_{-1}^i\right)\right]}{\mathrm{Var}\left(g_i(\mathbf{U}^i)\right)} = \frac{\mathbb{E}\left[\mathrm{Var}\left(\eta(\mathbf{X})|(\mathbf{X}_{-i}|X_i)\right)\right]}{\mathrm{Var}\left(\eta(\mathbf{X})\right)} \ , \tag{2.14}$$

  where $\mathbf{X}_{-i}|X_i$ represents 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

$$S_i^{ind} = \frac{\mathrm{Var}\left(\mathbb{E}\left[g_{i+1}(\mathbf{U}^{i+1})|U_d^{i+1}\right]\right)}{\mathrm{Var}\left(g_{i+1}(\mathbf{U}^{i+1})\right)} = \frac{\mathrm{Var}\left(\mathbb{E}\left[\eta(\mathbf{X})|(X_i|\mathbf{X}_{-i})\right]\right)}{\mathrm{Var}\left(\eta(\mathbf{X})\right)} \ , \tag{2.15}$$

$$ST_i^{ind} = \frac{\mathbb{E}\left[\mathrm{Var}\left(g_{i+1}(\mathbf{U}^{i+1})|\mathbf{U}_{-d}^{i+1}\right)\right]}{\mathrm{Var}\left(g_{i+1}(\mathbf{U}^{i+1})\right)} = \frac{\mathbb{E}\left[\mathrm{Var}\left(\eta(\mathbf{X})|\mathbf{X}_{-i}\right)\right]}{\mathrm{Var}\left(\eta(\mathbf{X})\right)} \ , \tag{2.16}$$

where $X_i|\mathbf{X}_{-i}$ represents 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 Rosenblatt Transformation, we can also define the sensitivity indices of $(X_i|\mathbf{X}_u), i = 1, \ldots, d$ and $u \subset \mathcal{D}\backslash\{i\}, u \neq \emptyset$ via U_u^i which represents the effect of X_i without its mutual dependent contribution with $\mathbf{X}_u$.

Hence, these four Sobol indices allow to determine whether an input X_i is important because of its independent contribution or due to its dependence with other variables. But, in practice, drawing a conclusion about the importance of a variable from four indices can be quite complex and it becomes more difficult when the number of variables is relatively large.

Recently, new potential indices called Shapley effects have been proposed in [Owen, 2014]. They present good properties in the presence of dependence such that they cannot be negative, they sum to the total output variance and they are easy to interpret as highlighted by Song et al. [2016]; Iooss and Prieur [2019].

2.3.3 Shapley effects

Shapley values have been introduced in game theory by Shapley [1953]. The motivation, in the context of cooperative game theory, was to define an attribution method to allocate fairly the value created by a team effort to its individual members. Turning now to variance-based sensitivity analysis, it appears that the idea of assigning a portion of the output variance to each input variable has some similarities. These were highlighted and brought to the SA community, in the context of variance-based sensitivity analysis, by Owen [2014].

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. Let $v^i = v^i(c)$, $i = 1, \ldots, d$, be the Shapley value for each player that will be defined below. According to Winter et al. [2002], an attribution method should have the next four compelling properties:

- **Efficiency:** $\sum_{i=1}^{d} v^i = c(\mathcal{D})$. The sum of the values attributed to the players must be equal to what the coalition of all the players can obtain.
- **Symmetry:** if $c(\mathcal{J} \cup i) = c(\mathcal{J} \cup j)$ for all $\mathcal{J} \subseteq \mathcal{D}\backslash\{i, j\}$, then $v^i = v^j$. The contribution of two players should be the same if they contribute equally to all possible coalitions.
- **Dummy:** if $c(\mathcal{J} \cup i) = c(\mathcal{J})$ for all $\mathcal{J} \subseteq \mathcal{D}\backslash\{i\}$, then $v^i = 0$. A player who does not change the predicted value, no matter to which coalition of players it is added, should have a contribution value of 0.

- **Additivity:** if the i-th player has a contribution v^i (resp. v'^i) in the coalitional game described by the gain function c (resp. c'). Then, the contribution of the i-th player in the new coalitional game described by the gain function $c + c'$ is $v^i + v'^i$ for $i \in \mathcal{D}$.

Shapley [1953] showed that the unique valuation v^i that satisfies these properties is the following formula

$$v^i = \sum_{\mathcal{J} \subseteq \mathcal{D} \backslash \{i\}} \frac{(d - |\mathcal{J}| - 1)!|\mathcal{J}|!}{d!} \left(c\left(\mathcal{J} \cup \{i\}\right) - c\left(\mathcal{J}\right)\right) \ , \tag{2.17}$$

defined for the player i with respect to the cost function $c(\cdot)$ and 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 $\mathcal{J} \subseteq \mathcal{D}\backslash\{i\}$. It should also be noted that the weight for each incremental cost of size-s subset of $\mathcal{D}\backslash\{i\}$ in (2.17) can be written as $\frac{(d-s-1)!s!}{d!} = \frac{1}{d}\binom{d-1}{s}^{-1}$.

In the framework of global sensitivity analysis, we can consider the set of inputs of $\eta(\cdot)$ as the set of players $\mathcal{D}$. We then need to define a $c(\cdot)$ cost 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 $c(\mathcal{D}) = \mathrm{Var}(Y)$.

Owen [2014] initially proposed the cost function $\tilde{c}(\mathcal{J}) = \mathrm{Var}\left(\mathbb{E}\left[Y|\mathbf{X}_{\mathcal{J}}\right]\right)$ for considering the Shapley value in the framework of variance-based sensitivity indices with independent inputs. As a matter of fact, in the case of a model not purely additive, the two most used measures, i.e. the first-order and total Sobol indices may fail to sum to the total variance because neither of them adequatly deals with interactions. The first-order index ignores interactions whereas the total one counts them multiplicatively according to Owen [2014]. He therefore proposed an alternative sensitivity measure, based on the Shapley value that always sums to the total variance.

Song et al. [2016] extended this new measure to the case of dependent inputs and showed in their Theorem 1 that the Shapley values defined using cost functions $\tilde{c}(\mathcal{J})$ and $c(\mathcal{J}) = \mathbb{E}\left[\mathrm{Var}\left(Y|\mathbf{X}_{-\mathcal{J}}\right)\right]$ are equivalent. They used the term *Shapley effects* to describe variance based Shapley values as new importance measures in SA. Note that in Owen [2014]; Song et al. [2016], the cost function is not normalized by the variance of Y, whereas, in this manuscript, we consider its normalized version to quantify relative importance of each input with respect to the output variance. We denote hereafter the *Shapley effect* by Sh^i and a generic *Shapley value* by v^i.

Thus, the purpose of the Sobol indices defined in Subsection 2.3.1 is to decompose $\mathrm{Var}(Y)$ and allocate it to *each subset* $\mathcal{J} \subseteq \mathcal{D}$ 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. The Shapley effects rely on an equitable allocation of part of the output variance to each input. During this allocation process, interaction and dependence effects of a subset of inputs are fairly shared with

each individual input within the subset. This fair allocation results in Shapley effects being non negative and sum-up to one, allowing an easy interpretation for ranking input factors. 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. Hence, the closer the index Sh^i is to 1, the more influential the variable is.

It has to be noted that several studies have been carried out on these new indices. In the first place, in case of independent inputs, Owen [2014] showed the Shapley effects are bounded by the Sobol indices, the first-order (resp. total) index as lower (resp. upper) bound. In the second place, in case of dependent inputs, several test-cases where the Shapley effects can be analytically computed have been investigated in Owen and Prieur [2017]; Iooss and Prieur [2019]; Benoumechiara and Elie-Dit-Cosaque [2019] in order to understand the effect of the dependence between inputs on the variance based Shapley values. These studies highlighted several properties of these indices. Benoumechiara and Elie-Dit-Cosaque [2019] have also compared the Shapley effects with the strategy proposed in Mara et al. [2015] based on the estimation of four sensitivity indices per input. In this last case, the Shapley effects can be a good alternative to the existing extensions of classical Sobol indices. Indeed, Shapley effects allow an apportionment of the interaction and dependence contributions between the input involved, making them condensed and easy-to-interpret indices. At last, we can mention that Rabitti and Borgonovo [2019] have recently introduced Shapley-Owen interaction effects which are a generalization of Shapley effects in order to study the synergistic/antagonistic nature of interactions among inputs.

2.4 Sensitivity indices based on contrast functions

In the previous section, we have reviewed variance-based measures. Nevertheless, even if these indices are extremely popular and informative importance measures, they suffer from a major theoretical limitation. Indeed, they only study the impact of the input variables on the expectation of the output distribution by considering the variance as distance measure. However, as highlighted by Borgonovo [2006], in some cases, this one poorly represents the variability/uncertainty of the output distribution. Different approaches have been developped to overcome this issue including, for example, moment independent importance measures proposed by Borgonovo [2007]; Borgonovo et al. [2011] that quantify the influence of an input over the whole distribution of the output.

Another approach presented in Fort et al. [2016] is to define indices that quantify the impact of inputs $\mathbf{X} = (X_1, \ldots, X_d)$ assumed independent on a feature of interest of the output distribution depending on the problem (mean, quantiles and so on). They refer to this method as *Goal-Oriented Sensitivity Analysis* (GOSA). These indices rely on contrast functions (defined below) and are members of a wider family containing sensitivity indices based on dissimilarity measures [Da Veiga, 2015].

The cornerstone of the new index proposed in Fort et al. [2016] is based on a generalization of the first-order Sobol index defined in (2.6). As a matter of fact, by using law of total variance, this one can be expressed as follows

$$S_i = \frac{\mathrm{Var}(Y) - \mathbb{E}\left[\mathrm{Var}\left(Y \middle| X_i\right)\right]}{\mathrm{Var}(Y)} \ . \tag{2.18}$$

Then, it is well-known that the mean $\mathbb{E}[Y]$ is the minimizer of the function $\Gamma_1 : \theta \mapsto \mathbb{E}[\gamma(Y, \theta)]$ with $\gamma : (y, \theta) \mapsto (y - \theta)^2$ the *contrast function*. The minimum value of Γ_1 is $\mathrm{Var}(Y)$. Conditioning by X_i, $\mathbb{E}[Y | X_i]$ is the minimizer of $\Gamma_2 : \theta \mapsto \mathbb{E}[\gamma(Y, \theta) | X_i]$ and the minimum value is $\mathrm{Var}(Y | X_i)$. Accordingly, by comparing the minimum of the function Γ_1 to the expected optimum of Γ_2, the index S_i quantifies the impact of X_i over the value of $\mathbb{E}[Y]$. The idea, followed by the authors in Fort et al. [2016] to set their new index, is to substitute the function γ in the previous expressions to quantify the influence of the inputs over another output's feature. This leads us to define the notion of contrast function.

Definition 2.4.1.
Let us assume that P_Y is some probability measure on the space Q_Y and Θ a feature space corresponding to a feature denoted by $\theta^*(Y)$ of the measure P_Y. A contrast function is then defined as any function ψ given by

$$\psi : \left| \begin{array}{ccc} \Theta & \longrightarrow & L^1(P_Y) \\ \theta & \longmapsto & \psi(\cdot, \theta) : y \in Q_Y \longmapsto \psi(y, \theta) \end{array} \right. , \tag{2.19}$$

and such that

$$\theta^*(Y) := \underset{\theta \in \Theta}{\arg\min}\, \Psi(Y, \theta) \ ,$$

is unique. The function $\Psi : \theta \mapsto \mathbb{E}[\psi(Y, \theta)]$ is called the averaged contrast function.

$\theta^*(Y)$ is the feature of interest of the output distribution on which we want to evaluate the impact of the inputs. For instance, the Sobol index set in (2.18) quantifies the influence of the inputs over the feature $\theta^*(Y) = \mathbb{E}[Y]$ related to the contrast function γ defined above. A list of contrast functions allowing to quantify the impact over various features of the output distribution is provided below:

- the mean: $\psi(y, \theta) = (y - \theta)^2 \quad \Longrightarrow \quad \theta^*(Y) = \mathbb{E}[Y]$
- the median: $\psi(y, \theta) = \frac{1}{2}|y - \theta| \quad \Longrightarrow \quad \theta^*(Y) = q^{0.5}(Y)$
- the α-quantile: $\psi(y, \theta) = (y - \theta)\left(\alpha - \mathbb{1}_{\{y \leqslant \theta\}}\right) \quad \Longrightarrow \quad \theta^*(Y) = q^{\alpha}(Y)$
- an excess probability at t-level: let $Z = \mathbb{1}_{\{Y \geqslant t\}}$ be a intermediate random variable then $\psi(z, \theta) = (z - \theta)^2 \quad \Longrightarrow \quad \theta^*(Z) = \mathbb{E}[Z] = \mathbb{P}(Y \geqslant t)$

We can note that the excess probability at t-level is nothing but a particular case of the mean by using the random variable $\mathbb{1}_{\{Y \geqslant t\}}$. For a more exhaustive list of contrast functions, the reader may refer to Rachdi [2011].

Thus, for the i-th input, the impact of the uncertainty of the parameter X_i over output's feature associated with the contrast function ψ is given by

$$S_i^{\psi} = \frac{\min_{\theta\in\Theta} \mathbb{E}\left[\psi\left(Y,\theta\right)\right] - \mathbb{E}\left[\min_{\theta\in\Theta} \mathbb{E}\left[\psi\left(Y,\theta\right)|\, X_i\right]\right]}{\min_{\theta\in\Theta} \mathbb{E}\left[\psi\left(Y,\theta\right)\right] - \mathbb{E}\left[\min_{\theta\in\Theta} \psi\left(Y,\theta\right)\right]} \,. \tag{2.20}$$

Moreover, as mentioned in Fort et al. [2016], for specific cases (e.g. constrast functions defined above), the second term in the denominator satisfies $\mathbb{E}\left[\min_{\theta\in\Theta} \psi\left(Y,\theta\right)\right] = 0$, which gives us

$$S_i^{\psi} = \frac{\min_{\theta\in\Theta} \mathbb{E}\left[\psi\left(Y,\theta\right)\right] - \mathbb{E}\left[\min_{\theta\in\Theta} \mathbb{E}\left[\psi\left(Y,\theta\right)|\, X_i\right]\right]}{\min_{\theta\in\Theta} \mathbb{E}\left[\psi\left(Y,\theta\right)\right]} \,. \tag{2.21}$$

Among several properties, one can mention that the index is nonnegative and satisfies $S_i^{\psi} \in [0,1]$, $S_i^{\psi} = 0$ if Y does not depend on X_i as well as $S_i^{\psi} = 1$ if Y is X_i measurable.

It is also possible, given a subset of input parameters $u \subseteq \mathcal{D}$ to define the higher-order contrast index given by

$$S_u^{\psi} = \frac{\min_{\theta\in\Theta} \mathbb{E}\left[\psi\left(Y,\theta\right)\right] - \mathbb{E}\left[\min_{\theta\in\Theta} \mathbb{E}\left[\psi\left(Y,\theta\right)|\, \mathbf{X}_u\right]\right]}{\min_{\theta\in\Theta} \mathbb{E}\left[\psi\left(Y,\theta\right)\right]} \,. \tag{2.22}$$

The latter quantifies the influence of the variation of the parameters indexed by u taken together on the feature $\theta^*\left(Y\right)$.

In what follows, we focus on α-quantile contrast function and give several properties for this specific index.

2.5 QOSA indices

The developments of this section are related to the Quantile Oriented Sensitivity Analysis (QOSA) indices measuring the impact of the inputs over the α-quantile of the output distribution. Given a level of quantile $\alpha \in]0,1[$, let us recall the expression of the first-order QOSA index:

$$S_i^{\alpha} = 1 - \frac{\mathbb{E}\left[\psi_\alpha\left(Y, q^\alpha\left(Y|\, X_i\right)\right)\right]}{\mathbb{E}\left[\psi_\alpha\left(Y, q^\alpha\left(Y\right)\right)\right]} \,, \tag{2.23}$$

where $\psi_\alpha : (y,\theta) \mapsto (y-\theta)\left(\alpha - \mathbb{1}_{\{y\leqslant\theta\}}\right)$ and the q's are the quantiles

$$q^\alpha\left(Y\right) = \arg\min_{\theta\in\mathbb{R}} \mathbb{E}\left[\psi_\alpha\left(Y,\theta\right)\right] \quad \text{and} \quad q^\alpha\left(Y|\, X_i = x_i\right) = \arg\min_{\theta\in\mathbb{R}} \mathbb{E}\left[\psi_\alpha\left(Y,\theta\right)|\, X_i = x_i\right] \,.$$

The first-order QOSA index has been studied in Browne et al. [2017] and Maume-Deschamps and Niang [2018] and may be rewritten as follows

$$S_i^\alpha = \frac{\mathbb{E}\left[Y\mathbb{1}_{\{Y\leqslant q^\alpha(Y|X_i)\}}\right] - \mathbb{E}\left[Y\mathbb{1}_{\{Y\leqslant q^\alpha(Y)\}}\right]}{\alpha\mathbb{E}[Y] - \mathbb{E}\left[Y\mathbb{1}_{\{Y\leqslant q^\alpha(Y)\}}\right]} = 1 - \frac{\alpha\mathbb{E}[Y] - \mathbb{E}\left[Y\mathbb{1}_{\{Y\leqslant q^\alpha(Y|X_i)\}}\right]}{\alpha\mathbb{E}[Y] - \mathbb{E}\left[Y\mathbb{1}_{\{Y\leqslant q^\alpha(Y)\}}\right]} .$$

The following lemma is useful, it is closely related to the proof of sub-additivity of TVaR in risk theory (see Marceau [2013] e.g.).

Lemma 2.5.1.
Consider any event E such that $\mathbb{P}(E) = \alpha$. Then, for any random variable X, we have

$$\mathbb{E}\left[X\mathbb{1}_{\{X\leqslant q^\alpha(X)\}}\right] \leqslant \mathbb{E}\left[X\mathbb{1}_E\right] ,$$

with $q^\alpha(X)$ the α-quantile of X.

Proof of Lemma 2.5.1.

We have:

$$\begin{aligned}\mathbb{E}\left[X\mathbb{1}_{\{X\leqslant q^\alpha(X)\}}\right] - \mathbb{E}\left[X\mathbb{1}_E\right] &= \mathbb{E}\left[X\left(\mathbb{1}_{\{X\leqslant q^\alpha(X)\}} - \mathbb{1}_E\right)\right] \\ &= \mathbb{E}\left[\left(X - q^\alpha(X)\right)\left(\mathbb{1}_{\{X\leqslant q^\alpha(X)\}} - \mathbb{1}_E\right)\right] \leqslant 0 .\end{aligned}$$

■

As a consequence, since $\mathbb{P}\left(Y \leqslant q^\alpha\left(Y|\, X_i\right)|\, X_i\right) = \alpha$ and $\mathbb{P}\left(Y \leqslant q^\alpha\left(Y\right)\right) = \alpha$, we get that

$$\alpha\mathbb{E}[Y] - \mathbb{E}\left[Y\mathbb{1}_{\{Y\leqslant q^\alpha(Y|X_i)\}}\right] = \mathbb{E}\left[\left(Y - q^\alpha\left(Y|\, X_i\right)\right)\left(\alpha - \mathbb{1}_{\{Y\leqslant q^\alpha(Y|X_i)\}}\right)\right] \geqslant 0 \quad (2.24)$$

$$\alpha\mathbb{E}[Y] - \mathbb{E}\left[Y\mathbb{1}_{\{Y\leqslant q^\alpha(Y)\}}\right] = \mathbb{E}\left[\left(Y - q^\alpha\left(Y\right)\right)\left(\alpha - \mathbb{1}_{\{Y\leqslant q^\alpha(Y)\}}\right)\right] \geqslant 0 \quad (2.25)$$

$$\mathbb{E}\left[Y\mathbb{1}_{\{Y\leqslant q^\alpha(Y|X_i)\}}\right] - \mathbb{E}\left[Y\mathbb{1}_{\{Y\leqslant q^\alpha(Y)\}}\right] \geqslant 0 . \quad (2.26)$$

This implies that $0 \leqslant S_i^\alpha \leqslant 1$. The index S_i^α also has the three following interesting properties.

Proposition 2.5.2.

1. *S_i^α is invariant with respect to translations of the output Y.*
2. *S_i^α is invariant by homothety with strictly positive ratio of the output Y.*
3. *A homothety with strictly negative ratio of the output Y gives the index $S_i^{1-\alpha}$ associated to the $1-\alpha$ level.*

Proof of Proposition 2.5.2.

Let us consider any model $Y = \eta(\mathbf{X})$. We will denote by $S_i'^{\alpha}$ the QOSA indices related to a random variable Y'.

1. Let $Y' = Y + k,\ k \in \mathbb{R}$. Then, we have $q^{\alpha}(Y') = q^{\alpha}(Y) + k$ and $q^{\alpha}(Y'|X_i) = q^{\alpha}(Y|X_i) + k$. It is easy to check that $S_i'^{\alpha} = S_i^{\alpha}$.

2. Let $Y' = k \times Y,\ k > 0$. Then we have, $q^{\alpha}(Y') = k \times q^{\alpha}(Y)$ and $q^{\alpha}(Y'|X_i) = k \times q^{\alpha}(Y|X_i)$. We can easily show that $S_i'^{\alpha} = S_i^{\alpha}$.

3. Let $Y' = k \times Y,\ k < 0$. Then we have, $q^{\alpha}(Y') = k \times q^{1-\alpha}(Y)$ and $q^{\alpha}(Y'|X_i) = k \times q^{1-\alpha}(Y|X_i)$. Finally, it can simply be proved that $S_i'^{\alpha} = S_i^{1-\alpha}$.

■

Now, we are going to investigate the sum $\mathcal{S}$ of the first-order QOSA indices:

$$\mathcal{S} = \sum_{i=1}^{d} S_i^{\alpha} = \frac{\sum_{i=1}^{d} \mathbb{E}\left[Y \mathbb{1}_{\{Y \leqslant q^{\alpha}(Y|X_i)\}}\right] - d\mathbb{E}\left[Y \mathbb{1}_{\{Y \leqslant q^{\alpha}(Y)\}}\right]}{\alpha \mathbb{E}[Y] - \mathbb{E}\left[Y \mathbb{1}_{\{Y \leqslant q^{\alpha}(Y)\}}\right]} .$$

We see that $\mathcal{S} \leqslant 1$ if and only if

$$\sum_{i=1}^{d} \mathbb{E}\left[Y \mathbb{1}_{\{Y \leqslant q^{\alpha}(Y|X_i)\}}\right] - d\mathbb{E}\left[Y \mathbb{1}_{\{Y \leqslant q^{\alpha}(Y)\}}\right] \leqslant \left(\alpha \mathbb{E}[Y] - \mathbb{E}\left[Y \mathbb{1}_{\{Y \leqslant q^{\alpha}(Y)\}}\right]\right) . \quad (2.27)$$

Or equivalently:

$$g(\alpha) = \alpha \mathbb{E}[Y] + (d-1)\mathbb{E}\left[Y \mathbb{1}_{\{Y \leqslant q^{\alpha}(Y)\}}\right] - \sum_{i=1}^{d} \mathbb{E}\left[Y \mathbb{1}_{\{Y \leqslant q^{\alpha}(Y|X_i)\}}\right] \geqslant 0 .$$

As proved in the following proposition, $\mathcal{S}$ is smaller than 1 in the case of an additive model [Stone, 1985; Hastie and Tibshirani, 1990] with $\mathbf{X}$ that has independent marginal laws. Unfortunately, this result is not true in the general case as showed with a counterexample in Subsection 2.5.1.1.

Proposition 2.5.3.

Let $\mathbf{X} = (X_1, \ldots, X_d)$ with the X_i's independent. Let $Y = m_0 + \sum_{i=1}^{d} m_i(X_i)$ be an additive model with m_i, $i = 1, \ldots, d$, the one-dimensional nonparametric functions operating on each element of the vector $\mathbf{X}$. Then, the sum of the first-order QOSA indices $\mathcal{S}$ satisfies $\mathcal{S} \leqslant 1$.

Proof of Proposition 2.5.3.

Given a random variable X, we denote by $q^\alpha(X)$ its α-quantile. For any $i = 1, \dots, d$, let $Xs_{(-i)} = \sum_{\substack{1 \leqslant j \leqslant d \\ j \neq i}} m_j(X_j)$, we have thanks to the independence

$$q^\alpha\left(Y \middle| X_i\right) = m_0 + m_i(X_i) + q^\alpha\left(Xs_{(-i)}\right), \text{ and } \left\{Y \leqslant q^\alpha\left(Y \middle| X_i\right)\right\} = \left\{Xs_{(-i)} \leqslant q^\alpha\left(Xs_{(-i)}\right)\right\}.$$

Consider $g(\alpha)$ as above,

$$\begin{aligned} g(\alpha) &= \alpha\mathbb{E}[Y] + (d-1)\mathbb{E}\left[Y\mathbb{1}_{\{Y \leqslant q^\alpha(Y)\}}\right] - \sum_{i=1}^{d} \mathbb{E}\left[Y\mathbb{1}_{\{Y \leqslant q^\alpha(Y|X_i)\}}\right] \\ &= \alpha\mathbb{E}[Y] + (d-1)\mathbb{E}\left[Y\mathbb{1}_{\{Y \leqslant q^\alpha(Y)\}}\right] - \sum_{i=1}^{d} \left(\alpha \times m_0 + \mathbb{E}\left[m_i(X_i)\mathbb{1}_{\left\{Xs_{-(i)} \leqslant q^\alpha\left(Xs_{(-i)}\right)\right\}}\right]\right. \\ &\quad \left. + \mathbb{E}\left[Xs_{(-i)}\mathbb{1}_{\left\{Xs_{-(i)} \leqslant q^\alpha\left(Xs_{(-i)}\right)\right\}}\right]\right). \end{aligned}$$

Now, the independence of the X_i's implies that $\mathbb{E}\left[m_i(X_i)\mathbb{1}_{\left\{Xs_{(-i)} \leqslant q^\alpha\left(Xs_{(-i)}\right)\right\}}\right] = \alpha\mathbb{E}[m_i(X_i)]$ and thus,

$$g(\alpha) = (d-1)\mathbb{E}\left[\left(\sum_{j=1}^{d} m_j(X_j)\right)\mathbb{1}_{\{Y \leqslant q^\alpha(Y)\}}\right] - \sum_{i=1}^{d} \mathbb{E}\left[Xs_{(-i)}\mathbb{1}_{\left\{Xs_{(-i)} \leqslant q^\alpha\left(Xs_{(-i)}\right)\right\}}\right].$$

Now, we use Lemma 2.5.1 which gives:

$$\begin{aligned} (d-1)\mathbb{E}\left[\left(\sum_{j=1}^{d} m_j(X_j)\right)\mathbb{1}_{\{Y \leqslant q^\alpha(Y)\}}\right] &= \sum_{i=1}^{d-1} \left(\mathbb{E}\left[m_i(X_i)\mathbb{1}_{\{Y \leqslant q^\alpha(Y)\}}\right] + \mathbb{E}\left[Xs_{(-i)}\mathbb{1}_{\{Y \leqslant q^\alpha(Y)\}}\right]\right) \\ &\geqslant \sum_{i=1}^{d-1} \left(\mathbb{E}\left[m_i(X_i)\mathbb{1}_{\{Y \leqslant q^\alpha(Y)\}}\right]\right. \\ &\quad \left. + \mathbb{E}\left[Xs_{(-i)}\mathbb{1}_{\left\{Xs_{(-i)} \leqslant q^\alpha\left(Xs_{(-i)}\right)\right\}}\right]\right). \end{aligned}$$

As a consequence,

$$\begin{aligned} g(\alpha) &\geqslant \sum_{i=1}^{d-1} \mathbb{E}\left[m_i(X_i)\mathbb{1}_{\{Y \leqslant q^\alpha(Y)\}}\right] - \mathbb{E}\left[Xs_{(-d)}\mathbb{1}_{\left\{Xs_{(-d)} \leqslant q^\alpha\left(Xs_{(-d)}\right)\right\}}\right] \\ &= \mathbb{E}\left[Xs_{(-d)}\mathbb{1}_{\{Y \leqslant q^\alpha(Y)\}}\right] - \mathbb{E}\left[Xs_{(-d)}\mathbb{1}_{\left\{Xs_{(-d)} \leqslant q^\alpha\left(Xs_{(-d)}\right)\right\}}\right] \\ &\geqslant 0 \text{ using again Lemma 2.5.1.} \end{aligned}$$

■

The first-order QOSA index presents some limitations because it only captures the main effect of the i-th input. Kala [2019] introduced the second-order QOSA index in order to assess the impact of the interaction effect of two inputs on the α-quantile. He uses the Equation (2.22) to measure the joint effect of the pair (X_i, X_j) on the α-quantile minus the first-order effects for the same factors in order to keep only the interaction effect and thereby defines

$$S_{ij}^{\alpha} = \frac{\min_{\theta\in\mathbb{R}} \mathbb{E}\left[\psi_{\alpha}\left(Y,\theta\right)\right] - \mathbb{E}\left[\min_{\theta\in\mathbb{R}} \mathbb{E}\left[\psi_{\alpha}\left(Y,\theta\right)\middle|\, X_i, X_j\right]\right]}{\min_{\theta\in\mathbb{R}} \mathbb{E}\left[\psi_{\alpha}\left(Y,\theta\right)\right]} - S_i^{\alpha} - S_j^{\alpha} \ .$$

The higher-order QOSA indices can be expressed analogously. Hence, one obtain a variance-like decomposition for quantiles in the case of independent inputs:

$$\sum_{i=1}^{d} S_i^{\alpha} + \sum_{1\leqslant i<j\leqslant d} S_{ij}^{\alpha} + \cdots + S_{1,\ldots,d}^{\alpha} = 1 \ . \tag{2.28}$$

Although this decomposition appears similar to that of Sobol indices described in (2.5), there is a key point to underline. Sobol indices stem from the univocal definition of the functional ANOVA decomposition whereas the decomposition above related to the QOSA indices is obtained by construction.

The indices $S_i^{\alpha}, S_{ij}^{\alpha}$ and higher order allow to assess thoroughly the impact of each input over the α-quantile of the output distribution. However, as for the Sobol indices, in the case of a large number of inputs, it would require the evaluation of $2^d - 1$ indices, which could be computationally demanding. Therefore, it is suitable to introduce the so-called total QOSA index as suggested by Kala [2019] that measures the contribution of an input, including its main effect as well as its interactions effects, of any order, with other input variables:

$$ST_i^{\alpha} = \frac{\mathbb{E}\left[\min_{\theta\in\mathbb{R}} \mathbb{E}\left[\psi_{\alpha}\left(Y,\theta\right)\middle|\, \mathbf{X}_{-i}\right]\right]}{\min_{\theta\in\mathbb{R}} \mathbb{E}\left[\psi_{\alpha}\left(Y,\theta\right)\right]} = \frac{\mathbb{E}\left[\psi_{\alpha}\left(Y, q^{\alpha}\left(Y\middle|\, \mathbf{X}_{-i}\right)\right)\right]}{\mathbb{E}\left[\psi_{\alpha}\left(Y, q^{\alpha}\left(Y\right)\right)\right]} \ .$$

The total QOSA index may be rewritten as follows

$$ST_i^{\alpha} = \frac{\alpha\mathbb{E}\left[Y\right] - \mathbb{E}\left[Y \mathbb{1}_{\{Y\leqslant q^{\alpha}(Y|\mathbf{X}_{-i})\}}\right]}{\alpha\mathbb{E}\left[Y\right] - \mathbb{E}\left[Y \mathbb{1}_{\{Y\leqslant q^{\alpha}(Y)\}}\right]} \ .$$

As for the first-order QOSA index, the total one has the three following interesting properties.

Proposition 2.5.4.

1. *ST_i^{α} is invariant with respect to translations of the output Y.*
2. *ST_i^{α} is invariant by homothety with strictly positive ratio of the output Y.*

3. *A homothety with strictly negative ratio of the output Y gives the index $ST_i^{1-\alpha}$ associated to the $1-\alpha$ level.*

Proof of Proposition 2.5.4.

Just adapting the steps of the proof of the Proposition 2.5.2 for the total QOSA index. ■

Proposition 2.5.5 below shows that the total QOSA index is greater than or equal to the first-order one for any α-level in the case of an additive model with $\mathbf{X}$ that has independent marginals. This is a major difference from variance-based methods, specifically Sobol indices. Indeed, it is well-known that for a purely additive model with independent inputs, we have for the Sobol indices $ST_i = S_i,\ \forall i \in \mathcal{D}$, which is not the case for the QOSA indices. It therefore appears that the total QOSA index captures some interaction between the inputs when using an additive model. The origin of this phenomenon is not yet understood at this stage and requires further analysis.
Besides, it should be noted that Proposition 2.5.5 is not verified in the general case as emphasized with a counterexample in Subsection 2.5.1.3.

Proposition 2.5.5.

Let $\mathbf{X} = (X_1, \ldots, X_d)$ with the X_i's independent. Let $Y = m_0 + \sum_{i=1}^{d} m_i(X_i)$ be an additive model with m_i, $i = 1, \ldots, d$, the one-dimensional nonparametric functions operating on each element of the vector $\mathbf{X}$. Then,

$$\forall\ \alpha \in]0,1[\,, \quad S_i^\alpha \leqslant ST_i^\alpha\ .$$

Proof of Proposition 2.5.5.

We have:

$$ST_i^\alpha - S_i^\alpha = \frac{\left(\alpha\mathbb{E}[Y] - \mathbb{E}\left[Y\mathbb{1}_{\{Y\leqslant q^\alpha(Y|\mathbf{X}_{-i})\}}\right]\right) - \left(\mathbb{E}\left[Y\mathbb{1}_{\{Y\leqslant q^\alpha(Y|X_i)\}}\right] - \mathbb{E}\left[Y\mathbb{1}_{\{Y\leqslant q^\alpha(Y)\}}\right]\right)}{\alpha\mathbb{E}[Y] - \mathbb{E}\left[Y\mathbb{1}_{\{Y\leqslant q^\alpha(Y)\}}\right]}\ .$$

As the denominator is positive according to Equation (2.25), we just have to show that the numerator is.

We denote the numerator by

$$g(\alpha) = \alpha\mathbb{E}[Y] - \mathbb{E}\left[Y\mathbb{1}_{\{Y\leqslant q^\alpha(Y|\mathbf{X}_{-i})\}}\right] + \mathbb{E}\left[Y\mathbb{1}_{\{Y\leqslant q^\alpha(Y)\}}\right] - \mathbb{E}\left[Y\mathbb{1}_{\{Y\leqslant q^\alpha(Y|X_i)\}}\right]\ ,$$

as well as $q^\alpha(X)$, the α-quantile of a random variable X. For any $i = 1, \ldots, d$, let $Xs_{(-i)} = \sum_{\substack{1\leqslant j\leqslant d \\ j\neq i}} m_j(X_j)$, we have thanks to the independence

$$q^\alpha(Y|X_i) = m_0 + m_i(X_i) + q^\alpha\left(Xs_{(-i)}\right),\ \text{and}\ \{Y \leqslant q^\alpha(Y|X_i)\} = \{Xs_{(-i)} \leqslant q^\alpha\left(Xs_{(-i)}\right)\}\ ,$$
$$q^\alpha(Y|\mathbf{X}_{-i}) = m_0 + Xs_{(-i)} + q^\alpha(m_i(X_i)),\ \text{and}\ \{Y \leqslant q^\alpha(Y|\mathbf{X}_{-i})\} = \{m_i(X_i) \leqslant q^\alpha(m_i(X_i))\}\ .$$

Then, we have

$$
\begin{aligned}
g(\alpha) &= \alpha\mathbb{E}[Y] - \mathbb{E}\left[Y\mathbb{1}_{\{Y\leqslant q^\alpha(Y|\mathbf{X}_{-i})\}}\right] + \mathbb{E}\left[Y\mathbb{1}_{\{Y\leqslant q^\alpha(Y)\}}\right] - \mathbb{E}\left[Y\mathbb{1}_{\{Y\leqslant q^\alpha(Y|X_i)\}}\right] \\
&= \alpha\mathbb{E}[Y] - \alpha m_0 - \mathbb{E}\left[m_i(X_i)\,\mathbb{1}_{\{m_i(X_i)\leqslant q^\alpha(m_i(X_i))\}}\right] - \mathbb{E}\left[Xs_{(-i)}\mathbb{1}_{\{m_i(X_i)\leqslant q^\alpha(m_i(X_i))\}}\right] \\
&\quad + \mathbb{E}\left[Y\mathbb{1}_{\{Y\leqslant q^\alpha(Y)\}}\right] - \alpha m_0 - \mathbb{E}\left[Xs_{(-i)}\mathbb{1}_{\left\{Xs_{(-i)}\leqslant q^\alpha\left(Xs_{(-i)}\right)\right\}}\right] \\
&\quad - \mathbb{E}\left[m_i(X_i)\,\mathbb{1}_{\left\{Xs_{(-i)}\leqslant q^\alpha\left(Xs_{(-i)}\right)\right\}}\right] .
\end{aligned}
$$

Now, the independence of the X_i's implies that $\mathbb{E}\left[Xs_{(-i)}\mathbb{1}_{\{m_i(X_i)\leqslant q^\alpha(m_i(X_i))\}}\right] = \alpha\mathbb{E}\left[Xs_{(-i)}\right]$ and $\mathbb{E}\left[m_i(X_i)\mathbb{1}_{\left\{Xs_{(-i)}\leqslant q^\alpha\left(Xs_{(-i)}\right)\right\}}\right] = \alpha\mathbb{E}[m_i(X_i)]$. Thus,

$$
\begin{aligned}
g(\alpha) &= \mathbb{E}\left[\left(\sum_{j=1}^{d} m_j(X_j)\right)\mathbb{1}_{\{Y\leqslant q^\alpha(Y)\}}\right] - \mathbb{E}\left[m_i(X_i)\,\mathbb{1}_{\{m_i(X_i)\leqslant q^\alpha(m_i(X_i))\}}\right] \\
&\quad - \mathbb{E}\left[Xs_{(-i)}\mathbb{1}_{\left\{Xs_{(-i)}\leqslant q^\alpha\left(Xs_{(-i)}\right)\right\}}\right] \\
&= \left(\mathbb{E}\left[m_i(X_i)\,\mathbb{1}_{\{Y\leqslant q^\alpha(Y)\}}\right] - \mathbb{E}\left[m_i(X_i)\,\mathbb{1}_{\{m_i(X_i)\leqslant q^\alpha(m_i(X_i))\}}\right]\right) \\
&\quad + \left(\mathbb{E}\left[Xs_{(-i)}\mathbb{1}_{\{Y\leqslant q^\alpha(Y)\}}\right] - \mathbb{E}\left[Xs_{(-i)}\mathbb{1}_{\left\{Xs_{(-i)}\leqslant q^\alpha\left(Xs_{(-i)}\right)\right\}}\right]\right) .
\end{aligned}
$$

Now, the two last terms are positive according to Lemma 2.5.1 which concludes the proof. ■

At last, let us mention that Kucherenko et al. [2019] have recently proposed new indices to assess the impact of inputs over the α-quantile of the output distribution. Instead of considering the expression of the first-order Sobol index based on a contrast function as in (2.18), they use the native form whose the numerator is $\mathrm{Var}\left(\mathbb{E}\left[Y\,|\,X_i\right]\right) = \mathbb{E}\left[\left(\mathbb{E}\left[Y\,|\,X_i\right] - \mathbb{E}[Y]\right)^2\right]$ and simply replace the expectations by α-quantiles to define the following indices

$$
\bar{q}^\alpha_{i,1} = \mathbb{E}\left[\left|q^\alpha(Y) - q^\alpha\left(Y\,|\,X_i\right)\right|\right] \quad \text{and} \quad \bar{q}^\alpha_{i,2} = \mathbb{E}\left[\left(q^\alpha(Y) - q^\alpha\left(Y\,|\,X_i\right)\right)^2\right] .
$$

They also provide the normalized versions as follows

$$
Q^\alpha_{i,1} = \frac{\bar{q}^\alpha_{i,1}}{\sum_{j=1}^{d} \bar{q}^\alpha_{j,1}} \quad \text{and} \quad Q^\alpha_{i,2} = \frac{\bar{q}^\alpha_{i,2}}{\sum_{j=1}^{d} \bar{q}^\alpha_{j,2}} .
$$

These measures thereby quantify the mean distance between quantiles $q^\alpha(Y)$ and $q^\alpha\left(Y\,|\,X_i\right)$ rather than the mean distance between average contrast functions like in the first-order QOSA index given in (2.23). They are still under investigation in order to understand their interpretation. We should pay attention that using these indices only allows to quantify the main effects and not the interaction ones that can nevertheless have an significant impact.

2.5.1 Special cases

In this subsection, we compute for some distributions the first-order and total QOSA indices for which we obtain the expressions in a closed or nearly closed-form. In particular, the examples with Gaussian inputs allow to investigate the behavior of the indices when there is some statistical dependence between the inputs.

2.5.1.1 Product of two exponential input variables

Let $Y = X_1 \cdot X_2$, with $X_1 \sim \mathcal{E}(\lambda)$, $X_2 \sim \mathcal{E}(\delta)$, these two variables being independent. After simple calculations, we get the first-order QOSA indices

$$S_1^\alpha = S_2^\alpha = 1 - \frac{(\alpha - 1)\log(1-\alpha)}{\alpha - \lambda\delta \cdot \mathbb{E}\left[Y \mathbb{1}_{\{Y \leqslant q^\alpha(Y)\}}\right]} , \tag{2.29}$$

and the total QOSA indices

$$ST_1^\alpha = ST_2^\alpha = \frac{(\alpha - 1)\log(1-\alpha)}{\alpha - \lambda\delta \cdot \mathbb{E}\left[Y \mathbb{1}_{\{Y \leqslant q^\alpha(Y)\}}\right]} . \tag{2.30}$$

The term $\mathbb{E}\left[Y \mathbb{1}_{\{Y \leqslant q^\alpha(Y)\}}\right]$ can be approximated by using a Monte-Carlo estimation or a numerical integration.

The equality of the first-order and total QOSA indices for both inputs is a particular case due to the exponential distribution. Indeed, let $X_2 \overset{\mathcal{L}}{=} \frac{\lambda}{\delta} X_1'$ with X_1' an independent copy of X_1. Then, the model writes

$$\begin{aligned} Y &\overset{\mathcal{L}}{=} \frac{\lambda}{\delta} X_1 \cdot X_1' \\ &\overset{\mathcal{L}}{=} kZ \end{aligned}$$

with $Z = X_1 \cdot X_1'$ and $k = \frac{\lambda}{\delta} > 0$. As the inputs X_1 and X_1' have the same distribution, their impact over the α-quantile of the model output Z is identical. Therefore, by using item 2. of Propositions 2.5.2 and 2.5.4 (because Y is just a homothety with strictly positive ratio of Z), that explains why both first-order and total QOSA indices of the inputs are equal.

Figure 2.1 below presents the behavior of the indices as a function of the level α for the model computed with $\lambda = 1/10$ and $\delta = 1$. The truncated expectation is estimated with a Monte-Carlo algorithm and a sample of size $n = 10^9$. We observe that the first-order and total QOSA indices vary in opposite directions. The first-order QOSA indices go to 1 when α tends to 1 while the total ones go to 0 when α tends to 1. It is interesting to notice that from $\alpha \approx 0.96$ the total QOSA indices are lower than the first-order ones and the sum of the first-order ones is greater than 1. That corroborates that Propositions 2.5.2 and 2.5.4 are not verified outside the additive model context.

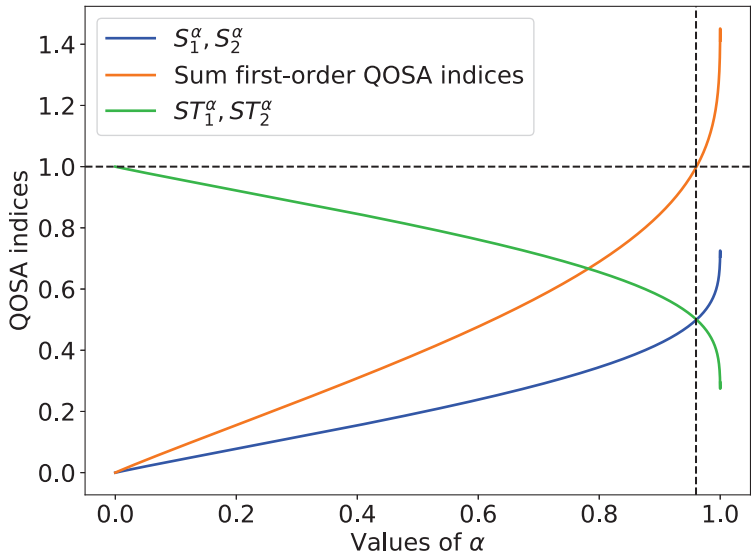


Figure 2.1 *Evolution of the first-order and total QOSA indices at different levels α for the product of two exponentials with $\lambda = 1/10$ for the fisrt input and $\delta = 1$ for the second one.*

2.5.1.2 Linear model with Gaussian input variables

We study in this subsection a linear model with Gaussian inputs which implies that the resulting output is also Gaussian. This framework facilitates calculations to obtain the analytical values given below.

Proposition 2.5.6.
If $Y = \eta(\mathbf{X}) = \beta_0 + \boldsymbol{\beta}^\mathsf{T}\mathbf{X}$ with $\beta_0 \in \mathbb{R}$, $\boldsymbol{\beta} \in \mathbb{R}^d$ and $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$ is a positive-definite matrix, then the first-order and total-order QOSA indices for the variable i at the α-level are

$$S_i^\alpha = 1 - \frac{\sqrt{\boldsymbol{\beta}_{-i}^\mathsf{T}\left(\boldsymbol{\Sigma}_{-i,-i} - \boldsymbol{\Sigma}_{-i,i}\boldsymbol{\Sigma}_{i,i}^{-1}\boldsymbol{\Sigma}_{i,-i}\right)\boldsymbol{\beta}_{-i}}}{\sigma_Y}, \tag{2.31}$$

$$ST_i^\alpha = \frac{|\boldsymbol{\beta}_i|\sqrt{\boldsymbol{\Sigma}_{i,i} - \boldsymbol{\Sigma}_{i,-i}\boldsymbol{\Sigma}_{-i,-i}^{-1}\boldsymbol{\Sigma}_{-i,i}}}{\sigma_Y}, \tag{2.32}$$

with $\sigma_Y^2 = \mathrm{Var}(Y) = \boldsymbol{\beta}^\mathsf{T}\boldsymbol{\Sigma}\boldsymbol{\beta}$.

We observe that as β_0 and $\boldsymbol{\mu}$ are translation parameters, they do not have any influence. Nevertheless, no general conclusion can be drawn from Equations (2.31) and (2.32) except that the values of the first-order and total QOSA indices are the same for all levels α. This phenomenon is specific to the Gaussian linear model and will be detailed hereafter on the next particular case.

Let us consider the case $d = 2$ with

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \ \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \text{ and } \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}, \ -1 \leq \rho \leq 1, \ \sigma_1 > 0, \ \sigma_2 > 0.$$

We have $\sigma_Y^2 = Var\,(Y) = \beta_1^2\sigma_1^2 + 2\rho\beta_1\beta_2\sigma_1\sigma_2 + \beta_2^2\sigma_2^2$ and obtain from Equations (2.31) and (2.32)

$$\begin{aligned} S_1^\alpha &= 1 - \frac{|\beta_2|\,\sigma_2\sqrt{1-\rho^2}}{\sigma_Y}\ , \\ S_2^\alpha &= 1 - \frac{|\beta_1|\,\sigma_1\sqrt{1-\rho^2}}{\sigma_Y}\ , \end{aligned} \tag{2.33}$$

and

$$\begin{aligned} ST_1^\alpha &= \frac{|\beta_1|\,\sigma_1\sqrt{1-\rho^2}}{\sigma_Y}\ , \\ ST_2^\alpha &= \frac{|\beta_2|\,\sigma_2\sqrt{1-\rho^2}}{\sigma_Y}\ . \end{aligned} \tag{2.34}$$

Why the index does not depend on the α-level ?

First of all, let $U \sim \mathcal{N}\,(\mu_U, \sigma_U^2)$ and $V \sim \mathcal{N}\,(\mu_V, \sigma_V^2)$ be two generic random variables. Then, their respective quantiles at the α-level are given by

$$\begin{aligned} q^\alpha\,(U) &= \mu_U + \sigma_U\Phi^{-1}\,(\alpha)\ , \\ q^\alpha\,(V) &= \mu_V + \sigma_V\Phi^{-1}\,(\alpha)\ , \end{aligned}$$

with Φ^{-1} the quantile function of the standard normal distribution $\mathcal{N}(0,1)$. The means μ_U and μ_V being just scale parameters, we could have considered centred laws, but for the sake of generality, we keep them. We then have the following link between the quantiles of U and V

$$\forall\ \alpha, \quad \frac{q^\alpha\,(U) - \mu_U}{q^\alpha\,(V) - \mu_V} = \frac{\sigma_U}{\sigma_V}\ . \tag{2.35}$$

Thus, whatever the α-level, the link between the quantiles of two Gaussian random variables is only determined by a variance ratio. This is the key point explaining why the indices do not depend on the α-level.

Let's go back to the case of $d = 2$ and try to understand why the index does not depend on α-level by looking only at the impact of the variable X_1.
We have that

$$Y|X_1 \sim \mathcal{N}\left(\beta_1 X_1 + \beta_2\mathbb{E}\,[\,X_2|\,X_1], \ \beta_2^2\mathrm{Var}\,(\,X_2|\,X_1)\right)\ .$$

As we work with Gaussian distributions, the conditional variance $\mathrm{Var}\,(\,X_2|\,X_1)$ does not depend on the specific value of X_1 and is $\mathrm{Var}\,(\,X_2|\,X_1) = \sigma_2^2\,(1-\rho^2)$. The conditional quantile of Y given X_1 has the following expression

$$q^\alpha\,(\,Y|\,X_1) = \beta_1 X_1 + \beta_2\mathbb{E}\,[\,X_2|\,X_1] + |\beta_2|\,\sqrt{\mathrm{Var}\,(\,X_2|\,X_1)}\Phi^{-1}\,(\alpha)\ .$$

One way to assess the impact of the variable X_1 on the quantile $q^\alpha(Y)$ would be to calculate the ratio

$$\frac{q^\alpha(Y|X_1) - (\beta_1 X_1 + \beta_2 \mathbb{E}[X_2|X_1])}{q^\alpha(Y) - \mathbb{E}[Y]},$$

for several values of x_1 and to observe its evolution. If, when one sets X_1 to several different values, the ratio varies a lot, then X_1 is highly responsible for the value $q^\alpha(Y)$. The issue with this measure is the choice of the value x_1 of X_1, which can be solved by considering the expectation of this quantity

$$\frac{\mathbb{E}[q^\alpha(Y|X_1)] - \mathbb{E}[Y]}{q^\alpha(Y) - \mathbb{E}[Y]}.$$

Thus, by using that $q^\alpha(Y) = \mathbb{E}[Y] + \sigma_Y \Phi^{-1}(\alpha)$, the previous ratio equals

$$\frac{\mathbb{E}[q^\alpha(Y|X_1)] - \mathbb{E}[Y]}{q^\alpha(Y) - \mathbb{E}[Y]} = \frac{|\beta_2|\sigma_2\sqrt{(1-\rho^2)}}{\sigma_Y}.$$

As Y is a Gaussian distribution as well as the conditional distribution of Y given X_1, then the influence of the input X_1 over the quantile $q^\alpha(Y)$ is only driven by the variances as showed in (2.35). We therefore recognize the term involved in the first-order QOSA index S_1^α in (2.33) as well as in the total QOSA index ST_2^α in (2.34).

In a more general way, the following equality holds for all variables $i = 1, \dots, d$ when using a **linear Gaussian model** and explains why the first-order and total QOSA indices do not depend on the α-level:

$$\frac{\alpha\mathbb{E}[Y] - \mathbb{E}\left[Y\mathbb{1}_{\{Y \leqslant q^\alpha(Y|X_i)\}}\right]}{\alpha\mathbb{E}[Y] - \mathbb{E}\left[Y\mathbb{1}_{\{Y \leqslant q^\alpha(Y)\}}\right]} = \frac{\mathbb{E}[q^\alpha(Y|X_i)] - \mathbb{E}[Y]}{q^\alpha(Y) - \mathbb{E}[Y]}.$$

We now study the particular case $\mu_1 = \mu_2 = 0, \beta_1 = \beta_2 = 1, \sigma_1 = 1$ and $\sigma_2 = 2$. The analytical values of the indices are depicted in Figure 2.2 on the left-hand graph for independent inputs and on the right-hand plot as a function of the correlation coefficient between the two inputs in order to investigate the influence of the dependence.

For the independent case, it appears that the variable X_2 has the higher impact over the α-quantile, which is consistent with the setting established. Besides, we have $S_i^\alpha \leqslant ST_i^\alpha,\ i = 1, 2$ as proved in Proposition 2.5.5.
Regarding the dependent case, we observe that the total QOSA indices tend to zero as $|\rho| \to 1$. It is also interesting to notice that we have $ST_i^\alpha \leqslant S_i^\alpha$ for some correlation coefficients. The behaviour of these indices is similar to that of the Sobol indices in the context of dependent inputs as studied in Kucherenko et al. [2012]; Iooss and Prieur [2019]. Indeed, by making an analogy with the method proposed by Mara et al. [2015] based on four Sobol indices, we could say that in the case of dependent inputs:

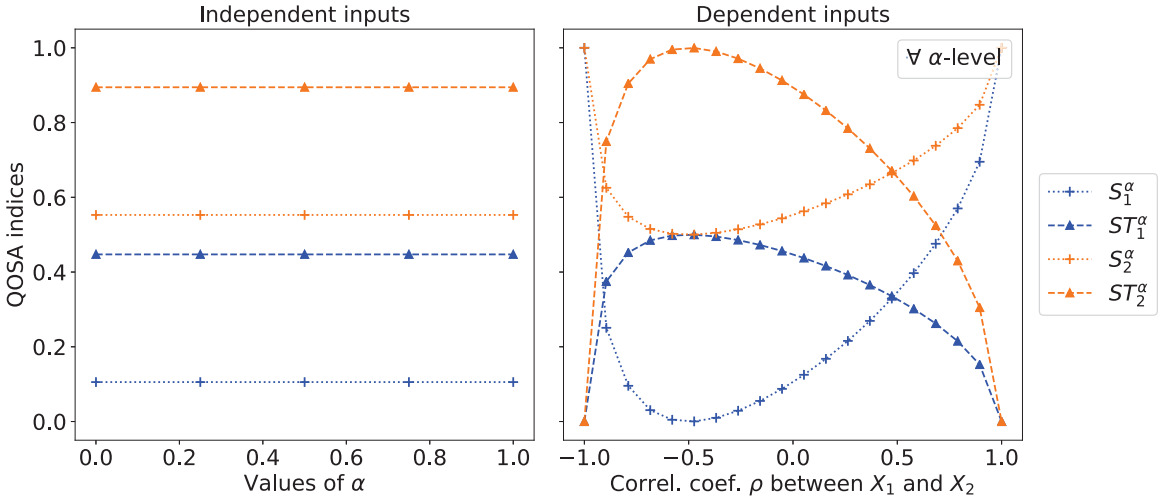


Figure 2.2 *First-order and total QOSA indices with independent (resp. dependent) inputs on the left (resp. right) graph.*

- the first-order QOSA index describes the influence of a variable including its dependence with other variables,
- the total QOSA index describes the influence of a variable without its dependence with other variables.

2.5.1.3 Gaussian input variables, exponential η

We analyze in this subsection a model with still Gaussian inputs but whose the resulting output is a Log-normal distribution so that we no longer have identical indices for any α-level. Using Gaussian inputs make calculations possible and we obtain the following analytical values.

Proposition 2.5.7.
If $Y = \eta(\mathbf{X}) = \exp\left(\beta_0 + \boldsymbol{\beta}^\mathsf{T}\mathbf{X}\right)$ *with* $\beta_0 \in \mathbb{R}$, $\boldsymbol{\beta} \in \mathbb{R}^d$ *and* $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *where* $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$ *is a positive-definite matrix, then the first-order and total-order QOSA indices for the variable* i *at the* α*-level are*

$$S_i^\alpha = 1 - \frac{\alpha - \Phi\left(\Phi^{-1}(\alpha) - \sqrt{\boldsymbol{\beta}_{-i}^\mathsf{T}\left(\boldsymbol{\Sigma}_{-i,-i} - \boldsymbol{\Sigma}_{-i,i}\boldsymbol{\Sigma}_{i,i}^{-1}\boldsymbol{\Sigma}_{i,-i}\right)\boldsymbol{\beta}_{-i}}\right)}{\alpha - \Phi\left(\Phi^{-1}(\alpha) - \sigma\right)}, \tag{2.36}$$

$$ST_i^\alpha = \frac{\alpha - \Phi\left(\Phi^{-1}(\alpha) - |\boldsymbol{\beta}_i|\sqrt{\boldsymbol{\Sigma}_{i,i} - \boldsymbol{\Sigma}_{i,-i}\boldsymbol{\Sigma}_{-i,-i}^{-1}\boldsymbol{\Sigma}_{-i,i}}\right)}{\alpha - \Phi\left(\Phi^{-1}(\alpha) - \sigma\right)}, \tag{2.37}$$

with $\sigma^2 = \boldsymbol{\beta}^\mathsf{T}\boldsymbol{\Sigma}\boldsymbol{\beta}$, Φ *the distribution function of the standard normal distribution* $\mathcal{N}(0,1)$ *and* Φ^{-1} *the quantile function.*

We observe that β_0 and $\boldsymbol{\mu}$ do not play any role as these are scale parameters in this example. While the indices vary as a function of α in this scheme, no conclusion can be reached from Equations (2.36) and (2.37). As a consequence, let us consider the particular case $d = 2$ with

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \ \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \text{ and } \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}, \ -1 \leq \rho \leq 1, \ \sigma_1 > 0, \ \sigma_2 > 0.$$

We have $\sigma^2 = \beta_1^2\sigma_1^2 + 2\rho\beta_1\beta_2\sigma_1\sigma_2 + \beta_2^2\sigma_2^2$ and obtain from Equations (2.36) and (2.37)

$$\begin{aligned} S_1^\alpha &= 1 - \frac{\alpha - \Phi\left(\Phi^{-1}(\alpha) - |\beta_2|\,\sigma_2\sqrt{1-\rho^2}\right)}{\alpha - \Phi\left(\Phi^{-1}(\alpha) - \sigma\right)}, \\ S_2^\alpha &= 1 - \frac{\alpha - \Phi\left(\Phi^{-1}(\alpha) - |\beta_1|\,\sigma_1\sqrt{1-\rho^2}\right)}{\alpha - \Phi\left(\Phi^{-1}(\alpha) - \sigma\right)}, \end{aligned} \tag{2.38}$$

and

$$\begin{aligned} ST_1^\alpha &= \frac{\alpha - \Phi\left(\Phi^{-1}(\alpha) - |\beta_1|\,\sigma_1\sqrt{1-\rho^2}\right)}{\alpha - \Phi\left(\Phi^{-1}(\alpha) - \sigma\right)}, \\ ST_2^\alpha &= \frac{\alpha - \Phi\left(\Phi^{-1}(\alpha) - |\beta_2|\,\sigma_2\sqrt{1-\rho^2}\right)}{\alpha - \Phi\left(\Phi^{-1}(\alpha) - \sigma\right)}. \end{aligned} \tag{2.39}$$

In all further tests, we take $\mu_1 = \mu_2 = 0, \beta_1 = \beta_2 = 1, \sigma_1 = 1$ and $\sigma_2 = 2$.
Figure 2.3 presents the analytical values of the first-order and total QOSA indices for both independent inputs and correlated inputs with $\rho_{1,2} = 0.75$. In the independent setting, the influence of the variable X_1 is close to 0 except for large values of α. We also note that the first-order and total QOSA indices vary in reverse direction and from some α-level, $ST_i^\alpha \leqslant S_i^\alpha$, $i = 1, 2$. This supports that Proposition 2.5.5 is not true outside the additive framework with independent inputs.
The behavior of the indices is similar in the dependent case. However, the influence of the input X_1 is reinforced in this scheme due to its large correlation with X_2 that is an influent variable. Indeed, the index S_1^α increases faster than in independent case. On the contrary, the index ST_2^α decreases to 0 quicker than in the independent case because of its high dependence with X_1.

To get another perspective on the impact of the dependence over the indices, we plot in Figure 2.4, for several levels α, the evolution of the latter as a function of the correlation coefficient. As for the linear Gaussian model, we observe that the total QOSA indices tend to zero as $|\rho| \to 1$ and they are lower than the first-order ones for some correlation coefficients.

Hence, we have $S_i \leqslant ST_i^\alpha$, $i = 1, \ldots, d$ for additive models with independent inputs. But this context is far from reality for many concrete examples and this inequality is no longer valid outside this framework as outlined by examples presented in Subsections 2.5.1.1 and 2.5.1.3. This therefore makes the interpretation of the

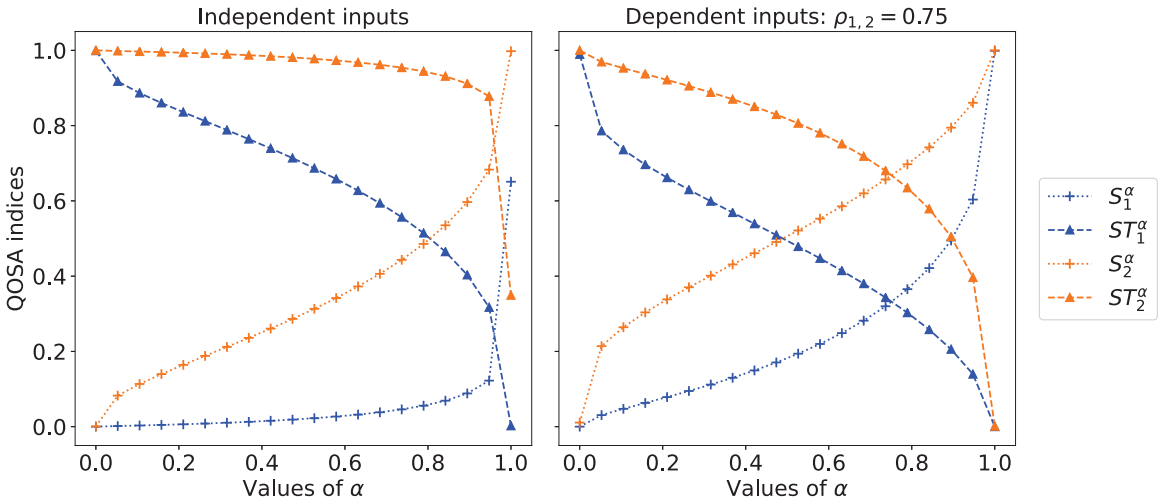


Figure 2.3 *First-order and total QOSA indices with independent (resp. dependent) inputs on the left (resp. right) graph.*

indices complicated.
Furthermore, in the case of dependent inputs, the behaviour of the QOSA indices should be compared to that of Sobol indices. Indeed, whatever the model (additive or not), it may happen in this scheme that the first-order QOSA indices are higher than the total ones depending on the correlation level. We have also observed that total indices tend to zero as the absolute value of the correlation goes to 1. These similar phenomenons observed for the Sobol indices were elucidated by Mara et al. [2015] thanks to a method based on the calculation of four sensitivity indices per input (see Subsection 2.3.2).

We could establish a similar strategy in order to better understand the impact of inputs in case of statistical dependence over the α-quantile, i.e., if their contribution derives from their marginal importance or their dependence with another variable. But, we prefer to turn to the Shapley values introduced Subsection 2.3.3 which present good properties for both independent and dependent inputs. Indeed, they allocate fairly to each input the interaction and/or dependency effect in which it is involved.

2.6 Quantile oriented Shapley effects

In this section, we propose to use the Shapley value defined in Equation (2.17) and recalled below for the sake of completness in order to quantify the impact of each input over the α-quantile of the output distribution

$$v^i = \sum_{\mathcal{J} \subseteq \mathcal{D} \setminus \{i\}} \frac{(d - |\mathcal{J}| - 1)!|\mathcal{J}|!}{d!} \left(c\left(\mathcal{J} \cup \{i\}\right) - c\left(\mathcal{J}\right) \right) , \tag{2.40}$$

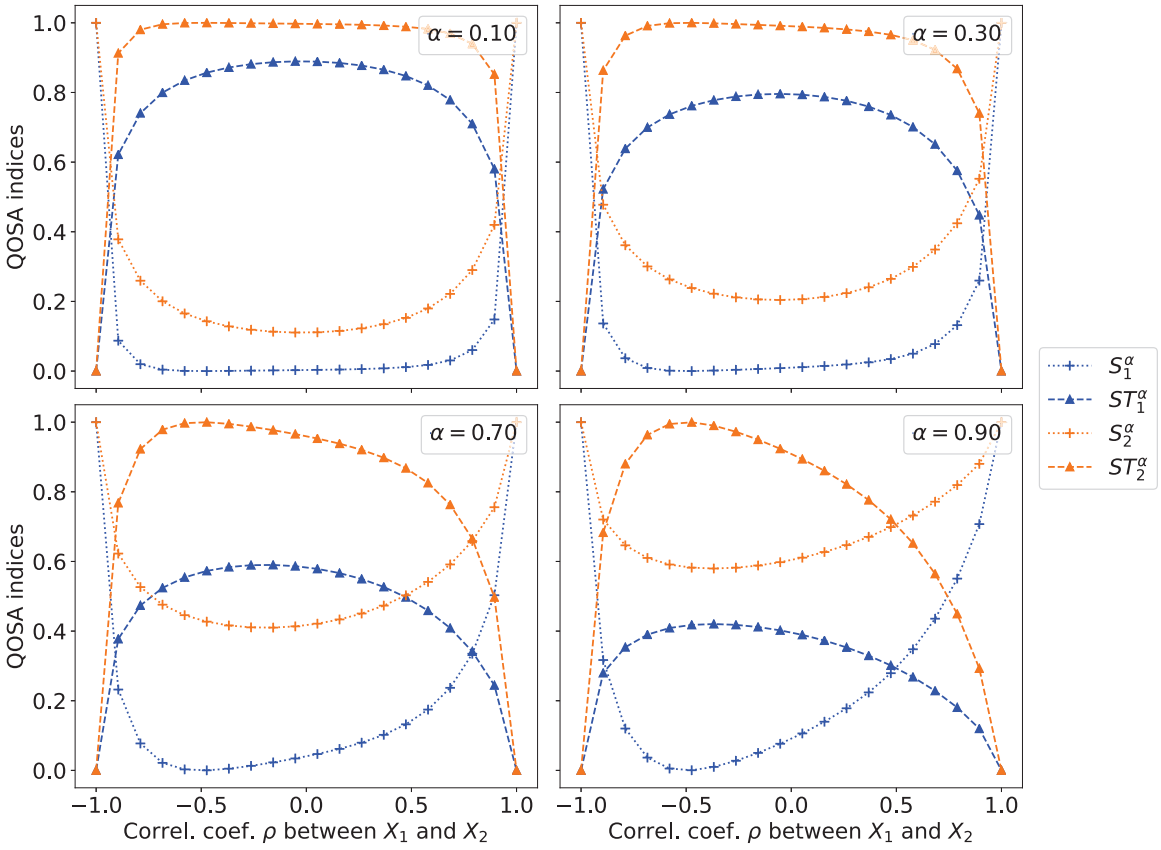


Figure 2.4 *Evolution of the first-order and total QOSA indices at different values of ρ for several levels α.*

with $c(\cdot)$ a generic cost function which maps the exploratory power generated by each subset $\mathcal{J} \subseteq \mathcal{D}$.

Shapley value was first adapted within the framework of variance-based sensitivity measures to measure how much of $\mathrm{Var}(Y)$ can be attributed to each X_i. Indeed, Owen [2014] and Song et al. [2016] proposed to use the two following unnormalized cost functions to measure the variance of Y caused by the uncertainty of the inputs in the subset $\mathcal{J} \subseteq \mathcal{D}$ also named as being the explanatory power created by $\mathcal{J}$:

$$\tilde{c}(\mathcal{J}) = \mathrm{Var}\left(\mathbb{E}\left[Y|\mathbf{X}_{\mathcal{J}}\right]\right) \text{ and } c(\mathcal{J}) = \mathbb{E}\left[\mathrm{Var}\left(Y|\mathbf{X}_{-\mathcal{J}}\right)\right] \ . \tag{2.41}$$

However, we have seen in Section 2.4 that measuring the variance of Y caused by the uncertainty of the inputs in $\mathcal{J}$ is equivalent to assess the impact of the inputs over the expected output. Thus, when using the cost functions given in (2.41), the feature of interest of the output considered is the expectation denoted by $\theta^*(Y) = \mathbb{E}[Y]$. We show in the left-hand column in Table 2.1 that both cost functions may be rewritten according to the contrast function related to the expectation as well as the conditional feature $\theta^*(Y|X_{\mathcal{J}}) = \mathbb{E}[Y|\mathbf{X}_{\mathcal{J}}]$ for the first cost function and $\theta^*(Y|X_{-\mathcal{J}}) = \mathbb{E}[Y|\mathbf{X}_{-\mathcal{J}}]$ for the second one.

Feature of interest	
$\theta^*(Y) = \mathbb{E}[Y]$	$\theta^*(Y) = q^\alpha(Y)$
Contrast function	
$\psi(y,\theta) = (y-\theta)^2$	$\psi(y,\theta) = (y-\theta)\left(\alpha - \mathbb{1}_{\{y\leqslant\theta\}}\right)$
Average contrast function	
$\mathrm{Var}(Y) = \mathbb{E}[\psi(Y,\ \mathbb{E}[Y])]$ $= \mathbb{E}[\psi(Y,\ \theta^*(Y))]$	$\Upsilon(Y) = \mathbb{E}[\psi(Y,\ \theta^*(Y))]$ $= \mathbb{E}[\psi(Y,\ q^\alpha(Y))]$
First cost function	
$\tilde{c}(\mathcal{J}) = \mathrm{Var}(\mathbb{E}[Y\,\vert\,\mathbf{X}_\mathcal{J}])$ $= \mathbb{E}[\psi(\mathbb{E}[Y\,\vert\,\mathbf{X}_\mathcal{J}],\ \mathbb{E}[Y])]$ $= \mathbb{E}[\psi(\theta^*(Y\,\vert\,X_\mathcal{J}),\ \theta^*(Y))]$ $\tilde{c}(\emptyset) = 0 \qquad \tilde{c}(\mathcal{D}) = \mathrm{Var}(Y)$	$\tilde{c}(\mathcal{J}) = \mathbb{E}[\psi(\theta^*(Y\,\vert\,X_\mathcal{J}),\ \theta^*(Y))]$ $= \mathbb{E}[\psi(q^\alpha(Y\,\vert\,\mathbf{X}_\mathcal{J}),\ q^\alpha(Y))]$ $\tilde{c}(\emptyset) = 0 \qquad \tilde{c}(\mathcal{D}) = \Upsilon(Y)$
Second cost function	
$c(\mathcal{J}) = \mathbb{E}[\mathrm{Var}(Y\,\vert\,\mathbf{X}_{-\mathcal{J}})]$ $= \mathbb{E}[\psi(Y,\ \mathbb{E}[Y\,\vert\,\mathbf{X}_{-\mathcal{J}}])]$ $= \mathbb{E}[\psi(Y,\ \theta^*(Y\,\vert\,X_{-\mathcal{J}}))]$ $c(\emptyset) = 0 \qquad c(\mathcal{D}) = \mathrm{Var}(Y)$	$c(\mathcal{J}) = \mathbb{E}[\psi(Y,\ \theta^*(Y\,\vert\,X_{-\mathcal{J}}))]$ $= \mathbb{E}[\psi(Y,\ q^\alpha(Y\,\vert\,\mathbf{X}_{-\mathcal{J}}))]$ $c(\emptyset) = 0 \qquad c(\mathcal{D}) = \Upsilon(Y)$

Table 2.1 *Analogy of the cost functions used for quantifying the impact of the inputs over the expectation for the case where the quantile is the feature of interest.*

Thus, in order to define indices for another feature of interest, the idea is to substitute the contrast function of the expectation by that associated with the feature of interest required.

Let us now use $\theta^*(Y)$, $\theta^*(Y\,\vert\,X_\mathcal{J})$ and $\theta^*(Y\,\vert\,X_{-\mathcal{J}})$ for $\mathcal{J}\subseteq\mathcal{D}$ as generic expressions to designate a feature of interest and the conditional ones related to a contrast function ψ. The impact of the inputs over $\theta^*(Y)$ is therefore assessed by measuring their contribution to the averaged contrast function $\mathbb{E}[\psi(Y,\ \theta^*(Y))]$. This one can be seen as a relevant distance allowing to quantify the variability around the feature of interest. The contributions of the inputs are then calculated with the following cost functions measuring the explanatory power of the subset $\mathcal{J}\subseteq\mathcal{D}$

$$\tilde{c}(\mathcal{J}) = \mathbb{E}[\psi(\theta^*(Y\,\vert\,X_\mathcal{J}),\ \theta^*(Y))] \text{ and } c(\mathcal{J}) = \mathbb{E}[\psi(Y,\ \theta^*(Y\,\vert\,X_{-\mathcal{J}}))]\ . \tag{2.42}$$

Nevertheless, for these cost functions to be a valid choices, they must satisfy that the empty set creates no value, and that all inputs generate $\mathbb{E}[\psi(Y,\ \theta^*(Y))]$. This is for example verified for all contrast functions listed in Section 2.4. Hence, this approach allows to propose ***Shapley effects subordinated to a contrast function*** and therefore subordinated to a specific feature of the output distribution.

Following this strategy, we propose to quantify the impact of the inputs over the

α-quantile of the model output by using the corresponding contrast function as well as the two unnormalized cost functions established in the right-hand column in Table 2.1. However, we define the ***Quantile oriented Shapley effects*** denoted by Sh_i^α with only the second cost function because it verifies that the incremental cost $c(\mathcal{J} \cup \{i\}) - c(\mathcal{J})$ is positive so that the index cannot be negative. Indeed, for $\mathcal{J} \subseteq \mathcal{D} \backslash \{i\}$, we have

$$\begin{aligned} c(\mathcal{J} \cup i) - c(\mathcal{J}) &= \left(\alpha \mathbb{E}[Y] - \mathbb{E}\left[Y \mathbb{1}_{\{Y \leqslant q^\alpha(Y|\mathbf{X}_{-\mathcal{J}\cup i})\}}\right]\right) - \left(\alpha \mathbb{E}[Y] - \mathbb{E}\left[Y \mathbb{1}_{\{Y \leqslant q^\alpha(Y|\mathbf{X}_{-\mathcal{J}})\}}\right]\right) \\ &= \mathbb{E}\left[\left(Y - q^\alpha\left(Y|\,\mathbf{X}_{-\mathcal{J}\cup i}\right)\right)\left(\mathbb{1}_{\{Y \leqslant q^\alpha(Y|X_{-\mathcal{J}})\}} - \mathbb{1}_{\{Y \leqslant q^\alpha(Y|\mathbf{X}_{-\mathcal{J}\cup i})\}}\right)\right] \geqslant 0\,. \end{aligned}$$

At this stage, this property has not yet been demonstrated for the first cost function.

We study in the next subsection examples whose analytical values of the index Sh_i^α are computed by using the cost function normalized by the quantity $\Upsilon(Y)$ introduced in Table 2.1, so that $\sum_{i=1}^{d} Sh_i^\alpha = 1$.

2.6.1 Special cases

Here, we reuse the special cases established in Subsections 2.5.1.2 and 2.5.1.3 for which we obtain a closed form for the Quantile oriented Shapley effects. The aim is to show that these new indices give sensible answers compared to the classical QOSA indices defined in Section 2.5.

2.6.1.1 Linear model with Gaussian input variables

We obtain the following analytical values for the linear model with Gaussian inputs:

Proposition 2.6.1.
If $Y = \eta(\mathbf{X}) = \beta_0 + \boldsymbol{\beta}^\mathsf{T}\mathbf{X}$ with $\beta_0 \in \mathbb{R}$, $\boldsymbol{\beta} \in \mathbb{R}^d$ and $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$ is a positive-definite matrix, then the Quantile oriented Shapley effect for the variable i at the α-level is

$$\begin{aligned} Sh_i^\alpha = \frac{1}{d \cdot \sigma_Y} \sum_{\mathcal{J} \subseteq \mathcal{D}\backslash\{i\}} \binom{d-1}{|\mathcal{J}|}^{-1} \Bigg[&\sqrt{\boldsymbol{\beta}_{\mathcal{J}+i}^\mathsf{T}\left(\boldsymbol{\Sigma}_{\mathcal{J}+i,\mathcal{J}+i} - \boldsymbol{\Sigma}_{\mathcal{J}+i,-\mathcal{J}-i}\boldsymbol{\Sigma}_{-\mathcal{J}-i,-\mathcal{J}-i}^{-1}\boldsymbol{\Sigma}_{-\mathcal{J}-i,\mathcal{J}+i}\right)\boldsymbol{\beta}_{\mathcal{J}+i}} \\ &- \sqrt{\boldsymbol{\beta}_{\mathcal{J}}^\mathsf{T}\left(\boldsymbol{\Sigma}_{\mathcal{J},\mathcal{J}} - \boldsymbol{\Sigma}_{\mathcal{J},-\mathcal{J}}\boldsymbol{\Sigma}_{-\mathcal{J},-\mathcal{J}}^{-1}\boldsymbol{\Sigma}_{-\mathcal{J},\mathcal{J}}\right)\boldsymbol{\beta}_{\mathcal{J}}} \Bigg] \end{aligned} \tag{2.43}$$

with $\sigma_Y^2 = \mathrm{Var}(Y) = \boldsymbol{\beta}^\mathsf{T}\boldsymbol{\Sigma}\boldsymbol{\beta}$, and $\mathcal{J} + i$ (resp. $-\mathcal{J} - i$), a notational compression for $\mathcal{J} \cup \{i\}$ (resp. $-\mathcal{J} \cup \{i\}$).

As for the QOSA index, we may notice that β_0 and $\boldsymbol{\mu}$ do not play any role as translation parameters and that the index does not depend on the α-level which is

a specificity of the linear Gaussian model as explained previously. Nevertheless, no general conclusion can be drawn, therefore we consider the case $d = 2$ with

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \ \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \text{ and } \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}, \ -1 \leq \rho \leq 1, \ \sigma_1 > 0, \ \sigma_2 > 0.$$

We have $\sigma_Y^2 = Var\,(Y) = \beta_1^2\sigma_1^2 + 2\rho\beta_1\beta_2\sigma_1\sigma_2 + \beta_2^2\sigma_2^2$ and obtain from(2.43)

$$\begin{aligned} Sh_1^\alpha &= \frac{1}{2} - \frac{|\beta_2|\,\sigma_2\sqrt{1-\rho^2}}{2\cdot\sigma_Y} + \frac{|\beta_1|\,\sigma_1\sqrt{1-\rho^2}}{2\cdot\sigma_Y}\,, \\ Sh_2^\alpha &= \frac{1}{2} - \frac{|\beta_1|\,\sigma_1\sqrt{1-\rho^2}}{2\cdot\sigma_Y} + \frac{|\beta_2|\,\sigma_2\sqrt{1-\rho^2}}{2\cdot\sigma_Y}\,. \end{aligned} \tag{2.44}$$

As expected, we have $\sum_{i=1}^{2} Sh_i^\alpha = 1$ and we observe that the correlation effects on the first-order QOSA indices (e.g. $\sigma_Y - |\beta_2|\,\sigma_2\sqrt{1-\rho^2}$ for X_1) and on the total QOSA indices (e.g. $|\beta_1|\,\sigma_1\sqrt{1-\rho^2}$ for X_1) are allocated half to the Quantile oriented Shapley effects. We also see that the Shapley effects are equal when the correlation is maximum (i.e. $|\rho| = 1$).

Figure 2.5 presents the first-order and total QOSA indices as well as the Quantile oriented Shapley effects for the particular case $\mu_1 = \mu_2 = 0$, $\beta_1 = \beta_2 = 1$, $\sigma_1 = 1$ and $\sigma_2 = 2$.
On the left-hand graph of the figure, we see that the Shapley effects are also constant and they are brackected by the first-order and total QOSA indices : $S_i^\alpha \leqslant Sh_i^\alpha \leqslant ST_i^\alpha$, $i = 1, 2$. This is due to the fact that the first-order QOSA index omits interaction effects, while the total one overcounts them relative to Shapley that fairly shares this effect on each variable involved within.

We illustrate on the right-hand graph the evolution of the indices as a function of the correlation between the two inputs. As X_2 is the more uncertain variable, its sensitivity indices are larger than those of X_1. Then, although the values are not identical, we can note that the shape of the curves is exactly the same as that observed for the variance-based Shapley effects calculated for the two-dimensional Gaussian linear model (with the same setting) in Iooss and Prieur [2019]. Indeed, we observe that in the presence of correlation, the Quantile oriented Shapley effects lie between the first-order QOSA indices and the total ones with either $S_i^\alpha \leqslant Sh_i^\alpha \leqslant ST_i^\alpha$ or $ST_i^\alpha \leqslant Sh_i^\alpha \leqslant S_i^\alpha$, $i = 1, 2$. This phenomenon is called the "sandwich effect" within the variance framework in Iooss and Prieur [2019]. Finally, as for the variance-based Shapley effects, it also seems that the dependence between the two inputs lead to a rebalancing of their respective Quantile oriented Shapley effects.

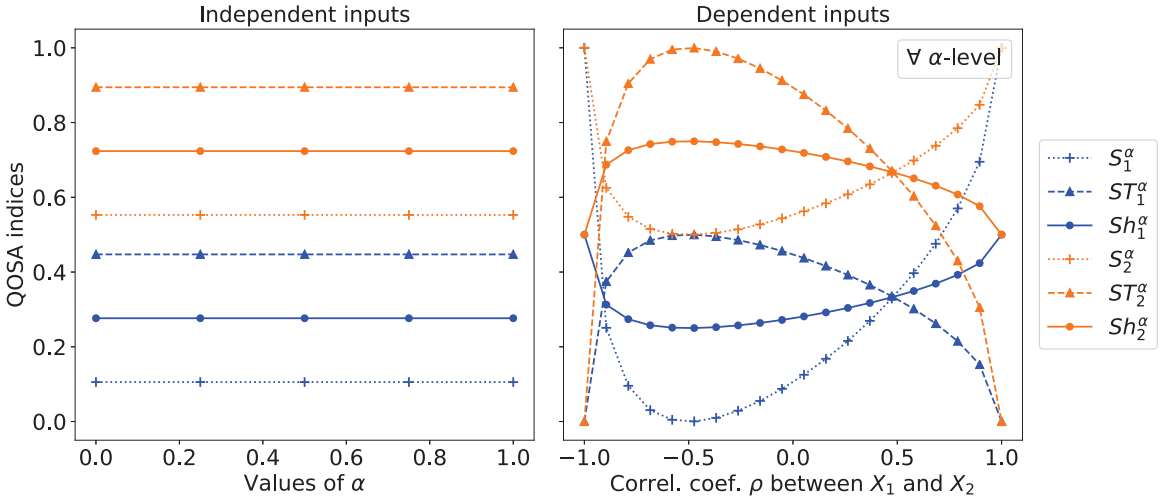


Figure 2.5 *First-order and total QOSA indices as well as the Quantile oriented Shapley effects with independent (resp. dependent) inputs on the left (resp. right) graph.*

2.6.1.2 Gaussian input variables, exponential η

We analyze in this subsection the analytical values below for the model with Gaussian inputs and the resulting output Log-normal distributed.

Proposition 2.6.2.
If $Y = \eta(\mathbf{X}) = \exp\left(\beta_0 + \boldsymbol{\beta}^\mathsf{T}\mathbf{X}\right)$ *with* $\beta_0 \in \mathbb{R}$, $\boldsymbol{\beta} \in \mathbb{R}^d$ *and* $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *where* $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$ *is a positive-definite matrix, then the Quantile oriented Shapley effect for the variable* i *at the* α*-level is*

$$Sh_i^\alpha = \frac{1}{d\cdot A}\sum_{\mathcal{J}\subseteq\mathcal{D}\setminus\{i\}}\binom{d-1}{|\mathcal{J}|}^{-1}\left[\Phi\left(\Phi^{-1}(\alpha) - B(\mathcal{J})\right) - \Phi\left(\Phi^{-1}(\alpha) - C(\mathcal{J},i)\right)\right] \tag{2.45}$$

with

$$A = \alpha - \Phi\left(\Phi^{-1}(\alpha) - \sigma\right) \text{ and } \sigma^2 = \boldsymbol{\beta}^\mathsf{T}\boldsymbol{\Sigma}\boldsymbol{\beta}\ ,$$
$$B(\mathcal{J}) = \sqrt{\boldsymbol{\beta}_{\mathcal{J}}^\mathsf{T}\left(\boldsymbol{\Sigma}_{\mathcal{J},\mathcal{J}} - \boldsymbol{\Sigma}_{\mathcal{J},-\mathcal{J}}\boldsymbol{\Sigma}_{-\mathcal{J},-\mathcal{J}}^{-1}\boldsymbol{\Sigma}_{-\mathcal{J},\mathcal{J}}\right)\boldsymbol{\beta}_{\mathcal{J}}}\ ,$$
$$C(\mathcal{J},i) = \sqrt{\boldsymbol{\beta}_{\mathcal{J}+i}^\mathsf{T}\left(\boldsymbol{\Sigma}_{\mathcal{J}+i,\mathcal{J}+i} - \boldsymbol{\Sigma}_{\mathcal{J}+i,-\mathcal{J}-i}\boldsymbol{\Sigma}_{-\mathcal{J}-i,-\mathcal{J}-i}^{-1}\boldsymbol{\Sigma}_{-\mathcal{J}-i,\mathcal{J}+i}\right)\boldsymbol{\beta}_{\mathcal{J}+i}}\ ,$$

where $\mathcal{J}+i$ *(resp.* $-\mathcal{J}-i$*) is a notational compression for* $\mathcal{J}\cup\{i\}$ *(resp.* $-\mathcal{J}\cup\{i\}$*).*

As for the QOSA indices, we observe that β_0 and $\boldsymbol{\mu}$ do not play any role and that the indices depend on α compared to the linear Gaussian model. However, it is difficult

to reach a conclusion from Equation (2.45). Accordingly, we consider the particular case $d = 2$ with

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \ \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \text{ and } \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}, \ -1 \leq \rho \leq 1, \ \sigma_1 > 0, \ \sigma_2 > 0.$$

We have $\sigma^2 = \beta_1^2\sigma_1^2 + 2\rho\beta_1\beta_2\sigma_1\sigma_2 + \beta_2^2\sigma_2^2$ and obtain from (2.45)

$$\begin{aligned} Sh_1^\alpha &= \frac{1}{2} + \frac{1}{2} \cdot \frac{\Phi\left(\Phi^{-1}(\alpha) - |\beta_2|\,\sigma_2\sqrt{1-\rho^2}\right)}{\alpha - \Phi\left(\Phi^{-1}(\alpha) - \sigma\right)} - \frac{1}{2} \cdot \frac{\Phi\left(\Phi^{-1}(\alpha) - |\beta_1|\,\sigma_1\sqrt{1-\rho^2}\right)}{\alpha - \Phi\left(\Phi^{-1}(\alpha) - \sigma\right)}, \\ Sh_2^\alpha &= \frac{1}{2} + \frac{1}{2} \cdot \frac{\Phi\left(\Phi^{-1}(\alpha) - |\beta_1|\,\sigma_1\sqrt{1-\rho^2}\right)}{\alpha - \Phi\left(\Phi^{-1}(\alpha) - \sigma\right)} - \frac{1}{2} \cdot \frac{\Phi\left(\Phi^{-1}(\alpha) - |\beta_2|\,\sigma_2\sqrt{1-\rho^2}\right)}{\alpha - \Phi\left(\Phi^{-1}(\alpha) - \sigma\right)}. \end{aligned} \tag{2.46}$$

We adopt the next settings in all further tests: $\mu_1 = \mu_2 = 0, \beta_1 = \beta_2 = 1, \sigma_1 = 1$ and $\sigma_2 = 2$.

The analytical values of the first-order, total QOSA indices and the Quantile oriented Shapley effects are illustrated in Figure 2.6 for both independent inputs and correlated inputs with $\rho_{1,2} = 0.75$. The "sandwich effect" which was noticed in the linear Gaussian model in the presence of correlation is also observed here. Indeed, both in the dependent and independent cases and for all the levels α, Shapley effects lie between the first-order and total QOSA indices.

Besides, with the three indices, we obtain the same ranking of the inputs for all α-levels but the Shapley effect is easier to interpret because it properly condenses all the information (dependence and interaction effects). For instance, let us focus over the input X_1 on the right-hand graph at the level $\alpha = 0.2$. If we use the first-order QOSA index S_1^α, we conclude that the impact of the input X_1 is low, but not so small because, conversely its total QOSA index is high enough. But, ultimately, it is difficult to quantify precisely on the basis of these two indices the contribution of the input X_1 at level $\alpha = 0.2$. The Shapley index, in contrast, contains the marginal contribution of the variable but also those due to dependence and interaction effects that are correctly allocated to it. It therefore makes easier to express an opinion on the impact of the variable by taking into account all possible contributions. This observation is valid for all the levels α.

Again, to get another insight on the impact of the dependence over the indices, we plot in Figure 2.7, for several levels α, the evolution of the latter as a function of the correlation coefficient. As explained before, the Quantile oriented Shapley effects give a condensed information of all contributions. That explains why we observe that the Shapley effects of both variables are almost equal for small values of α. Conversely, for large values, the variable X_2 is the most influential overall except when $|\rho| \to 1$ where both inputs have the same contribution.

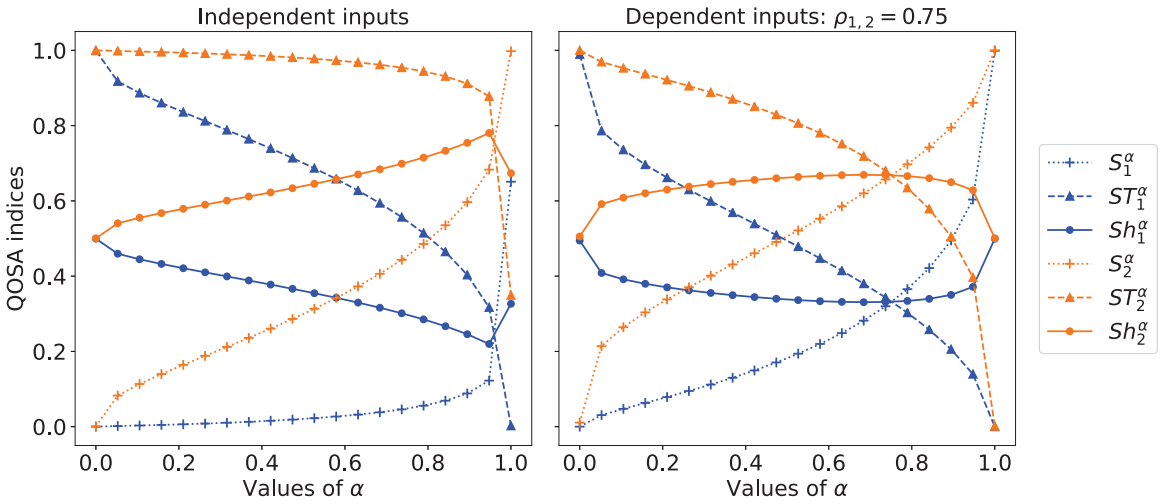


Figure 2.6 *First-order and total QOSA indices as well as the Quantile oriented Shapley effects with independent (resp. dependent) inputs on the left (resp. right) graph.*

Perspective

As a conclusion of this preliminary work, Quantile oriented Shapley effects appear to be a good alternative to the classical QOSA indices. Indeed, they make possible to overcome the various problems encountered with the QOSA indices such as $ST_i^\alpha \leqslant S_i^\alpha$ outside the additive framework or when using dependent inputs. Besides, the Sobol indices stem, for example, from the functional ANOVA decomposition but there is no such a decomposition for the quantiles. Hence, as an allocation method, these indices therefore allow to quantify precisely the contribution of each input at the α-quantile while taking into account interaction and dependency effects.
An additional work would be to compute analytical values of the indices for the Gaussian examples with models of dimension 3 or larger to study their interpretation in higher dimension. Last but not least, the development of an estimation algorithm to compute these new indices would be a significant contribution, but also a difficult task because they involve all subsets of the inputs.

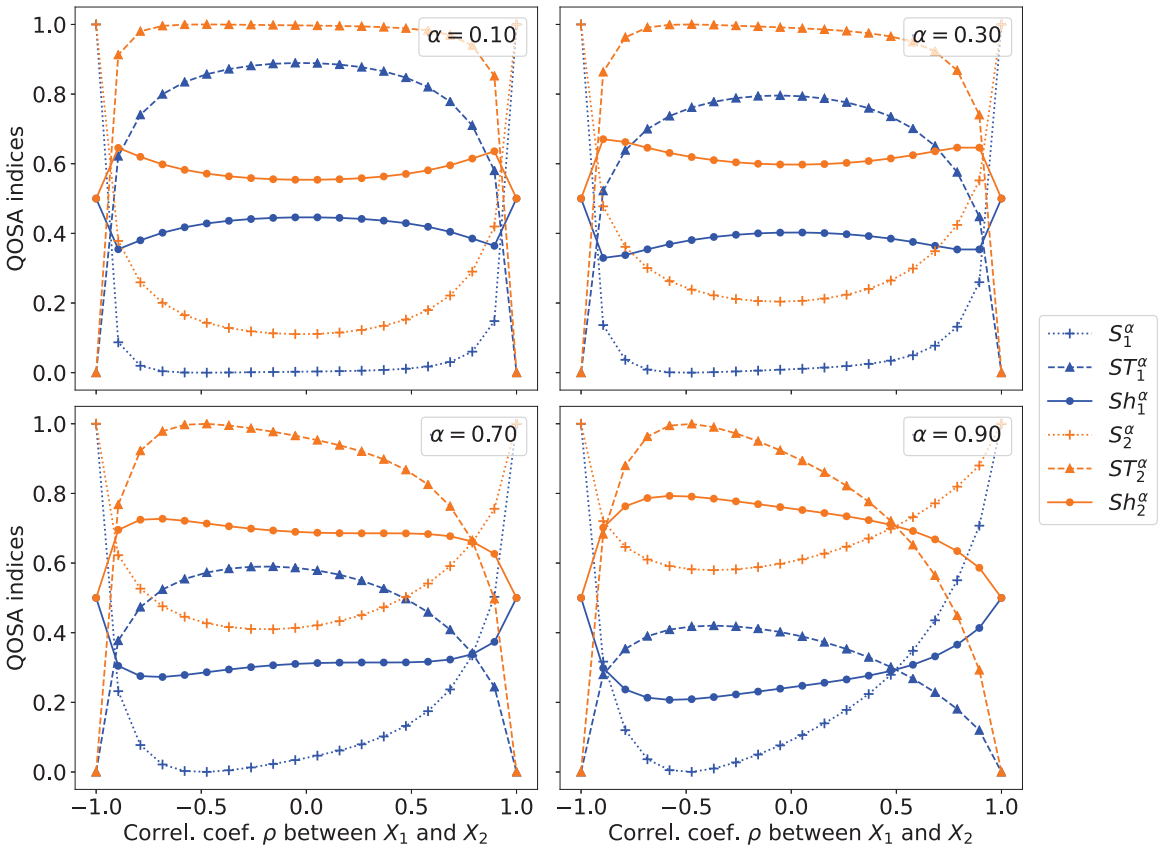


Figure 2.7 *Evolution of the first-order and total QOSA indices as well as the Quantile oriented Shapley effects at different values of ρ for several levels α.*

Chapter 3

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

This chapter consists in the published article Benoumechiara and Elie-Dit-Cosaque [2019].

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

3.1 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 results 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 address 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 inputs, such as Owen and Prieur [2017]; Iooss and Prieur [2019]. 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 dependence contributions between the inputs 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 decreases 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 3.2 recalls the basic concept of Sobol' indices in the independent and dependent configuration; Section 3.3 introduces the Shapley values and their links with sensitivity analysis; Section 3.4 theoretically compares the Sobol' indices and the Shapley effects for two toy examples; Section 3.5 studies the quality of the estimated Shapley effects and their confidence intervals; Section 3.6 introduces the kriging model and how the kriging and Monte Carlo errors can be separated from the overall error; Section 3.7 compares the indice performances using a kriging model on two toy examples; finally, Section 3.8 synthesizes this work and suggests some perspectives.

3.2 Sobol' sensitivity indices

3.2.1 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

$$\eta(\mathbf{X}) = \eta_\emptyset + \sum_{i=1}^{d} \eta_i(X_i) + \sum_{1 \leqslant i < j \leqslant d} \eta_{i,j}(X_i, X_j) + \cdots + \eta_{1,\dots,d}(\mathbf{X}) \ , \tag{3.1}$$

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

$$\int_0^1 \eta_{i_1,i_2,\dots,i_s}(x_{i_1}, x_{i_2}, \dots, x_{i_s}) \mathrm{d}x_{i_w} = 0 \ , \tag{3.2}$$

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 (3.1) are orthogonal to one another. This property implies the independence of the random variables X_i in the stochastic configuration and allows to obtain the following expressions for the functions $\eta_{i_1,i_2,\dots,i_s}$ of (3.1) :

$$\eta_\emptyset = \mathbb{E}(Y) \ , \tag{3.3}$$

$$\eta_i(X_i) = \mathbb{E}_{\mathbf{X}_{\sim i}}(Y|X_i) - \mathbb{E}(Y) \ , \tag{3.4}$$

$$\eta_{i,j}(X_i, X_j) = \mathbb{E}_{\mathbf{X}_{\sim ij}}(Y|X_i, X_j) - \eta_i - \eta_j - \mathbb{E}(Y) \ , \tag{3.5}$$

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 (3.1) leads to the functional ANalysis Of VAriance (ANOVA) which consists in expanding the global variance into a sum of partial variances such as

$$\mathrm{Var}(Y) = \sum_{i=1}^{d} \mathrm{Var}[\eta_i(X_i)] + \sum_{i=1}^{d} \sum_{i<j}^{d} \mathrm{Var}[\eta_{i,j}(X_i, X_j)] + \cdots + \mathrm{Var}[\eta_{1,\dots,d}(\mathbf{X})] \ . \tag{3.6}$$

The so-called Sobol' indices [Sobol, 1993] can be derived from (3.6) by dividing both sides with $\mathrm{Var}(Y)$. This operation results in the following property:

$$\sum_{i=1}^{d} S_i + \sum_{i=1}^{d}\sum_{i<j}^{d} S_{ij} + \cdots + S_{1,\ldots,d} = 1 \ , \tag{3.7}$$

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

$$S_i = \frac{\mathrm{Var}[\eta_i(X_i)]}{\mathrm{Var}(Y)}, \qquad S_{ij} = \frac{\mathrm{Var}[\eta_{i,j}(X_i, X_j)]}{\mathrm{Var}(Y)}, \qquad \ldots \tag{3.8}$$

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 index S_{ij} measures 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

$$ST_i = S_i + \sum_{i\neq j} S_{ij} + \cdots + S_{1,\ldots,d} = 1 - \frac{\mathrm{Var}_{\mathbf{X}_{\sim i}}[\mathbb{E}_{X_i}(Y|\mathbf{X}_{\sim i})]}{\mathrm{Var}(Y)} = \frac{\mathbb{E}_{\mathbf{X}_{\sim i}}[\mathrm{Var}_{X_i}(Y|\mathbf{X}_{\sim i})]}{\mathrm{Var}(Y)} \ . \tag{3.9}$$

The property (3.7) does not always hold for the total indices as summing total indices for all variables introduces redundant interactions terms appearing only once in (3.7). 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, (3.6) 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.

3.2.2 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 $\mathbf{U} \sim \mathcal{U}([0,1]^d)$ 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

$$T : \mathbb{R}^d \to [0,1]^d$$
$$\mathbf{x} \mapsto \mathbf{u} = \begin{pmatrix} F_1(x_1) \\ \vdots \\ F_{k|1,\dots,k-1}(x_k|x_1,\dots,x_{k-1}) \\ \vdots \\ F_{d|1,\dots,d-1}(x_d|x_1,\dots,x_{d-1}) \end{pmatrix}$$

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, 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

$$(X_i, X_{i+1},\dots,X_d,X_1,\dots,X_{i-1}) \sim p_{\mathbf{X}} \xrightarrow{T} (U_1^i,\dots,U_d^i) \sim \mathcal{U}([0,1]^d) \ . \tag{3.10}$$

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

$$g_i(\mathbf{U}^i) = g_\emptyset + \sum_{k=1}^{d} g_k(U_k^i) + \sum_{k=1}^{d}\sum_{k<l}^{d} g_{k,l}(U_k^i,U_l^i) + \cdots + g_{1,\dots,d}(U_1^i,\dots,U_d^i) \ , \tag{3.11}$$

where $g_\emptyset = \mathbb{E}[g_i(\mathbf{U}^i)]$. Because the summands in (3.11) are orthogonal, the variance based decomposition can be derived, such that

$$\mathrm{Var}(Y) = \mathrm{Var}[g_i(\mathbf{U}^i)] = \sum_{k=1}^{d} V_k + \sum_{k=1}^{d}\sum_{k<l}^{d} V_{k,l} + \cdots + V_{1,\dots,d} \ , \tag{3.12}$$

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 (3.12) with the total variance such that

$$S_k^i = \frac{\mathrm{Var}[\mathbb{E}[g_i(\mathbf{U}^i)|U_k^i]]}{\mathrm{Var}[g_i(\mathbf{U}^i)]} \ . \tag{3.13}$$

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

$$ST_k^i = \frac{\mathbb{E}[\mathrm{Var}[g_i(\mathbf{U}^i)|\mathbf{U}_{\sim k}^i]]}{\mathrm{Var}[g_i(\mathbf{U}^i)]} \ , \tag{3.14}$$

where $\mathbf{U}^i_{\sim k} = \mathbf{U}^i_{\mathcal{D}\backslash\{k\}}$. We remind that (3.13) and (3.14) are derived from the RT of the ordered set $(X_i, X_{i+1}, \ldots, X_d, X_1, \ldots, X_{i-1})$. The RT in equation (3.10) determines the following mapping between $\mathbf{X}$ and $\mathbf{U}^i$:

$$\Big[(X_i), (X_{i+1}|X_i), \ldots, (X_1|X_i, X_{i+1}, \ldots, X_d), \ldots, (X_{i-1}|\mathbf{X}_{\sim(i-1)})\Big] \longleftrightarrow (U^i_1, U^i_2, \ldots, U^i_d) \ ,$$

where $U^i_1 = F_i(X_i)$, $U^i_2 = F_{i+1|i}(X_{i+1}|X_i)$ and so on for other variables. From here, we only consider the variables U^i_1 and U^i_d because they present interesting properties. Indeed, the variable U^i_1 is representative of the behavior of X_i taking into account the dependence with other variables. On the opposite, the variable U^i_d represents the effects of X_{i-1} that are 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

$$S_i^{full} = \frac{\mathrm{Var}[\mathbb{E}[g_i(\mathbf{U}^i)|U^i_1]]}{\mathrm{Var}[g_i(\mathbf{U}^i)]} = \frac{\mathrm{Var}[\mathbb{E}[\eta(\mathbf{X})|X_i]]}{\mathrm{Var}[\eta(\mathbf{X})]} \ , \tag{3.15}$$

$$ST_i^{full} = \frac{\mathbb{E}[\mathrm{Var}[g_i(\mathbf{U}^i)|\mathbf{U}^i_{\sim 1}]]}{\mathrm{Var}[g_i(\mathbf{U}^i)]} = \frac{\mathbb{E}[\mathrm{Var}[\eta(\mathbf{X})|(\mathbf{X}_{\sim i}|X_i)]]}{\mathrm{Var}[\eta(\mathbf{X})]} \ , \tag{3.16}$$

  where $\mathbf{X}_{\sim i}|X_i$ represents 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

$$S_i^{ind} = \frac{\mathrm{Var}[\mathbb{E}[g_{i+1}(\mathbf{U}^{i+1})|U^{i+1}_d]]}{\mathrm{Var}[g_{i+1}(\mathbf{U}^{i+1})]} = \frac{\mathrm{Var}[\mathbb{E}[\eta(\mathbf{X})|(X_i|\mathbf{X}_{\sim i})]]}{\mathrm{Var}[\eta(\mathbf{X})]} \ , \tag{3.17}$$

$$ST_i^{ind} = \frac{\mathbb{E}[\mathrm{Var}[g_{i+1}(\mathbf{U}^{i+1})|\mathbf{U}^{i+1}_{\sim d}]]}{\mathrm{Var}[g_{i+1}(\mathbf{U}^{i+1})]} = \frac{\mathbb{E}[\mathrm{Var}[\eta(\mathbf{X})|\mathbf{X}_{\sim i}]]}{\mathrm{Var}[\eta(\mathbf{X})]} \ , \tag{3.18}$$

  where $X_i|\mathbf{X}_{\sim i}$ represents 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^i_u which represents 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.

3.2.3 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:

$$\widehat{S}_i^{full} = \frac{\frac{1}{N}\sum_{j=1}^{N} g_i(A)_j g_i(\mathbf{B}_{\mathbf{A}}^{(1)})_j - g_{i_0}^2}{\widehat{V}} , \tag{3.19}$$

$$\widehat{ST}_i^{full} = 1 - \frac{\frac{1}{N}\sum_{j=1}^{N} g_i(B)_j g_i(\mathbf{B}_{\mathbf{A}}^{(1)})_j - g_{i_0}^2}{\widehat{V}} , \tag{3.20}$$

$$\widehat{S}_{i-1}^{ind} = \frac{\frac{1}{N}\sum_{j=1}^{N} g_i(A)_j g_i(\mathbf{B}_{\mathbf{A}}^{(d)})_j - g_{i_0}^2}{\widehat{V}} , \tag{3.21}$$

$$\widehat{ST}_{i-1}^{ind} = 1 - \frac{\frac{1}{N}\sum_{j=1}^{N} g_i(B)_j g_i(\mathbf{B}_{\mathbf{A}}^{(d)})_j - g_{i_0}^2}{\widehat{V}} , \tag{3.22}$$

  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.

3.3 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.

3.3.1 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

$$v^i = \sum_{\mathcal{J} \subseteq \mathcal{D} \backslash \{i\}} \frac{(d - |\mathcal{J}| - 1)!|\mathcal{J}|!}{d!} \left(c\left(\mathcal{J} \cup \{i\}\right) - c\left(\mathcal{J}\right) \right) , \tag{3.23}$$

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 $\mathcal{J} \subseteq \mathcal{D} \backslash \{i\}$.

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.

3.3.2 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

$$v^i = \frac{1}{d!} \sum_{\pi \in \Pi(\mathcal{D})} \left[c\left(P_i(\pi) \cup \{i\}\right) - c\left(P_i(\pi)\right) \right] . \tag{3.24}$$

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

$$\widehat{v}^i = \frac{1}{m} \sum_{l=1}^{m} \Delta_i c(\pi_l) , \tag{3.25}$$

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

$$\widehat{Sh}^i = \frac{1}{m}\sum_{l=1}^{m}\left[\widehat{c}\left(P_i(\pi_l) \cup \{i\}\right) - \widehat{c}\left(P_i(\pi_l)\right)\right] \ , \tag{3.26}$$

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 $C = N_v + m(d-1)N_oN_i$ 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 [2019] have also illustrated the convergence of two numerical algorithms for estimating Shapley effects.

3.3.3 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 $\widehat{\theta}(X_1, \ldots, X_n)$ 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:

$$\left[\widehat{G}^{-1} \circ \Phi(2\widehat{z}_0 + z_{\alpha/2}) \quad ; \quad \widehat{G}^{-1} \circ \Phi(2\widehat{z}_0 - z_{\alpha/2})\right] \tag{3.27}$$

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 1 in order to obtain the distribution of $\widehat{Sh}^i$ to calculate the confidence interval for Sh^i.

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 Iooss and Prieur [2019] 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.

Algorithme 1 : 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 $\mathrm{Var}(Y)$;
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 (3.26) ;
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 by bloc with replacement a realization $\tilde{\mathbf{y}}^{(2)}$ of $\mathbf{y}^{(2)}$;
    Compute $\widehat{Sh}_b^i$ thanks to Equation (3.26). ;
end
Compute confidence intervals for Sh^i with (3.27).

• The first one is to use the CLT like Iooss and Prieur [2019]. Indeed, in Castro et al. [2009] the CLT gives us:

$$\widehat{Sh}^i \xrightarrow[m\to\infty]{\mathcal{L}} \mathcal{N}\left(Sh^i, \frac{\sigma^2}{m}\right) \tag{3.28}$$

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 to estimate the confidence intervals doing a bootstrap on the permutations. We describe in Algorithm 2 the procedure allowing to do that.

3.4 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.

Algorithme 2 : 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 (3.26) ;
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 $\widehat{Sh}^i_b$ thanks to Equation (3.26). ;
end
Compute confidence intervals for Sh^i with (3.27).

3.4.1 Interactive model with two inputs

Let us consider an interactive model

$$Y = (\beta_1 X_1) \times (\beta_2 X_2) \ , \tag{3.29}$$

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 3.1.

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 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 represents marginal contribution due to the dependence. We see in these terms that the

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 3.1 *Sensitivity indices of independent and dependent Gaussian models.*

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 [2019] 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 [2019] 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 $ST_i^{ind} = \frac{\mathbb{E}[\text{Var}(Y|\mathbf{X}_{\sim i})]}{\text{Var}(Y)} = 0$ and $ST_i^{full} = \frac{\mathbb{E}[\text{Var}(Y|(\mathbf{X}_{\sim i}|X_i))]}{\text{Var}(Y)} = 0$ and as the variance is always a positive function, that implies $\text{Var}(Y|\mathbf{X}_{\sim i}) = 0$ and $\text{Var}\left(Y|\left(\mathbf{X}_{\sim i}|X_i\right)\right) = 0$. 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 $\text{Var}(Y)$ and the FF setting. Sure enough, if $Sh^i = 0$, then we have $\forall \mathcal{J} \subseteq \mathcal{D}\backslash\{i\}, \text{Var}\left[Y|\mathbf{X}_{-(\mathcal{J}\cup\{i\})}\right] = \text{Var}\left[Y|\mathbf{X}_{-\mathcal{J}}\right]$ 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 index. However, this share allocation makes the Shapley effects easier to interpret than Sobol' indices and might be a great alternative to the four Sobol' indices. Thus, in the sequel, we will compare Sobol' indices and the Shapley effects on a basic example to see if they make correctly the factor prioritization.

3.4.2 Linear model with three inputs

Let us consider

$$Y = \beta_0 + \boldsymbol{\beta}^{\mathsf{T}}\mathbf{X} \ , \tag{3.30}$$

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

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

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

$$\sigma^2 S_1^{full} = \sigma^2 ST_1^{full} = \left(\beta_1\sigma_1 + \alpha\beta_2\sigma_2 + \rho\beta_3\sigma_3\right)^2 ,$$
$$\sigma^2 S_2^{full} = \sigma^2 ST_2^{full} = \left(\alpha\beta_1\sigma_1 + \beta_2\sigma_2 + \gamma\beta_3\sigma_3\right)^2 ,$$
$$\sigma^2 S_3^{full} = \sigma^2 ST_3^{full} = \left(\rho\beta_1\sigma_1 + \gamma\beta_2\sigma_2 + \beta_3\sigma_3\right)^2 .$$

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

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

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 3.9.1. 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 3.2 *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 [2019] 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 3.3 *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.

3.5 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.

3.5.1 Parametrization of the Shapley algorithms

As defined in Section 3.3.2, 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 index value inside the confidence intervals of the estimation. We consider the three dimensional linear Gaussian model of Section 3.4.2 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 3.1 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_oN_im, 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.

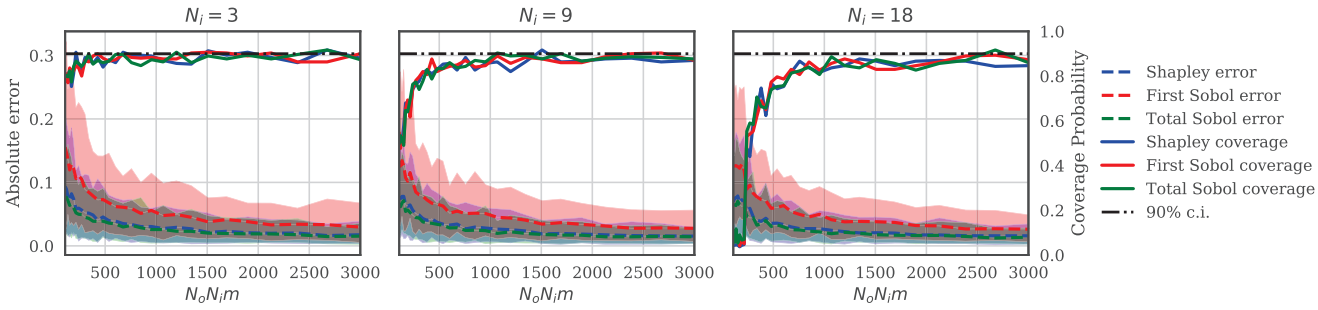


Figure 3.1 *Variation of the absolute error and the POC with N_o for three values of $N_i = 3, 9, 18$ for the exact permutation algorithm ($m = d! = 6$).*

We observe that the estimation error is similar for the three different values of N_i and decreases 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_oN_im,

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 3.2 is similar to Figure 3.1 but for the random permutation algorithm and by fixing $N_o = 1$ and by varying the number of permutations.

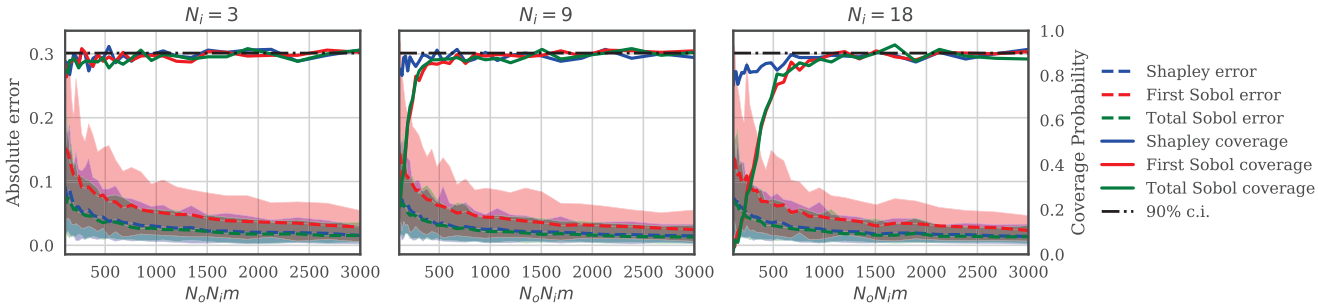


Figure 3.2 *Variation of the absolute error and the POC with m for three values of $N_i = 3, 9, 18$ and $N_o = 1$ for the random permutation algorithm.*

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, 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 3.3 is the same as Figure 3.2 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.

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 3.7.2 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 in Song et al. [2016] explained within 3.3.2 and the suggestion analyzed in Iooss and Prieur [2019].

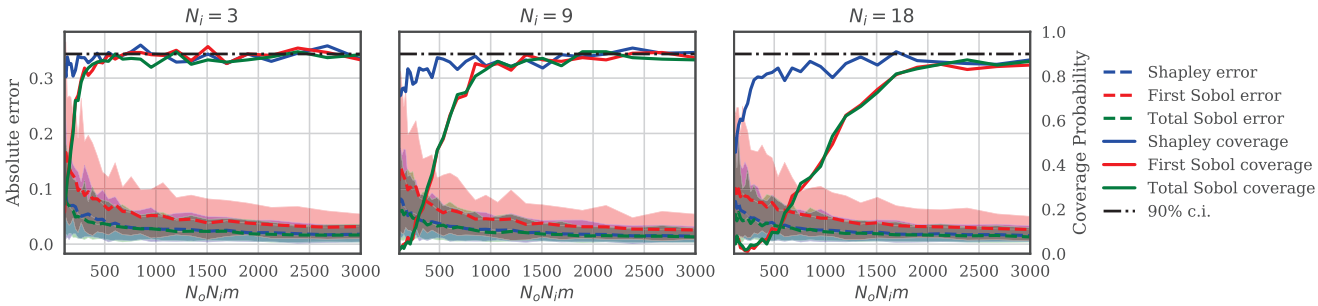


Figure 3.3 *Variation of the absolute error and the POC with m for three values of $N_i = 3, 9, 18$ and $N_o = 3$ for the random permutation algorithm.*

3.5.2 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 3.4 illustrates the same experiment as Figure 3.1 but with a correlation of $\gamma = 0.9$ between X_2 and X_3. We observed that the POC of the total Sobol' index starts to tend to the true probability (at 90%) before slowly decreasing. Thus, it seems that the confidence intervals are underestimated or the index estimation is biased.

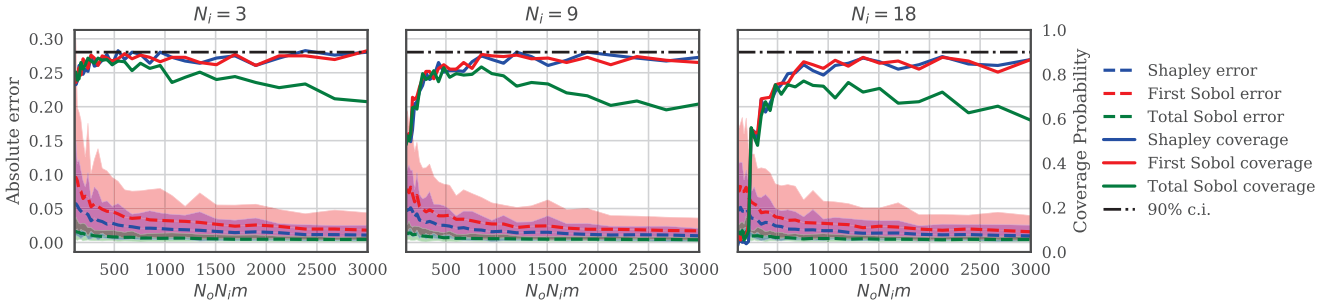


Figure 3.4 *Variation of the absolute error and the POC with N_o for three values of $N_i = 3, 9, 18$ for the exact permutation algorithm and a correlation $\gamma = 0.9$ between X_2 and X_3.*

To verify this observation, Figure 3.5 shows the estimation of the total Sobol' index for $N_v = 20000$, $N_o = 10000$, $N_i = 3$ with the histogram from the bootstrap sampling in blue, the estimated index ST_i in red line and the true index 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 3.4. 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.

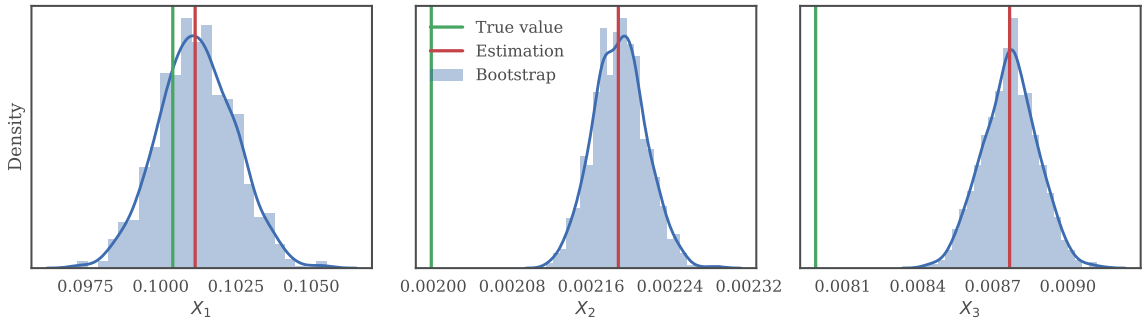


Figure 3.5 *Estimated bootstrap estimations of the total Sobol' indices from the exact Shapley algorithm with a correlation of 0.99 between X_2 and X_3.*

3.5.3 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 3.5.1 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 3.5.1 by increasing the sample-size N. The Figure 3.6 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.

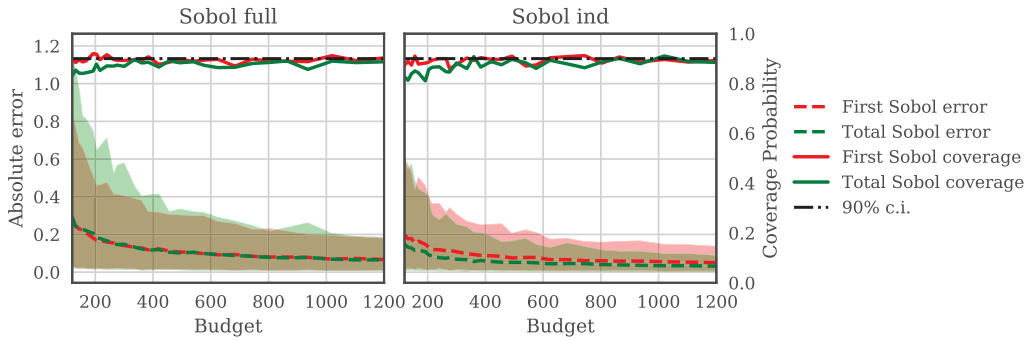


Figure 3.6 *Variation of the absolute error and the POC with the computational cost for the RT method.*

We recall from Section 3.3 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 3.7 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.

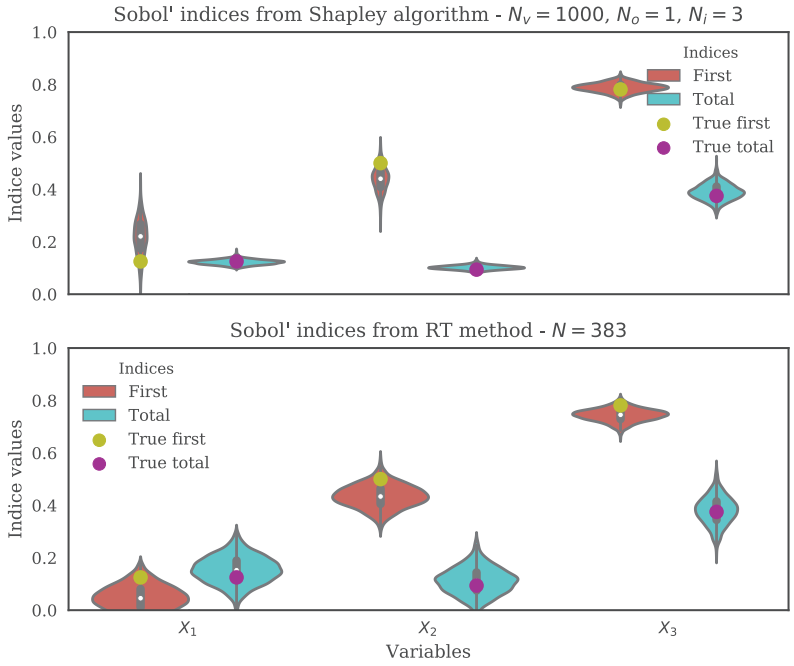


Figure 3.7 *Sobol' indices estimation from the exact permutation method of the Shapley algorithm (top) and the RT method (bottom) using the Janon estimator for similar number of evaluation:* $N_v + N_o N_i m(d-1) = 4Nd = 4600$.

We recall from Section 3.2.3 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 3.8. We observed that the estimation errors from the RT method depends of the used estimator and this error is lower using estimator from Figure 3.8 than the one from Figure 3.7.

The Figure 3.9 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.

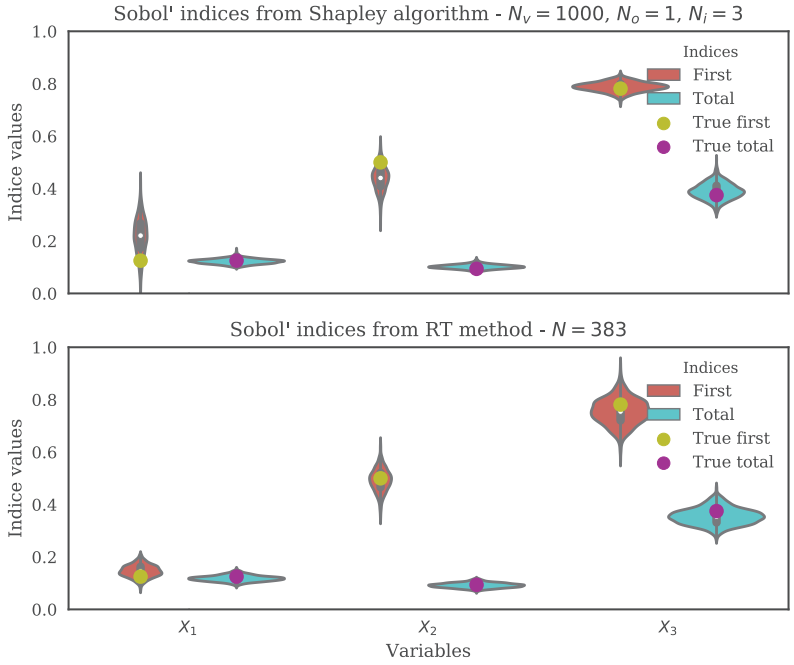


Figure 3.8 *Sobol' indices estimation from the exact permutation method of the Shapley algorithm (top) and the RT method (bottom) using the estimator from Mara et al. [2015] for similar number of evaluation:* $N_v + N_o N_i m(d-1) = 4Nd = 4600$.

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.

3.6 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 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

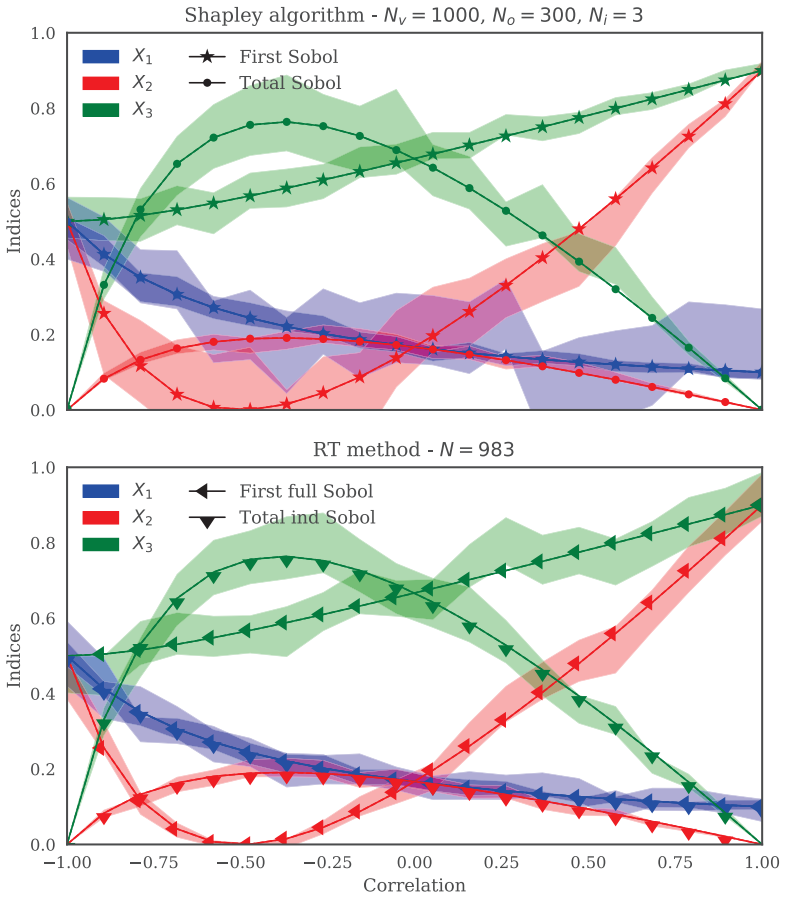


Figure 3.9 *Sobol' indices estimations from the exact permutation method of the Shapley algorithm (top) and the RT method (bottom) in fonction of γ.*

a kriging surrogate model taking into account both the Monte Carlo error and the surrogate model error.

3.6.1 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 conditionally to the value taken by η on a design of experiments $\mathbf{D}$.

Thus, we consider that our expensive function $\eta(x)$ can be modeled by a Gaussian process $H(x)$ whose 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

$$H_n(x) \sim GP\left(m_n(x), s_n^2(x, \tilde{x})\right), \tag{3.31}$$

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

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

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

3.6.2 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 (3.24) which leads to

$$Sh_n^i = \frac{1}{d!}\sum_{\pi\in\Pi(\mathcal{D})}\left[c_n\left(P_i(\pi)\cup\{i\}\right) - c_n\left(P_i(\pi)\right)\right] \ , \tag{3.32}$$

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 (3.26) 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

$$\widehat{Sh}_n^i = \frac{1}{m}\sum_{l=1}^{m}\left[\widehat{c}_n\left(P_i(\pi_l)\cup\{i\}\right) - \widehat{c}_n\left(P_i(\pi_l)\right)\right] \ , \tag{3.33}$$

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

3.6.3 Estimation of errors : Monte Carlo and surrogate model

The estimator (3.33) 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. Note that by decomposing the variance of $\widehat{Sh}_n^i$ in this way, we implicitly assume that there is no interaction effect between the metamodel $H_n(x)$ and the Monte Carlo sample $(\mathbf{X}_{\kappa_l})_{l=1,\dots,B}$.

In section 4 of the article Le Gratiet et al. [2014], they proposed the algorithm (3) that we adapted here to estimate each of these contributions.

The output $\left(\widehat{Sh}_{n,k,l}^i\right)_{\substack{k=1,\dots,N_H \\ l=1,\dots,B}}$ of the algorithm (3) is a sample of size $N_H \times B$ representative of the distribution of $\widehat{Sh}_n^i$ 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.

3.7 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 3.6.3

Algorithme 3 : 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 (3.32) 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 (3.32) from $\{\tilde{\mathbf{y}}^{(1)}, \tilde{\mathbf{y}}^{(2)}\}$;
    end
end
return $\left(\widehat{Sh}^i_{n,k,l}\right)_{\substack{k=1,\ldots,N_H \\ l=1,\ldots,B}}$

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.

3.7.1 Gaussian framework

We use the same configuration as in the Section 3.5.3 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 3.10 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.

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 3.11 allows to distinguish which part the overall error is due to

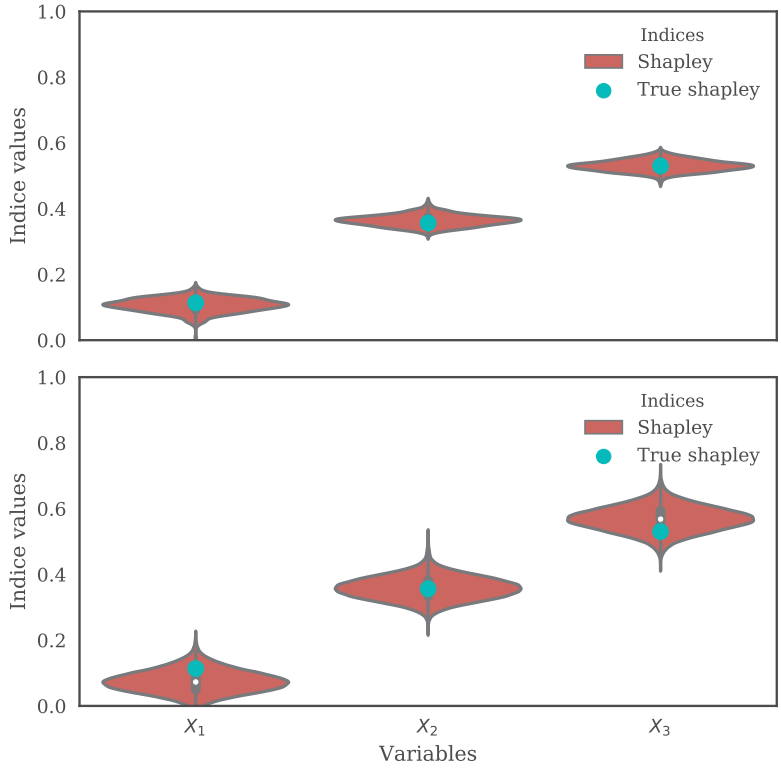


Figure 3.10 *Estimation of the Shapley effects with the exact permutation algorithm. The top and bottom figures respectively show the estimation results with the true function and the kriging model with* $Q^2 = 0.90$.

the kriging. We see immediately that 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 3.10.

3.7.2 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

$$\eta(\mathbf{x}) = \sin(x_1) + 7\sin^2(x_2) + 0.1x_3^4 \sin(x_1) \ . \tag{3.34}$$

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

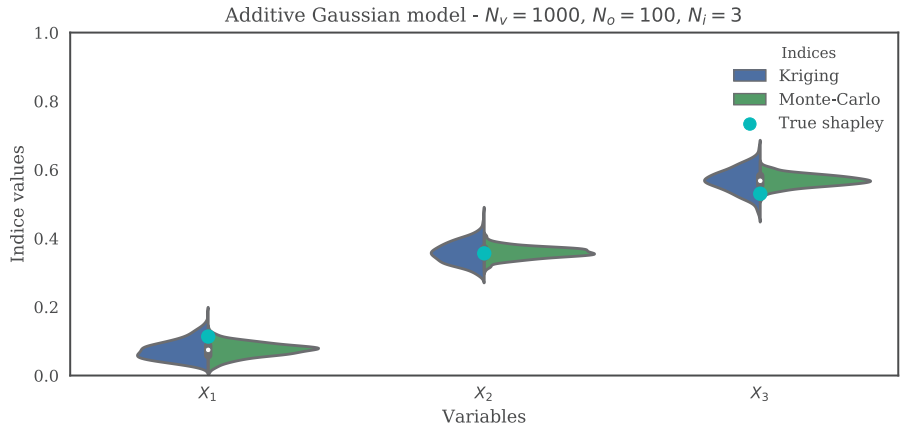


Figure 3.11 *Separation of the uncertainty from the Monte Carlo estimation and the kriging model approximation.*

cumulative distribution function F of $\mathbf{X}$ can be written as

$$F(x_1, x_2, x_3) = C_{\boldsymbol{\rho}}\left(F_1(x_1), F_2(x_2), F_3(x_3)\right) \ , \tag{3.35}$$

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 3.12 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 3.13, 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.

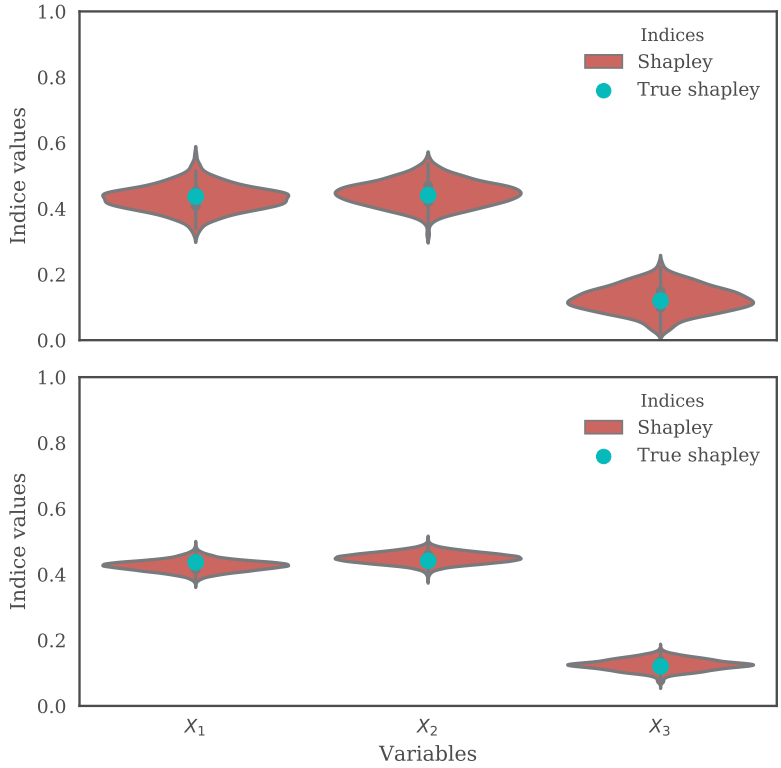


Figure 3.12 *Estimation of the Shapley effects with the exact permutation algorithm. The top and bottom figures respectively show the estimation results with the true function and the kriging model with* $Q^2 = 0.98$.

3.8 Conclusion

Throughout this article, we studied the Shapley effects and the *independent* and *full* Sobol' indices defined in Mara et al. [2015] for the models with a dependence structure on the input variables. The comparison between these indices revealed that:

- 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 include 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

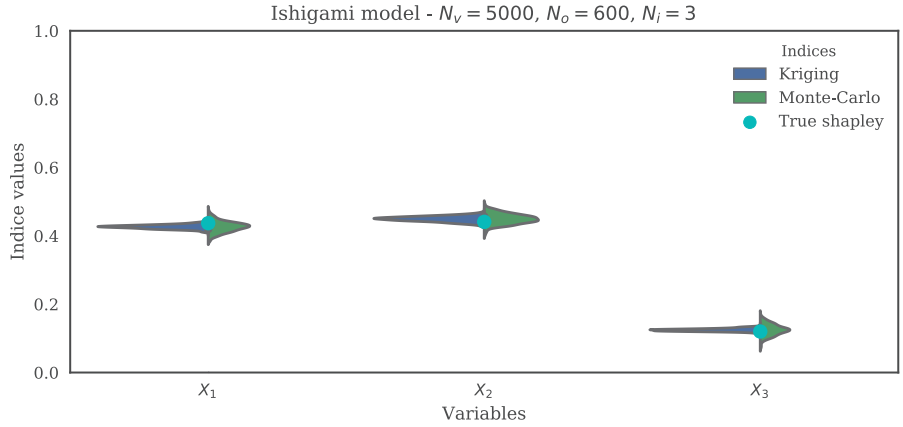


Figure 3.13 *Separation of the uncertainty from the Monte Carlo estimation and the kriging model approximation.*

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 requires large sample sizes 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 [2019] 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 designed 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

We are particularly grateful to Bertrand Iooss, Roman Sueur, Veronique Maume-Deschamps, Clémentine Prieur, Andrés Cuberos and Ecaterina Nisipasu for their supervising during the research session of the CEMRACS'17 and even more. We are also grateful to an anonymous reviewer who considerably helped in improving the manuscript. We would like to thank EDF R&D for the financial support of this project and the organizers of the CEMRACS'17. A Python library shapley has been developed to perform the numerical estimations of the Shapley effects and Sobol' indices with some dependencies such as OpenTURNS and GPflow. This library has been tested with the help of the sensitivity package of the R software.

3.9 Appendix

3.9.1 Gaussian framework: linear model

Let us consider

$$Y = \beta_0 + \boldsymbol{\beta}^{\mathsf{T}}\mathbf{X} \ , \tag{3.36}$$

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

$$\begin{aligned}
\sigma^2 ST_{1|2} &= -\frac{\left(\beta_1\sigma_1\left(\alpha^2 - 1\right) + \beta_3\sigma_3(\alpha\gamma - \rho)\right)^2}{\alpha^2 - 1} \quad \sigma^2 ST_{1|3} = -\frac{\left(\beta_1\sigma_1\left(\rho^2 - 1\right) + \beta_2\sigma_2(\gamma\rho - \alpha)\right)^2}{\rho^2 - 1} \ , \\
\sigma^2 ST_{2|1} &= -\frac{\left(\beta_2\sigma_2\left(\alpha^2 - 1\right) + \beta_3\sigma_3(\alpha\rho - \gamma)\right)^2}{\alpha^2 - 1} \quad \sigma^2 ST_{2|3} = -\frac{\left(\beta_2\sigma_2\left(\gamma^2 - 1\right) + \beta_1\sigma_1(\gamma\rho - \alpha)\right)^2}{\gamma^2 - 1} \ , \\
\sigma^2 ST_{3|1} &= -\frac{\left(\beta_3\sigma_3\left(\rho^2 - 1\right) + \beta_2\sigma_2(\alpha\rho - \gamma)\right)^2}{\rho^2 - 1} \quad \sigma^2 ST_{3|2} = -\frac{\left(\beta_3\sigma_3\left(\gamma^2 - 1\right) + \beta_1\sigma_1(\alpha\gamma - \rho)\right)^2}{\gamma^2 - 1} \ .
\end{aligned}$$

• For $j = 1, 2, 3$, from the definition of Shapley effects, we have

$$\begin{aligned}
Sh_1 &= \frac{1}{3}\left(\left(\tilde{c}(1) - \tilde{c}(\emptyset)\right) + \frac{1}{2}\left(\tilde{c}(1,2) - \tilde{c}(2)\right) + \frac{1}{2}\left(\tilde{c}(1,3) - \tilde{c}(3)\right) + \left(\tilde{c}(1,2,3) - \tilde{c}(2,3)\right)\right) \\
&= \frac{1}{3}\left(S_1^{full} + \frac{1}{2}\left(S_{1,2}^{full} - S_2^{full}\right) + \frac{1}{2}\left(S_{1,3}^{full} - S_3^{full}\right) + \left(S_{1,2,3}^{full} - S_{2,3}^{full}\right)\right) \\
&= \frac{1}{3}\left(S_1^{full} + \frac{1}{2}ST_{1|2} + \frac{1}{2}ST_{1|3} + ST_1^{ind}\right) \ ,
\end{aligned}$$

$$
\begin{aligned}
Sh_2 &= \frac{1}{3}\left((\tilde{c}(2)-\tilde{c}(\emptyset))+\frac{1}{2}\left(\tilde{c}(1,2)-\tilde{c}(1)\right)+\frac{1}{2}\left(\tilde{c}(2,3)-\tilde{c}(3)\right)+\left(\tilde{c}(1,2,3)-\tilde{c}(1,3)\right)\right)\\
&= \frac{1}{3}\left(S_2^{full}+\frac{1}{2}\left(S_{1,2}^{full}-S_1^{full}\right)+\frac{1}{2}\left(S_{2,3}^{full}-S_3^{full}\right)+\left(S_{1,2,3}^{full}-S_{1,3}^{full}\right)\right)\\
&= \frac{1}{3}\left(S_2^{full}+\frac{1}{2}ST_{2|1}+\frac{1}{2}ST_{2|3}+ST_2^{ind}\right)\ ,
\end{aligned}
$$

$$
\begin{aligned}
Sh_3 &= \frac{1}{3}\left((\tilde{c}(3)-\tilde{c}(\emptyset))+\frac{1}{2}\left(\tilde{c}(1,3)-\tilde{c}(1)\right)+\frac{1}{2}\left(\tilde{c}(2,3)-\tilde{c}(2)\right)+\left(\tilde{c}(1,2,3)-\tilde{c}(1,2)\right)\right)\\
&= \frac{1}{3}\left(S_3^{full}+\frac{1}{2}\left(S_{1,3}^{full}-S_1^{full}\right)+\frac{1}{2}\left(S_{2,3}^{full}-S_2^{full}\right)+\left(S_{1,2,3}^{full}-S_{1,2}^{full}\right)\right)\\
&= \frac{1}{3}\left(S_3^{full}+\frac{1}{2}ST_{3|1}+\frac{1}{2}ST_{3|2}+ST_3^{ind}\right)\ .
\end{aligned}
$$

Chapter 4

Random forest estimation of conditional distribution functions and conditional quantiles

This chapter consists in the article Elie-Dit-Cosaque and Maume-Deschamps [2020] submitted for publication.

Abstract
We propose a theoretical study of two realistic estimators of conditional distribution functions and conditional quantiles using random forests. The estimation process uses the bootstrap samples generated from the original dataset when constructing the forest. Bootstrap samples are reused to define the first estimator, while the second requires only the original sample, once the forest has been built. We prove that both proposed estimators of the conditional distribution functions are consistent uniformly a.s. To the best of our knowledge, it is the first proof of consistency including the bootstrap part. We also illustrate the estimation procedures on a numerical example.

Contents

4.1 Introduction

Conditional distribution functions and conditional quantiles estimation is an important task in several domains including environment, insurance or industry. It is also an important tool for Quantile Oriented Sensitivity Analysis (QOSA), see e.g., Fort et al. [2016]; Maume-Deschamps and Niang [2018]; Browne et al. [2017]. In order to estimate conditional quantiles, various methods exist such as kernel based estimation or quantile regression [Koenker and Hallock, 2001] but they present some limitations. Indeed, the performance of kernel methods strongly depends on the bandwidth parameter selection and quickly breaks down as the number of covariates increases. On the other hand, quantile regression is not adapted in a non-gaussian setting since the true conditional quantile is not necessarily a linear combination of the input variables [Maume-Deschamps et al., 2017]. To overcome these issues, we propose to explore the Random Forest estimation of conditional quantiles [Meinshausen, 2006].

Random forest algorithms allow a flexible modeling of interactions in high dimension by building a large number of regression trees and averaging their predictions. The most famous random forest algorithm is that of Breiman [2001] whose construction is based on the seminal work of Amit and Geman [1997]; Ho [1998]; Dietterich [2000]. Breiman's random forest estimate is a combination of two essential components: Bagging and Classification And Regression Trees (CART)-split criterion [Breiman et al., 1984]. Bagging for *bootstrap-aggregating* was proposed by Breiman [1996a] in order to improve the performance of weak or unstable learners.

Random forests are also related to some local averaging algorithms such as nearest neighbors methods [Lin and Jeon, 2006; Biau and Devroye, 2010] or kernel estimators [Scornet, 2016c]. More precisely, thanks to Lin and Jeon [2006], the random forest method can be seen as an adaptive neighborhood regression procedure and therefore the prediction (estimation of the conditional mean) can be formulated as a weighted average of the observed response variables.

Based on that approach, we develop a Weighted Conditional Empirical Cumulative Distribution Function (W_C_ECDF) approximating the Conditional Cumulative Distribution Function (C_CDF). Then, α-quantile estimates are obtained by using W_C_ECDF instead of C_CDF. Meinshausen [2006] defined a W_C_ECDF with weights using the original dataset whereas we allow to construct the weights using the bootstrap samples, as it is done practically in regression random forests. We prove the almost sure consistency of these estimators. Both estimators have several advantages over methods such as kernel methods [Nadaraya, 1964; Watson, 1964]. Due to the intrinsic tree building process, random forest estimators can easily handle both univariate and multivariate data with few parameters to tune. Besides, these methods have good predictive power and can outperform standard kernel methods [Davies and Ghahramani, 2014; Scornet, 2016c]. Lastly, being based on the random forest algorithm, they are also easily parallelizable and can handle large dataset. A implementation of both algorithms is made available within a `Julia` package called `ConditionalDistributionForest` [Fabrège and Maume-Deschamps, 2020] as well as

a `python` package named `qosa-indices`, [Elie-Dit-Cosaque, 2020].

The C_CDF can be seen as a regression function. On this basis, we reviewed the literature dealing with the consistency of random forest estimates in order to show the convergence of ours estimators.
Several authors such as Breiman [2004]; Biau [2012]; Wager and Walther [2015]; Scornet et al. [2015b]; Mentch and Hooker [2016]; Wager and Athey [2018]; Goehry [2019] have established asymptotic properties of particular variants and simplifications of the original Breiman's random forest algorithm. Facing some theoretical issues with the bootstrap, most studies replace it by subsampling, assuming that each tree is grown with $s_n < n$ observations randomly chosen without replacement from the original dataset. Most of the time, in order to ensure the convergence of the simplified model, the subsampling rate s_n/n is assumed to tend to zero at some prescribed rate, assumption that excludes the bootstrap mode. Besides, the consistency is generally showed by assuming that the number of trees goes to infinity which is not fully relevant in practice. Under some conditions, Scornet [2016a] showed that if the infinite random forest regression estimator is $\mathbb{L}^2$ consistent then so does the finite random forest regression estimator when the number of trees goes to infinity in a controlled way.
Recent attempts to bridge the gap between theory and practice, provide some results on random forest algorithms at the price of fairly strong conditions. For example, Scornet et al. [2015b] showed the $\mathbb{L}^2$ consistency of random forests in an additive regression framework by replacing the bootstrap step by subsampling. Their result rests on a fundamental lemma developped in Scornet et al. [2015a] which reviews theoretical random forest. Highlighted by a counterexample developed in Section 4.4, assumptions are required to get out of the additive framework. Furthermore, consistency and asymptoctic normality of the whole algorithm were recently proved under strong conditions by Wager and Athey [2018] replacing bootstrap by subsampling and simplifying the splitting step. One of the strong conditions used in the Theorem 3.1. of Wager and Athey [2018] is that the individual trees satisfy a condition called *honesty*. An example of an honest tree given by the authors is *one where the tree is grown using one subsample, while the predictions at the leaves of the tree are estimated using a different subsample.* Due to this assumption, the authors admit that their theorems are not valid for the practical applications most of the time because almost all implementations of random forests use the training sample twice.

Thus, despite an active investigation during the last decade, the consistency of the original (i.e. with the bootstrap samples) Breiman's random forest method is not fully proved. This motivated our work.
Our major contribution is the proof of the almost everywhere uniform convergence of the estimator W_C_ECDF both using the bootstrap samples (Theorem 4.4.1) or the original one (Theorem 4.4.2). To the best of our knowledge, this is the first consistency result under realistic assumptions for a method based on bootstrap samples in the random forest field. Notice that Meinshausen [2006] gave a proof of the consistency in probability of the W_C_ECDF for a simplified model where the weights are considered as constant while they are indeed random variables heavily data-dependent.

The paper is organized as follows. Breiman's random forest algorithm is detailed in Section 4.2 and notations are stated. The random forest estimations of C_CDF based both on bootstrap samples and the original dataset are introduced in Section 4.3 as a natural generalization of regression random forests. The main consistency results are presented in Section 4.4 and the proofs of those are gathered in Section 4.5. Section 4.6 is devoted to a short simulation study and a conclusion is given in Section 5.7.

4.2 Breiman's random forest

The aim of this section is to present the Breiman's random forest algorithm as well as notations used throughout this paper.

Random forest is a generic term to name an aggregation scheme of decision trees allowing to deal with both supervised classification and regression tasks. We are only concerned with the regression task.
The general framework is the nonparametric regression estimation where an input random vector $\mathbf{X} \in \mathcal{X} \subset \mathbb{R}^d$ is observed and a response $Y \in \mathbb{R}$ is predicted by estimating the regression function $m(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$. We assume that we are given a training sample $\mathcal{D}_n = (\mathbf{X}^j, Y^j)_{j=1,\dots,n}$ of independent random variables distributed as the prototype pair $(\mathbf{X}, Y)$ which is a $(d+1)$-dimensional random vector. The purpose is to use the dataset $\mathcal{D}_n$ to construct an estimator $m_n : \mathcal{X} \mapsto \mathbb{R}$ of the function m.

Random forests proposed by Breiman [2001] build a predictor consisting of a collection of k randomized regression trees grown based on the CART algorithm.
The CART-split criterion of Breiman et al. [1984] is used in the construction of the individual trees to recursively partition the input space $\mathcal{X}$ in a dyadic manner. More precisely, at each step of the partitioning, a part of the space is divided into two sub-parts according to the best cut perpendicular to the axes. This best cut is selected in each node of the tree by optimizing the CART-split criterion over the d variables, i.e. minimizing the prediction squared error in the two child nodes. The trees are thus grown until reaching a stopping rule. There are several rules, but one generally proposed is that the tree construction continues while leaves contain at least $min\_samples\_leaf$ elements. This criterion is implemented in the `RandomForestRegressor` class of the `python` package `Scikit-Learn` [Pedregosa et al., 2011] or in the `build_forest` function of the `Julia` [Bezanson et al., 2017] package `DecisionTree`.

Building several different trees from a single dataset requires to randomize the tree building process. Breiman [2001] proposed to inject some randomness both in the dataset and in the tree construction. First of all, prior to the construction of each tree, a resampling step is done by bootstrapping [Efron, 1979] from the original dataset, that is, by choosing uniformly at random n times from n observations with replacement. Only these bootstrap observations are taken into account in the tree building. Accordingly, the $min\_samples\_leaf$ hyperparameter introduced previously refers, in the random forest method, to the minimum number of bootstrap observations

contained in each leaf of a tree. Secondly, at each step of the tree construction, instead of optimizing the CART-split criterion over the d variables, a number of variables called $max\_features$ is selected uniformly at random among the d variables. Then, the best split is chosen as the one optimizing the CART-split criterion only along the $max\_features$ preselected variables in each node.

For any query point $\mathbf{x} \in \mathcal{X}$, the ℓ-th tree estimates $m(\mathbf{x})$ as follows

$$m_n^b\left(\mathbf{x}; \Theta_\ell, \mathcal{D}_n\right) = \sum_{j \in \mathcal{D}_n^\star(\Theta_\ell)} \frac{\mathbb{1}_{\{\mathbf{X}^j \in A_n(\mathbf{x}; \Theta_\ell, \mathcal{D}_n)\}}}{N_n^b\left(\mathbf{x}; \Theta_\ell, \mathcal{D}_n\right)} Y^j \ , \tag{4.1}$$

where:

- $\Theta_\ell, \ell = 1, \ldots, k$ are independent random vectors, distributed as a generic random vector $\Theta = (\Theta^1, \Theta^2)$ and independent of $\mathcal{D}_n$. Θ^1 contains indexes of observations that are used to build each tree, i.e. the bootstrap sample and Θ^2 indexes of splitting candidate variables in each node,
- $\mathcal{D}_n^\star\left(\Theta_\ell\right)$ is the bootstrap sample selected prior to the tree construction,
- $A_n\left(\mathbf{x}; \Theta_\ell, \mathcal{D}_n\right)$ is the tree cell (subspace of $\mathcal{X}$) containing $\mathbf{x}$,
- $N_n^b\left(\mathbf{x}; \Theta_\ell, \mathcal{D}_n\right)$ is the number of elements of $\mathcal{D}_n^\star\left(\Theta_\ell\right)$ that fall into $A_n\left(\mathbf{x}; \Theta_\ell, \mathcal{D}_n\right)$.

The trees are then combined to form the finite forest estimator

$$m_{k,n}^b\left(\mathbf{x}; \Theta_1, \ldots, \Theta_k, \mathcal{D}_n\right) = \frac{1}{k} \sum_{\ell=1}^{k} m_n^b\left(\mathbf{x}; \Theta_\ell, \mathcal{D}_n\right) \ . \tag{4.2}$$

We may now present the conditional distribution function estimators.

4.3 Conditional Distribution Forests

We aim to estimate $F(y|\mathbf{x}) = \mathbb{P}(Y \leqslant y|\mathbf{X} = \mathbf{x})$. Two estimators may be defined. One uses the bootstrap samples both in the forest construction and in the estimation. The other uses the original sample in the estimation part. Once the distribution function has been estimated, the conditional quantiles may be estimated straightforwardly.

4.3.1 Bootstrap samples based estimator

First of all, let us define the random variable $B_j\left(\Theta_\ell^1, \mathcal{D}_n\right)$ as the number of times that the observation $(\mathbf{X}^j, Y^j)$ has been drawn from the original dataset for the ℓ-th tree construction. Thanks to it, the conditional mean estimator in Equation (4.2) may be

rewritten as

$$m_{k,n}^{b}\left(\mathbf{x};\Theta_1,\ldots,\Theta_k,\mathcal{D}_n\right)=\sum_{j=1}^{n}\left(\frac{1}{k}\sum_{\ell=1}^{k}\frac{B_j\left(\Theta_\ell^1,\mathcal{D}_n\right)\mathbb{1}_{\{\mathbf{X}^j\in A_n(\mathbf{x};\Theta_\ell,\mathcal{D}_n)\}}}{N_n^b\left(\mathbf{x};\Theta_\ell,\mathcal{D}_n\right)}\right)Y^j$$
$$=\sum_{j=1}^{n}w_{n,j}^{b}\left(\mathbf{x};\Theta_1,\ldots,\Theta_k,\mathcal{D}_n\right)Y^j\ , \tag{4.3}$$

where the weights are defined by

$$w_{n,j}^{b}\left(\mathbf{x};\Theta_1,\ldots,\Theta_k,\mathcal{D}_n\right)=\frac{1}{k}\sum_{\ell=1}^{k}\frac{B_j\left(\Theta_\ell^1,\mathcal{D}_n\right)\mathbb{1}_{\{\mathbf{X}^j\in A_n(\mathbf{x};\Theta_\ell,\mathcal{D}_n)\}}}{N_n^b\left(\mathbf{x};\Theta_\ell,\mathcal{D}_n\right)}\ . \tag{4.4}$$

Note that the weights $w_{n,j}^{b}\left(\mathbf{x};\Theta_1,\ldots,\Theta_k,\mathcal{D}_n\right)$ are nonnegative random variables as functions of $\Theta_1,\ldots,\Theta_k,\mathcal{D}_n$ and their sum for $j=1,\ldots,n$ equals 1.

The random forest estimator (4.3) can be seen as a local averaging estimate. Indeed, as mentioned by Scornet [2016b], the regression trees make an average of the observations located in a neighborhood of $\mathbf{x}$, this neighborhood being defined as the leaf of the tree containing $\mathbf{x}$. The forest, which aggregates several trees, also operates by calculating a weighted average of the observations in a neighborhood of $\mathbf{x}$. However, in the case of forests, this neighborhood results from the superposition of the neighborhoods of each tree, and therefore has a more complex shape. Several works have tried to study the random forest algorithm from this point of view (local averaging estimate) such as Lin and Jeon [2006] who was the first to point out the connection between the random forest and the adaptive nearest-neighbors methods, further developped by Biau and Devroye [2010]. Some works such as Scornet [2016c] have also studied random forests through their link with the kernel methods.

We are interested in the Conditional Cumulative Distribution Function (C_CDF) of Y given $\mathbf{X}=\mathbf{x}$ in order to obtain the conditional quantiles. Pairing the following equality

$$F\left(y|\,\mathbf{X}=\mathbf{x}\right)=\mathbb{P}\left(Y\leqslant y|\,\mathbf{X}=\mathbf{x}\right)=\mathbb{E}\left[\mathbb{1}_{\{Y\leqslant y\}}\middle|\,\mathbf{X}=\mathbf{x}\right]\ , \tag{4.5}$$

with the weighted approach described above, we propose to estimate the C_CDF as follows

$$F_{k,n}^{b}\left(y|\,\mathbf{X}=\mathbf{x};\Theta_1,\ldots,\Theta_k,\mathcal{D}_n\right)=\sum_{j=1}^{n}w_{n,j}^{b}\left(\mathbf{x};\Theta_1,\ldots,\Theta_k,\mathcal{D}_n\right)\mathbb{1}_{\{Y^j\leqslant y\}}\ . \tag{4.6}$$

Hence, given a level $\alpha\in\,]0,1[$, the conditional quantile estimator $\widehat{q}^{\,\alpha}\left(Y|\,\mathbf{X}=\mathbf{x}\right)$ is defined as follows

$$\widehat{q}^{\,\alpha}\left(Y|\,\mathbf{X}=\mathbf{x}\right)=\inf\left\{Y^p,p=1,\ldots,n:F_{k,n}^{b}\left(Y^p|\,\mathbf{X}=\mathbf{x};\Theta_1,\ldots,\Theta_k,\mathcal{D}_n\right)\geqslant\alpha\right\}\ .$$

Let us turn now to the estimator using the original sample.

4.3.2 Original sample based estimator

Trees are still grown with their respective bootstrap sample $\mathcal{D}_n^\star(\Theta_\ell), \ell = 1, \ldots, k$. But instead of considering them in the estimation, we may use the original sample $\mathcal{D}_n$. Consider the weights

$$w_{n,j}^o(\mathbf{x}; \Theta_1, \ldots, \Theta_k, \mathcal{D}_n) = \frac{1}{k} \sum_{\ell=1}^{k} \frac{\mathbb{1}_{\{\mathbf{X}^j \in A_n(\mathbf{x}; \Theta_\ell, \mathcal{D}_n)\}}}{N_n^o(\mathbf{x}; \Theta_\ell, \mathcal{D}_n)} , \tag{4.7}$$

where $N_n^o(\mathbf{x}; \Theta_\ell, \mathcal{D}_n)$ is the number of points of $\mathcal{D}_n$ that fall into $A_n(\mathbf{x}; \Theta_\ell, \mathcal{D}_n)$. As previously, the weights $w_{n,j}^o(\mathbf{x}; \Theta_1, \ldots, \Theta_k, \mathcal{D}_n)$ are nonnegative random variables as functions of $\Theta_1, \ldots, \Theta_k, \mathcal{D}_n$ and their sum over $j = 1, \ldots, n$ equals 1.

It was proposed in Meinshausen [2006] to estimate the C_CDF with

$$F_{k,n}^o(y | \mathbf{X} = \mathbf{x}; \Theta_1, \ldots, \Theta_k, \mathcal{D}_n) = \sum_{j=1}^{n} w_{n,j}^o(\mathbf{x}; \Theta_1, \ldots, \Theta_k, \mathcal{D}_n) \mathbb{1}_{\{Y^j \leqslant y\}} . \tag{4.8}$$

The conditional quantiles are then estimated by plugging $F_{k,n}^o(y | \mathbf{X} = \mathbf{x}; \Theta_1, \ldots, \Theta_k, \mathcal{D}_n)$ instead of $F(y | \mathbf{X} = \mathbf{x})$ as before.

A complete description of the procedure for computing conditional quantile estimates via the C_CDF with both previous estimators can be found in Algorithm 4. A `python` library named `qosa-indices` has also been developed to perform the numerical estimations of conditional distributions and quantiles for both methods. It is available at Elie-Dit-Cosaque [2020] and uses `Scikit-Learn, Numpy, Numba`. Both approaches are also implemented in a `Julia` package based on the library `DecisionTree` and that is available at Fabrège and Maume-Deschamps [2020].

It has to be noted that a package called `quantregForest` has been made available in `R` [R Core Team, 2019] and can be found at Meinshausen [2019]. The estimation method currently implemented in `quantregForest` is different from the method described in Meinshausen [2006]. As a matter of fact, for a new observation $\mathbf{x}$ and the ℓ-th tree, one element of $\mathcal{D}_n = (\mathbf{X}^j, Y^j)_{j=1,\ldots,n}$ falling into in the leaf node $A_n(\mathbf{x}; \Theta_\ell, \mathcal{D}_n)$ is chosen at random. This gives, k values of Y and allows to estimate the conditional distribution function with the classical Empirical Cumulative Distribution Function associated with the empirical measure. The performance of this method seems weak and no theoretical guarantees are available.

4.4 Consistency results

In this section, we state our main results, which are the uniform a.s. consistency of both estimators $F_{k,n}^b$ and $F_{k,n}^o$ of the conditional distribution function. It constitutes the most interesting result of this paper because it handles the bootstrap component and

Algorithme 4 : Conditional Distribution Forest algorithm

Input :

- Training sample: $\mathcal{D}_n$
- Number of trees: $k \in \mathbb{N}^\star$
- Number of features to be considered for the best split in a node: $max\_features \in \{1, \dots, d\}$
- Minimum number of samples required in a leaf node: $min\_samples\_leaf \in \{1, \dots, n\}$
- Point where the conditional distribution function or the conditional quantile is required: $\mathbf{x} \in \mathcal{X}$
- The order of the conditional quantile: $\alpha \in [0, 1]$

Output : Estimated value of the conditional quantile of $\mathbf{x}$ at the α-order.

1 **for** $\ell = 1, \dots, k$ **do**
2 Select uniformly with replacement n data points among $\mathcal{D}_n$. Only these observations will be used in the tree construction.
3 **begin** Tree construction
4 Consider the whole space $\mathcal{X}$ as root node.
5 Select uniformly without replacement $max\_features$ coordinates among $\{1, \dots, d\}$.
6 Select the split minimizing the CART-split criterion [Breiman et al., 1984; Biau and Scornet, 2016] along the pre-selected $max\_features$ directions.
7 Cut the current node at the selected split in two child nodes.
8 Repeat the lines (5)-(7) for the two resulting nodes until each node of the tree contains at least $min\_samples\_leaf$ observations.
9 **end**
10 Save in which leaf node of the tree fall each observation of the training sample $\mathcal{D}_n$.
11 **end**
12 Drop $\mathbf{x}$ through all trees and calculate for each observation in $\mathcal{D}_n$ its weigthed average through the forest as in Equation (4.4) or (4.7) according to the estimator used.
13 Sort the calculated weights according to $\left(Y^{(j)}\right)_{j=1,\dots,n}$ that are the order statistics of $\left(Y^j\right)_{j=1,\dots,n}$.
14 Compute the cumulative sum of the sorted weights which gives us a Weighted Conditional Empirical Cumulative Distribution Function (W_C_ECDF).
15 Get the α-conditional quantile of $\mathbf{x}$ thanks to the W_C_ECDF.

gives the almost sure uniform convergence. Indeed, most of the studies [Scornet et al., 2015b; Wager and Athey, 2018; Goehry, 2019] replaces the bootstrap by subsampling without replacement in order to avoid the mathematical difficulties induced by this one and therefore differ slightly from the procedure used in practice.

Meinshausen [2006] showed the uniform convergence in probability of a simplified version of the estimator $F^o_{k,n}$. In Meinshausen [2006], the weights $w^o_{n,j}(\mathbf{x}; \Theta_1, \ldots, \Theta_k, \mathcal{D}_n)$ are in fact considered to be non-random while they are indeed random variables depending on $(\Theta_\ell)_{\ell=1,\ldots,k}$ and $\mathcal{D}_n$.

Overall, proving the consistency of the forest methods whose construction depends both on the $\mathbf{X}^j$'s and on the Y^j's is a difficult task. This feature makes the resulting estimate highly data-dependent, and therefore difficult to analyze. A simplification widely used by most authors from a theoretical point of view is to work with random forest estimates whose form of the tree depends only on $\mathbf{X}^j$'s which Devroye et al. [2013] called the $\mathbf{X}$-property but the Y^j's are still used to compute the prediction, either the conditional mean or the conditional distribution function for example. One of the first results dealing with data-dependent random forest estimator of the regression function is Scornet et al. [2015b] who showed the $\mathbb{L}^2$ consistency in an additive regression framework by replacing the bootstrap by subsampling. Thanks to the following assumptions, we go further by showing the consistency of our estimators in a general framework and not only in the additive regression scheme.

Assumption 4.4.1.
For all $\ell \in [\![1, k]\!]$, we assume that the variation of the conditional cumulative distribution function within any cell goes to 0:

$$\forall \mathbf{x} \in \mathcal{X}, \forall y \in \mathbb{R}, \quad \sup_{\mathbf{z} \in A_n(\mathbf{x}; \Theta_\ell, \mathcal{D}_n)} |F(y|\,\mathbf{z}) - F(y|\,\mathbf{x})| \xrightarrow[n\to\infty]{a.s.} 0 \ .$$

We shall discuss further on Assumption 4.4.1 but let us remark that it is satisfied, for example, provided that the diameter of each tree cell goes to zero and for all y, $F(y|\cdot)$ is continuous.

Assumption 4.4.2.
We shall make the following assumptions on k (number of trees) and $N^b_n(\mathbf{x}; \Theta, \mathcal{D}_n)$ (number of bootstrap observations in a leaf node):

1. $k = \mathcal{O}(n^\alpha)$, with $\alpha > 0$.
2. $\forall \mathbf{x} \in \mathcal{X}, \quad N^b_n(\mathbf{x}; \Theta, \mathcal{D}_n) = \Omega^1\left(\sqrt{n}\,(\ln(n))^\beta\right)$, with $\beta > 1$, *a.s.*

   or

3. $\forall \mathbf{x} \in \mathcal{X}, \quad \mathbb{E}\left[N^b_n(\mathbf{x}; \Theta, \mathcal{D}_n)\right] = \Omega\left(\sqrt{n}\,(\ln(n))^\beta\right)$, with $\beta > 1$, and

[1] $f(n) = \Omega(g(n)) \iff \exists k > 0, \exists n_0 > 0 \mid \forall n \geqslant n_0 \quad |f(n)| \geqslant k \cdot |g(n)|$

$$\forall \mathbf{x} \in \mathcal{X}, \quad \mathrm{CV}^2\left(N_n^b\left(\mathbf{x}; \Theta, \mathcal{D}_n\right)\right) = \mathcal{O}\left(\frac{1}{n^{(\alpha+1)/2}\left(\ln\left(n\right)\right)^{\gamma/2}}\right), \text{ with } \gamma > 1.$$

Remark 4.4.1.
In order to prove our main consistency result, either Assumption 4.4.2 item 2. or item 3. is needed. Item 2. may seem much stronger than item 3. but it has to be noted that the number of bootstrap observations in a tree leaf is a construction parameter of the forest, so that it can be controlled. Using item 2. simplifies the proof but item 3. is sufficient.

Assumption 4.4.3.
For every $\mathbf{x} \in \mathcal{X}$, the conditional cumulative distribution function $F\left(y|\, \mathbf{X} = \mathbf{x}\right)$ is continuous and strictly increasing in y.

The two theorems below give the uniform a.s. consistency of our two estimators.

Theorem 4.4.1.
Consider a random forest which satisfies Assumptions 4.4.1 to 4.4.3. Then,

$$\forall \mathbf{x} \in \mathcal{X}, \quad \sup_{y \in \mathbb{R}} \left| F_{k,n}^b\left(y|\, \mathbf{X} = \mathbf{x}\right) - F\left(y|\, \mathbf{X} = \mathbf{x}\right)\right| \xrightarrow[n\to\infty]{a.s.} 0 \ .$$

Theorem 4.4.2.
Consider a random forest which satisfies Assumptions 4.4.1 to 4.4.3. Then,

$$\forall \mathbf{x} \in \mathcal{X}, \quad \sup_{y \in \mathbb{R}} \left| F_{k,n}^o\left(y|\, \mathbf{X} = \mathbf{x}\right) - F\left(y|\, \mathbf{X} = \mathbf{x}\right)\right| \xrightarrow[n\to\infty]{a.s.} 0 \ .$$

Remark 4.4.2.
Using standard arguments, the consistency of quantile estimates stems from Assumption 4.4.3 as well as the uniform convergence of the conditional distribution function estimators obtained above.

Let us make some comments on the assumptions above.

Assumption 4.4.1 ensures a control on the approximation error of the estimators. It is drawn from the Proposition 2 of Scornet et al. [2015b] who shows the consistency of Breiman's random forest estimate in an additive regression framework. Their Proposition 2 allows to manage the approximation error of the estimator by showing that the variation of the regression function m within a cell of a random empirical tree is small provided n is large enough. This result is based on the fundamental Lemma 1 of Scornet et al. [2015b] which states that the variation of the regression function m within a cell of a random theoretical tree goes to zero for an additive regression model. A random theoretical tree is grown as a random empirical tree, except that the theoretical equivalent of the empirical CART-split criterion [Biau and Scornet, 2016]

[2] $\mathrm{CV}\left(X\right) = \sigma_X / \mathbb{E}\left[X\right]$

defined in any node A below is used to choose the best split

$$\begin{aligned} L_A^\star(i,z) =& \operatorname{Var}(Y|\,\mathbf{X}\in A) \\ & - \mathbb{P}(X_i < z|\,\mathbf{X}\in A)\operatorname{Var}(Y|\,X_i < z, \mathbf{X}\in A) \\ & - \mathbb{P}(X_i \geqslant z|\,\mathbf{X}\in A)\operatorname{Var}(Y|\,X_i \geqslant z, \mathbf{X}\in A)\ . \end{aligned}$$

Hence, a theoretical tree is obtained thanks to the best consecutive cuts $(i^\star, z^\star)$ optimizing the previous criterion $L^\star(\cdot,\cdot)$.

General results on standard partitioning estimators whose construction is independent of the label in the training set (see Chapter 4 in Györfi et al. [2006] or Chapter 6 in Devroye et al. [2013]) state that a necessary condition to prove the consistency is that the diameter of the cells tend to zero as $n \to \infty$. Instead of such a geometrical assumption, Proposition 2 in Scornet et al. [2015b] ensures that the variation of m inside a node is small thanks to their Lemma 1. But the cornerstone of the Lemma 1 is the Technical Lemma 1 of Scornet et al. [2015a] recalled below for completness.

Technical Lemma.
Assume that:

- $Y = m(\mathbf{X}) + \varepsilon$ *with* $m(\mathbf{X}) = \sum_{i=1}^{d} m_i(X_i), \mathbf{X} \sim \mathcal{U}\left([0,1]^d\right)$ *and* $\varepsilon \sim \mathcal{N}(0,\sigma^2)$,
- $L_A^\star(i,z) = 0 \quad \forall i, \forall z \in [a_i, b_i] \ \ (0 \leqslant a_i < b_i \leqslant 1)$,

then the regression function m is constant on A.

This lemma states that if the theoretical split criterion is zero for all cuts in a node, then the regression function m is constant on this node, i.e. the variation of this one within the cell is zero. But, we will see in the sequel a counterexample where $L_A^\star(i,z) = 0 \quad \forall i, \forall z \in [a_i, b_i]$ and yet, the regression function is not constant.

First of all, let us look under which conditions $L_A^\star(i,z) = 0 \quad \forall i, \forall z \in [a_i, b_i]$ in a node A. Remark that

$$\begin{aligned} \operatorname{Var}(Y|\,\mathbf{X}\in A) =& \mathbb{P}(X_i < z|\,\mathbf{X}\in A)\operatorname{Var}(Y|\,X_i < z, \mathbf{X}\in A) \\ & + \mathbb{P}(X_i \geqslant z|\,\mathbf{X}\in A)\operatorname{Var}(Y|\,X_i \geqslant z, \mathbf{X}\in A) \\ & + \mathbb{P}(X_i < z|\,\mathbf{X}\in A)\left(\mathbb{E}[Y|\,X_i < z, \mathbf{X}\in A] - \mathbb{E}[Y|\,\mathbf{X}\in A]\right)^2 \\ & + \mathbb{P}(X_i \geqslant z|\,\mathbf{X}\in A)\left(\mathbb{E}[Y|\,X_i \geqslant z, \mathbf{X}\in A] - \mathbb{E}[Y|\,\mathbf{X}\in A]\right)^2\ . \end{aligned}$$

Thus, $L_A^\star(i,z) = 0 \ \ \forall i, \forall z \in [a_i, b_i]$ if and only if $\mathbb{E}[Y|\,\mathbf{X}\in A] = \mathbb{E}[Y|\,X_i < z, \mathbf{X}\in A] = \mathbb{E}[Y|\,X_i \geqslant z, \mathbf{X}\in A] \quad \forall i, \forall z \in [a_i, b_i]$.

Within a standard cell of the form $A = \prod_{i=1}^{d} A_i = \prod_{i=1}^{d} [a_i, b_i]$ as well as for a generic model of the type $Y = m(\mathbf{X}) + \varepsilon$ with $\mathbf{X}$, independent random inputs and ε, an independent centered noise of $\mathbf{X}$, the condition above is equivalent to

$$\mathbb{E}\left[m(\mathbf{X})\,\mathbb{1}_{\{\mathbf{X}\in A\}}\right] = \mathbb{P}(X_i \in A_i)\,\mathbb{E}\left[m(\mathbf{X}_{-i}, z)\,\mathbb{1}_{\{\mathbf{X}_{-i}\in A_{-i}\}}\right] \quad \forall i, \forall z \in [a_i, b_i]\ .$$

This result is obtained by deriving with respect to z the following function

$$\Phi(z) = \mathbb{P}\left(X_i < z, \mathbf{X} \in A\right) \mathbb{E}\left[m\left(\mathbf{X}\right) \mathbb{1}_{\{\mathbf{X} \in A\}}\right] - \mathbb{P}\left(\mathbf{X} \in A\right) \mathbb{E}\left[m\left(\mathbf{X}\right) \mathbb{1}_{\{X_i < z, \mathbf{X} \in A\}}\right] \ .$$

Let us consider a two-dimensional example, let $A = A_1 \times A_2 = [a_1, b_1] \times [a_2, b_2]$ and suppose that the response Y is

$$Y = X_1 X_2 + c_1 X_1 + c_2 X_2 + \varepsilon \ ,$$

with

- $\mathbf{X} = (X_1, X_2)$ independent random inputs,
- $c_1 = -\dfrac{\mathbb{E}\left[X_2 \mathbb{1}_{\{X_2 \in A_2\}}\right]}{\mathbb{P}\left(X_2 \in A_2\right)}$ and $c_2 = -\dfrac{\mathbb{E}\left[X_1 \mathbb{1}_{\{X_1 \in A_1\}}\right]}{\mathbb{P}\left(X_1 \in A_1\right)}$,
- and ε a centered noise independent of $\mathbf{X}$.

It can be shown for this model that within the node A, $L^\star \equiv 0$ for all $i \in \{1, 2\}$, for all $z \in [a_i, b_i]$ and yet the regression function m is not constant.

Accordingly, the technical lemma above is well-designed for an additive regression framework. But this context is far from reality for many concrete examples. Outside this framework and as highlighted by our counterexample, an additional assumption is necessary in order to control the approximation error of the estimator. Theorem 1 in Meinshausen [2006] handles the approximation error of its estimator based on a simplified random forest model thanks to a restrictive assumption on the proportion of observations selected in each split and for each direction (see Assumption 3 in Meinshausen [2006]) and a Lipschitz assumption on the conditional distribution function, which is not always true as mentionned in Biau and Scornet [2016]. We use Assumption 4.4.1 instead.

On the other hand, Assumption 4.4.2 allows us to control the estimation error of our estimators and expresses that cells should contain a sufficiently large number of points so that averaging among the observations is effective.

Finally, Assumption 4.4.3 is necessary to get uniform convergence of the estimators.

Hence, thanks to all these suitable assumptions we get the consistency of our estimators in Theorems 4.4.1 and 4.4.2. As far as we know, Theorem 4.4.1 is the first consistency result for a method based on the original Breiman's random forest algorithm (i.e. using the bootstrap samples).

The next section is devoted to the proofs of the two Theorems 4.4.1 and 4.4.2.

4.5 Proofs of the main theorems

The proofs of Theorems 4.4.1 and 4.4.2 are close. We begin with that of Theorem 4.4.1 and then sketch that of Theorem 4.4.2 which is a bit simpler.

4.5.1 Proof of Theorem 4.4.1

The main ingredient of the proof is to use a second sample $\mathcal{D}_n^\diamond$ in order to deal with the data-dependent aspect. Thus, we first define a dummy estimator based on two samples $\mathcal{D}_n$ and $\mathcal{D}_n^\diamond$ which will be used below. The trees are grown as in Algorithm 4 using $\mathcal{D}_n$, but we consider another sample $\mathcal{D}_n^\diamond$ (independent of $\mathcal{D}_n$ and Θ) which is used to define a dummy estimator

$$F_{k,n}^\diamond\left(y|\,\mathbf{X}=\mathbf{x};\Theta_1,\ldots,\Theta_k,\mathcal{D}_n^\diamond,\mathcal{D}_n\right)=\sum_{j=1}^n w_{n,j}^\diamond\left(\mathbf{x};\Theta_1,\ldots,\Theta_k,\mathbf{X}^{\diamond 1},\ldots,\mathbf{X}^{\diamond n},\mathcal{D}_n\right)\mathbb{1}_{\{Y^{\diamond j}\leqslant y\}}\,, \tag{4.9}$$

where the weights are

$$w_{n,j}^\diamond\left(\mathbf{x};\Theta_1,\ldots,\Theta_k,\mathbf{X}^{\diamond 1},\ldots,\mathbf{X}^{\diamond n},\mathcal{D}_n\right)=\frac{1}{k}\sum_{\ell=1}^k\frac{\mathbb{1}_{\{\mathbf{X}^{\diamond j}\in A_n(\mathbf{x};\Theta_\ell,\mathcal{D}_n)\}}}{N_n^\diamond\left(\mathbf{x};\Theta_\ell,\mathbf{X}^{\diamond 1},\ldots,\mathbf{X}^{\diamond n},\mathcal{D}_n\right)},\ j=1,\ldots,n,$$

with $N_n^\diamond\left(\mathbf{x};\Theta_\ell,\mathbf{X}^{\diamond 1},\ldots,\mathbf{X}^{\diamond n},\mathcal{D}_n\right)$, the number of elements of $\mathcal{D}_n^\diamond$ that fall into $A_n\left(\mathbf{x};\Theta_\ell,\mathcal{D}_n\right)$. Throughout this section, we shall use the convention $\frac{0}{0}=0$ in case $N_n^\diamond\left(\mathbf{x};\Theta_\ell,\mathbf{X}^{\diamond 1},\ldots,\mathbf{X}^{\diamond n},\mathcal{D}_n\right)=0$ and thus $\mathbb{1}_{\{\mathbf{X}^{\diamond j}\in A_n(\mathbf{x};\Theta_\ell,\mathcal{D}_n)\}}=0$ for $j=1,\ldots n$.
The weights $w_{n,j}^\diamond\left(\mathbf{x};\Theta_1,\ldots,\Theta_k,\mathbf{X}^{\diamond 1},\ldots,\mathbf{X}^{\diamond n},\mathcal{D}_n\right)$ are nonnegative random variables, as function of $\Theta_1,\ldots,\Theta_k,\mathbf{X}^{\diamond 1},\ldots,\mathbf{X}^{\diamond n},\mathcal{D}_n$. To lighten the notation in the sequel, we will simply write $F_{k,n}^\diamond\left(y|\,\mathbf{X}=\mathbf{x}\right)=\sum_{j=1}^n w_j^\diamond\left(\mathbf{x}\right)\mathbb{1}_{\{Y^{\diamond j}\leqslant y\}}$ instead of (4.9).

Let $\mathbf{x}\in\mathcal{X}$ and $y\in\mathbb{R}$, we have

$$\begin{aligned}\left|F_{k,n}^b\left(y|\,\mathbf{X}=\mathbf{x}\right)-F\left(y|\,\mathbf{X}=\mathbf{x}\right)\right|\leqslant&\left|F_{k,n}^\diamond\left(y|\,\mathbf{X}=\mathbf{x}\right)-F\left(y|\,\mathbf{X}=\mathbf{x}\right)\right|\\&+\left|F_{k,n}^\diamond\left(y|\,\mathbf{X}=\mathbf{x}\right)-F_{k,n}^b\left(y|\,\mathbf{X}=\mathbf{x}\right)\right|\ .\end{aligned}$$

The convergence of the two right-hand terms is handled separately into the following Proposition 4.5.1 and Lemma 4.5.2.

Proposition 4.5.1.
Consider a random forest which satisfies Assumptions 4.4.1 and 4.4.2. Then,

$$\forall\mathbf{x}\in\mathcal{X},\forall y\in\mathbb{R},\quad F_{k,n}^\diamond\left(y|\,\mathbf{X}=\mathbf{x}\right)\xrightarrow[n\to\infty]{a.s.}F\left(y|\,\mathbf{X}=\mathbf{x}\right)\ .$$

Hence, Proposition 4.5.1 establishes the consistency for a random forest estimator based on a second sample $\mathcal{D}_n^\diamond$ independent of $\mathcal{D}_n$ and Θ. Wager and Athey [2018]

proved that estimators built from honest forests are asymptotically Gaussian. Remark that in Wager and Athey [2018], it is also required to control the proportion of chosen observations at each split and in each direction. In our case, going through a kind of honest trees is just a theoretical tool. We go one step further with the following lemma by showing that the estimators built with honest and non-honest trees are close.

Lemma 4.5.2.
Consider a random forest which satisfies Assumption 4.4.2. Then,

$$\forall \mathbf{x} \in \mathcal{X}, \forall y \in \mathbb{R}, \quad \left| F^{\diamond}_{k,n}\left(y \middle| \mathbf{X} = \mathbf{x}\right) - F^{b}_{k,n}\left(y \middle| \mathbf{X} = \mathbf{x}\right) \right| \xrightarrow[n\to\infty]{a.s.} 0 \ .$$

Hence, according to Proposition 4.5.1 and Lemma 4.5.2, we get

$$\forall \mathbf{x} \in \mathcal{X}, \forall y \in \mathbb{R}, \quad F^{b}_{k,n}\left(y \middle| \mathbf{X} = \mathbf{x}\right) \xrightarrow[n\to\infty]{a.s.} F\left(y \middle| \mathbf{X} = \mathbf{x}\right) \ . \tag{4.10}$$

Now, thanks to Dini's second theorem, let us sketch how to obtain the almost sure uniform convergence relative to y of the estimator.

Note that $\{Y^j \leqslant y\} = \left\{U_j \leqslant F_{Y|\mathbf{X}=\mathbf{x}}(y)\right\}$ under Assumption 4.4.3 with $U_j = F_{Y|\mathbf{X}=\mathbf{x}}\left(Y^j\right), j = 1, \ldots, n$ which are i.i.d random variables. Then, (4.10) is equivalent to

$$\forall \mathbf{x} \in \mathcal{X}, \forall s \in [0,1], \quad \sum_{j=1}^{n} w^{b}_{j}(\mathbf{x}) \mathbb{1}_{\{U_j \leqslant s\}} \xrightarrow[n\to\infty]{a.s.} s \ .$$

As in the proof of Glivenko–Cantelli's Theorem, using that $s \mapsto \sum_{j=1}^{n} w^{b}_{j}(\mathbf{x}) \mathbb{1}_{\{U_j(\omega) \leqslant s\}}$ is increasing and Dini's second theorem, we get the uniform convergence almost everywhere, which concludes the proof of the theorem. ■

We now turn to the proofs of Proposition 4.5.1 and Lemma 4.5.2. To that aim, the following lemma, based on Vapnik-Chervonenkis classes [Vapnik and Chervonenkis, 1971] is a key tool.

Lemma 4.5.3.
Consider $\mathcal{D}_n$ and $\mathcal{D}^{\diamond}_n$, two independent datasets of independent n samples of $(\mathbf{X}, Y)$. Build a tree using $\mathcal{D}_n$ with bootstrap and bagging procedure driven by Θ. As before, $N^b\left(A_n(\Theta)\right) = N^b_n(\mathbf{x}; \Theta, \mathcal{D}_n)$ is the number of bootstrap observations of $\mathcal{D}_n$ that fall into in $A_n(\Theta) = A_n(\mathbf{x}; \Theta, \mathcal{D}_n)$ and $N^{\diamond}\left(A_n(\Theta)\right) = N^{\diamond}_n(\mathbf{x}; \Theta, \mathbf{X}^{\diamond 1}, \ldots, \mathbf{X}^{\diamond n}, \mathcal{D}_n)$, the number of observations of $\mathcal{D}^{\diamond}_n$ that fall into in $A_n(\Theta)$. Then,

$$\forall \varepsilon > 0, \quad \mathbb{P}\left(\left| N^b\left(A_n(\Theta)\right) - N^{\diamond}\left(A_n(\Theta)\right) \right| > \varepsilon\right) \leqslant 24(n+1)^{2d} e^{-\varepsilon^2/288n} \ .$$

Proof of Lemma 4.5.3.

Let $\varepsilon > 0$ and $\mathbf{x} \in \mathcal{X}$, we have

$$
\begin{aligned}
&\mathbb{P}\left(\left|N^b\left(A_n\left(\Theta\right)\right) - N^\diamond\left(A_n\left(\Theta\right)\right)\right| > \varepsilon\right) \\
&\leqslant \mathbb{P}\left(\left|\frac{N^b\left(A_n\left(\Theta\right)\right)}{n} - \frac{1}{n}\sum_{j=1}^{n}\mathbb{1}_{\{\mathbf{X}^j \in A_n(\Theta)\}}\right| > \frac{\varepsilon}{3n}\right) + \mathbb{P}\left(\left|\frac{1}{n}\sum_{j=1}^{n}\mathbb{1}_{\{\mathbf{X}^j \in A_n(\Theta)\}} - \mathbb{P}_{\mathbf{X}}\left(\mathbf{X} \in A_n\left(\Theta\right)\right)\right| > \frac{\varepsilon}{3n}\right) \\
&\quad + \mathbb{P}\left(\left|\frac{N^\diamond\left(A_n\left(\Theta\right)\right)}{n} - \mathbb{P}_{\mathbf{X}}\left(\mathbf{X} \in A_n\left(\Theta\right)\right)\right| > \frac{\varepsilon}{3n}\right) \\
&\leqslant \mathbb{P}\left(\sup_{A\in\mathcal{B}}\left|\frac{1}{n}\sum_{j=1}^{n}B_j\left(\Theta^1, \mathcal{D}_n\right)\mathbb{1}_{\{\mathbf{X}^j \in A\}} - \frac{1}{n}\sum_{j=1}^{n}\mathbb{1}_{\{\mathbf{X}^j \in A\}}\right| > \frac{\varepsilon}{3n}\right) \\
&\quad + \mathbb{P}\left(\sup_{A\in\mathcal{B}}\left|\frac{1}{n}\sum_{j=1}^{n}\mathbb{1}_{\{\mathbf{X}^j \in A\}} - \mathbb{P}_{\mathbf{X}}\left(\mathbf{X} \in A\right)\right| > \frac{\varepsilon}{3n}\right) + \mathbb{P}\left(\sup_{A\in\mathcal{B}}\left|\frac{1}{n}\sum_{j=1}^{n}\mathbb{1}_{\{\mathbf{X}^{\diamond j} \in A\}} - \mathbb{P}_{\mathbf{X}}\left(\mathbf{X} \in A\right)\right| > \frac{\varepsilon}{3n}\right)
\end{aligned}
$$

with $\mathcal{B} = \left\{\prod_{i=1}^{d}[a_i, b_i] : a_i, b_i \in \overline{\mathbb{R}}\right\}$. The last two right-hand terms are handled thanks to a direct application of the Theorem of Vapnik and Chervonenkis [1971] over the class $\mathcal{B}$ whose Vapnik-Chervonenkis dimension is $2d$. This class is nothing more than an extension of the class $\mathcal{R}$ of rectangles in $\mathbb{R}^d$. Following the lines of the proof of Theorem 13.8 in Devroye et al. [2013], one sees that the classes $\mathcal{R}$ and $\mathcal{B}$ have the same Vapnik-Chervonenkis dimension.

A special attention should be given to the first right hand-term. The bootstrap component is represented with the random vector $\left(B_j\left(\Theta^1, \mathcal{D}_n\right)\right)_{j=1,\ldots,n}$ referring to the number of times that the observation $(\mathbf{X}^j, Y^j)$ has been chosen from the original dataset. Conditionally to $\mathcal{D}_n$, this random vector has a multinomial distribution with parameters $\mathcal{M}\left(n; 1/n, \ldots, 1/n\right)$. As stated in Arenal-Gutiérrez et al. [1996], the bootstrap component can also be represented thanks to the variables selected with replacement from the set $\mathcal{D}_n = \{(\mathbf{X}^1, Y^1), \ldots, (\mathbf{X}^n, Y^n)\}$. Let $Z^1, \ldots, Z^n$ be these elements which are distributed as $Z = (Z_1, Z_2)$ that has a discrete uniform distribution over $\mathcal{D}_n$ conditionally to $\mathcal{D}_n$. The first right hand-term is rewritten as

$$
\begin{aligned}
&\mathbb{P}\left(\sup_{A\in\mathcal{B}}\left|\frac{1}{n}\sum_{j=1}^{n}B_j\left(\Theta^1, \mathcal{D}_n\right)\mathbb{1}_{\{\mathbf{X}^j \in A\}} - \frac{1}{n}\sum_{j=1}^{n}\mathbb{1}_{\{\mathbf{X}^j \in A\}}\right| > \frac{\varepsilon}{3n}\right) \\
&= \mathbb{P}\left(\sup_{A\in\mathcal{B}}\left|\frac{1}{n}\sum_{j=1}^{n}\mathbb{1}_{\{Z_1^j \in A\}} - \frac{1}{n}\sum_{j=1}^{n}\mathbb{1}_{\{\mathbf{X}^j \in A\}}\right| > \frac{\varepsilon}{3n}\right) \\
&= \mathbb{E}\left[\mathbb{P}\left(\sup_{A\in\mathcal{B}}\left|\frac{1}{n}\sum_{j=1}^{n}\mathbb{1}_{\{Z_1^j \in A\}} - \frac{1}{n}\sum_{j=1}^{n}\mathbb{1}_{\{\mathbf{X}^j \in A\}}\right| > \frac{\varepsilon}{3n}\middle|\, \mathcal{D}_n\right)\right] \\
&= \mathbb{E}\left[\mathbb{P}\left(\sup_{A\in\mathcal{B}}\left|\frac{1}{n}\sum_{j=1}^{n}\mathbb{1}_{\{Z_1^j \in A\}} - \mathbb{P}\left(Z_1 \in A|\, \mathcal{D}_n\right)\right| > \frac{\varepsilon}{3n}\middle|\, \mathcal{D}_n\right)\right] .
\end{aligned}
$$

By applying Vapnik-Chervonenkis' Theorem under the conditional distribution given $\mathcal{D}_n$, we get

$$\mathbb{P}\left(\sup_{A\in\mathcal{B}}\left|\frac{1}{n}\sum_{j=1}^{n}\mathbb{1}_{\left\{Z_1^j\in A\right\}}-\mathbb{P}\left(Z_1\in A|\,\mathcal{D}_n\right)\right|>\frac{\varepsilon}{3n}\middle|\,\mathcal{D}_n\right)\leqslant 8(n+1)^{2d}e^{-\varepsilon^2/288n}\ .$$

Therefore,

$$\mathbb{P}\left(\sup_{A\in\mathcal{B}}\left|\frac{1}{n}\sum_{j=1}^{n}\mathbb{1}_{\left\{Z_1^j\in A\right\}}-\frac{1}{n}\sum_{j=1}^{n}\mathbb{1}_{\left\{\mathbf{X}^j\in A\right\}}\right|>\frac{\varepsilon}{3n}\right)\leqslant 8(n+1)^{2d}e^{-\varepsilon^2/288n}\ .$$

Finally, we get the overall upper bound

$$\mathbb{P}\left(\left|N^b\left(A_n\left(\Theta\right)\right)-N^\diamond\left(A_n\left(\Theta\right)\right)\right|>\varepsilon\right)\leqslant 24(n+1)^{2d}e^{-\varepsilon^2/288n}\ .$$

■

Lemma 4.5.3 is the main ingredient of the proof of Proposition 4.5.1.

Proof of Proposition 4.5.1.

We aim to prove

$$\forall\mathbf{x}\in\mathcal{X},\forall y\in\mathbb{R},\quad \mathbb{P}\left(\omega\in\Omega:\ F_{k,n}^{\diamond}\left(y|\,\mathbf{X}=\mathbf{x}\right)\underset{n\to\infty}{\longrightarrow}F\left(y|\,\mathbf{X}=\mathbf{x}\right)\right)=1\ .$$

Let $\mathbf{x}\in\mathcal{X}$ and $y\in\mathbb{R}$, we have

$$\left|F_{k,n}^{\diamond}\left(y|\,\mathbf{x}\right)-F\left(y|\,\mathbf{x}\right)\right|\leqslant\left|\sum_{j=1}^{n}w_j^{\diamond}\left(\mathbf{x}\right)\left(\mathbb{1}_{\{Y^{\diamond j}\leqslant y\}}-F\left(y|\,\mathbf{X}^{\diamond j}\right)\right)\right|+\left|\sum_{j=1}^{n}w_j^{\diamond}\left(\mathbf{x}\right)\left(F\left(y|\,\mathbf{X}^{\diamond j}\right)-F\left(y|\,\mathbf{x}\right)\right)\right|\ .$$

Define $W_n=\sum_{j=1}^{n}w_j^{\diamond}\left(\mathbf{x}\right)\left(\mathbb{1}_{\{Y^{\diamond j}\leqslant y\}}-F\left(y|\,\mathbf{X}^{\diamond j}\right)\right)=\sum_{j=1}^{n}w_j^{\diamond}\left(\mathbf{x}\right)Z_j^{\diamond}$ with $Z_j^{\diamond}=\mathbb{1}_{\{Y^{\diamond j}\leqslant y\}}-F\left(y|\,\mathbf{X}^{\diamond j}\right)$, n i.i.d random variables and $V_n=\sum_{j=1}^{n}w_j^{\diamond}\left(\mathbf{x}\right)\left(F\left(y|\,\mathbf{X}^{\diamond j}\right)-F\left(y|\,\mathbf{x}\right)\right)$. Remark that $\mathbb{E}\left[Z_j^{\diamond}\middle|\,\mathbf{X}^{\diamond j}\right]=0$.

1. We first show that $(W_n)_{n\geqslant 1}$ goes to 0 a.s. in the case of Assumption 4.4.2 item 2. This is achieved by adapting Hoeffding inequality's proof to our random weighted sum context.

For any $\varepsilon>0$, $t\in\mathbb{R}_+^*$, we have

$$\mathbb{P}\left(W_n>\varepsilon\right)\leq\mathbb{E}\left[e^{tW_n}\right]\cdot e^{-t\varepsilon}\ .$$

We shall make use of the folklore lemma below.

Lemma.
Let X be a centred random variable, a.s. bounded by 1. *Then, for any* $t \in \mathbb{R}$, $\mathbb{E}\left[e^{tX}\right] \leqslant e^{\frac{t^2}{2}}$.

Let $t > 0$, we have

$$\mathbb{E}\left[e^{tW_n}\right] = \mathbb{E}\left[\prod_{j=1}^{n} e^{tw_j^\diamond(\mathbf{x})Z_j^\diamond}\right] = \mathbb{E}\left[\mathbb{E}\left[\prod_{j=1}^{n} e^{tw_j^\diamond(\mathbf{x})Z_j^\diamond}\middle| \mathcal{D}_n, \Theta_1, \ldots, \Theta_k, \mathbf{X}^{\diamond 1}, \ldots, \mathbf{X}^{\diamond n}\right]\right]$$

conditionally to $\mathcal{D}_n, \Theta_1, \ldots, \Theta_k, \mathbf{X}^{\diamond 1}, \ldots, \mathbf{X}^{\diamond n}$, the $w_j^\diamond$ are constant and the $Z_j^\diamond$ are centred, independent and bounded by 1. Thus, using the folklore lemma,

$$\mathbb{E}\left[e^{tW_n}\right] = \mathbb{E}\left[\prod_{j=1}^{n} \mathbb{E}\left[e^{tw_j^\diamond(\mathbf{x})Z_j^\diamond}\middle| \mathcal{D}_n, \Theta_1, \ldots, \Theta_k, \mathbf{X}^{\diamond 1}, \ldots, \mathbf{X}^{\diamond n}\right]\right] \leqslant \mathbb{E}\left[\prod_{j=1}^{n} e^{t^2 w_j^\diamond(\mathbf{x})^2/2}\right].$$

Let $K > 0$ be such that for all $\ell = 1, \ldots, k$, $N_n^b\left(A_n(\ell)\right) = N_n^b\left(\mathbf{x}; \Theta_\ell, \mathcal{D}_n\right) \geqslant K\sqrt{n}\left(\ln(n)\right)^\beta$ a.s. by using Assumption 4.4.2 item 2. Denote $\Gamma(\ell)$ the event $\left\{N_n^\diamond(A_n(\ell)) < \frac{K\sqrt{n}(\ln(n))^\beta}{2}\right\}$. Remark that $\Gamma(\ell) \subset \left\{\left|N_n^\diamond(A_n(\ell)) - N_n^b\left(A_n(\ell)\right)\right| > \frac{K\sqrt{n}(\ln(n))^\beta}{2}\right\}$. Thus, using Lemma 4.5.3, we have that $\mathbb{P}\left(\Gamma(\ell)\right) \leq 24\left(n+1\right)^{2d}\exp\left[-\frac{K^2(\ln(n))^{2\beta}}{1152}\right]$.

We have

$$\begin{aligned}\sum_{j=1}^{n} w_j^\diamond(\mathbf{x})^2 &= \sum_{j=1}^{n} \frac{w_j^\diamond(\mathbf{x})}{k}\left(\sum_{\ell=1}^{k} \frac{\mathbb{1}_{\{\mathbf{X}^{\diamond j} \in A_n(\ell)\}}}{N_n^\diamond(A_n(\ell))}\left(\mathbb{1}_{\{\Gamma(\ell)^c\}} + \mathbb{1}_{\{\Gamma(\ell)\}}\right)\right) \\ &\leqslant \sum_{j=1}^{n} w_j^\diamond(\mathbf{x})\left(\frac{2}{K\sqrt{n}(\ln n)^\beta} + \frac{1}{k}\sum_{\ell=1}^{k} \mathbb{1}_{\{\mathbf{X}^{\diamond j} \in A_n(\ell)\}}\mathbb{1}_{\{\Gamma(\ell)\}}\right).\end{aligned}$$

So that,

$$\begin{aligned}\mathbb{E}\left[\prod_{j=1}^{n} e^{t^2 w_j^\diamond(\mathbf{x})^2/2}\right] &\leqslant \exp\left[t^2/\left(K\sqrt{n}\left(\ln(n)\right)^\beta\right)\right] \times \mathbb{E}\left[\exp\left(\frac{t^2}{2}\cdot \mathbb{1}_{\left\{\bigcup_{\ell=1}^{k}\Gamma(\ell)\right\}}\right)\right] \\ &\leqslant \exp\left[t^2/\left(K\sqrt{n}\left(\ln(n)\right)^\beta\right)\right] \times \left(1 + e^{t^2/2}\sum_{\ell=1}^{k}\mathbb{P}\left(\Gamma(\ell)\right)\right) \\ &\leqslant \exp\left[t^2/\left(K\sqrt{n}\left(\ln(n)\right)^\beta\right)\right] \times \left(1 + 24k\left(n+1\right)^{2d}\exp\left[\frac{t^2}{2} - \frac{K^2\left(\ln(n)\right)^{2\beta}}{1152}\right]\right).\end{aligned}$$

Taking $t^2 = \frac{K^2(\ln(n))^{2\beta}}{576}$ leads to

$$\mathbb{P}\left(W_n > \varepsilon\right) \leqslant \left(1 + 24k\left(n+1\right)^{2d}\right)\exp\left[\frac{K\left(\ln(n)\right)^\beta}{576\sqrt{n}} - \frac{\varepsilon K\left(\ln(n)\right)^\beta}{24}\right].$$

The same upper bound is obtained for $\mathbb{P}(W_n < -\varepsilon)$ by using that $\mathbb{P}(W_n < -\varepsilon) = \mathbb{P}(-W_n > \varepsilon)$.

Thus, by using Assumption 4.4.2, item 1., $k = O(n^\alpha)$ so that the right hand side is summable, we conclude that W_n goes to 0 almost surely.

2 Let us now show that $(W_n)_{n\geqslant 1}$ goes to 0 a.s. in the case where Assumption 4.4.2 item 3 is satisfied.
★ Let us first show that $(W_{n^2})_{n\geqslant 1}$ goes to 0 a.s.

$$\begin{aligned}
\mathbb{E}\left[(W_n)^2\right] &= \mathbb{E}\left[\left(\sum_{j=1}^{n} w_j^\diamond(\mathbf{x})\, Z_j^\diamond\right)^2\right] \\
&= \sum_{j=1}^{n}\sum_{m=1}^{n} \mathbb{E}\left[w_j^\diamond(\mathbf{x})\, w_m^\diamond(\mathbf{x})\, Z_j^\diamond Z_m^\diamond\right] \\
&= \sum_{j=1}^{n} \mathbb{E}\left[w_j^{\diamond^2}(\mathbf{x})\, Z_j^{\diamond^2}\right] + \sum_{\substack{1\leqslant j,m\leqslant n \\ j\neq m}} \mathbb{E}\left[w_j^\diamond(\mathbf{x})\, w_m^\diamond(\mathbf{x})\, Z_j^\diamond Z_m^\diamond\right] \\
&\overset{\text{def}}{=} I_n + J_n
\end{aligned}$$

$$\begin{aligned}
I_n &= \mathbb{E}\left[\sum_{j=1}^{n} w_j^{\diamond^2}(\mathbf{x})\, Z_j^{\diamond^2}\right] \\
&\leqslant \mathbb{E}\left[\sum_{j=1}^{n} w_j^{\diamond^2}(\mathbf{x})\right] \\
&\leqslant \mathbb{E}\left[\frac{1}{\min_{\ell=1,\ldots,k} N_n^\diamond\left(\mathbf{x};\Theta_\ell,\mathbf{X}^{\diamond 1},\ldots,\mathbf{X}^{\diamond n},\mathcal{D}_n\right)}\right] \text{(recall that } \sum_{j=1}^{n} w_j^\diamond(\mathbf{x}) = 1\text{)} \\
&\leqslant \mathbb{E}\left[\frac{1}{\min_{\ell=1,\ldots,k} N_n^\diamond\left(\mathbf{x};\Theta_\ell,\mathbf{X}^{\diamond 1},\ldots,\mathbf{X}^{\diamond n},\mathcal{D}_n\right)} \mathbb{1}\{\forall \ell \setminus N^\diamond(A_n(\Theta_\ell)) > \lambda\}\right] \\
&\quad + \mathbb{E}\left[\frac{1}{\min_{\ell=1,\ldots,k} N_n^\diamond\left(\mathbf{x};\Theta_\ell,\mathbf{X}^{\diamond 1},\ldots,\mathbf{X}^{\diamond n},\mathcal{D}_n\right)} \mathbb{1}\{\exists \ell \setminus N^\diamond(A_n(\Theta_\ell)) \leqslant \lambda\}\right]
\end{aligned}$$

where $\lambda = \dfrac{\mathbb{E}\left[N^b\left(A_n\left(\Theta\right)\right)\right]}{4}$.

$$
\begin{aligned}
I_n &\leqslant \frac{1}{\lambda} + \mathbb{P}\left(\exists \ell \setminus N^{\diamond}\left(A_n\left(\Theta_\ell\right)\right) \leqslant \lambda\right) \\
&\leqslant \frac{1}{\lambda} + k\mathbb{P}\left(N^{\diamond}\left(A_n\left(\Theta\right)\right) \leqslant \lambda\right) \\
&\leqslant \frac{1}{\lambda} + k\left[\mathbb{P}\left(N^{\diamond}\left(A_n\left(\Theta\right)\right) \leqslant \lambda,\ \left|N^b\left(A_n\left(\Theta\right)\right) - N^{\diamond}\left(A_n\left(\Theta\right)\right)\right| \leqslant \lambda\right)\right. \\
&\quad \left. + \mathbb{P}\left(N^{\diamond}\left(A_n\left(\Theta\right)\right) \leqslant \lambda,\ \left|N^b\left(A_n\left(\Theta\right)\right) - N^{\diamond}\left(A_n\left(\Theta\right)\right)\right| > \lambda\right)\right] \\
&\leqslant \frac{1}{\lambda} + k\left[\mathbb{P}\left(N^b\left(A_n\left(\Theta\right)\right) \leqslant 2\lambda\right) + \mathbb{P}\left(\left|N^b\left(A_n\left(\Theta\right)\right) - N^{\diamond}\left(A_n\left(\Theta\right)\right)\right| > \lambda\right)\right] .
\end{aligned}
$$

Using Bienaymé-Tchebychev's inequality, we get

$$
\begin{aligned}
\mathbb{P}\left(N^b\left(A_n\left(\Theta\right)\right) \leqslant 2\lambda\right) &\leqslant 4\frac{\operatorname{Var}\left(N^b\left(A_n\left(\Theta\right)\right)\right)}{\left(\mathbb{E}\left[N^b\left(A_n\left(\Theta\right)\right)\right]\right)^2} \\
&\leqslant 4\left(\operatorname{CV}\left(N^b\left(A_n\left(\Theta\right)\right)\right)\right)^2 .
\end{aligned}
$$

Finally, thanks to Lemma 4.5.3 and Assumption 4.4.2 items 1. and 3., there exist C, K and M positive constants such that

$$
\begin{aligned}
I_n &\leqslant \frac{4}{\mathbb{E}\left[N^b\left(A_n\left(\Theta\right)\right)\right]} + 4k\left(\operatorname{CV}\left(N^b\left(A_n\left(\Theta\right)\right)\right)\right)^2 + k\mathbb{P}\left(\left|N^b\left(A_n\left(\Theta\right)\right) - N^{\diamond}\left(A_n\left(\Theta\right)\right)\right| > \lambda\right) \\
&\leqslant \frac{4}{K\sqrt{n}\left(\ln\left(n\right)\right)^{\beta}} + \frac{4CM^2}{n\left(\ln\left(n\right)\right)^{\gamma}} + 24Cn^{\alpha}(n+1)^{2d}\exp\left[-\frac{K^2\left(\ln\left(n\right)\right)^{2\beta}}{4608}\right] .
\end{aligned}
$$

The trick of using a second sample $\mathcal{D}_n^{\diamond}$ independent of the first-one and the random variable Θ is really important to handle the J_n term. Indeed, we have $J_n = 0$ while the equivalent term encountered in the proof of the Theorem 2 developed by Scornet et al. [2015b] is handled using a conjecture regarding the correlation behavior of the CART algorithm that is difficult to verify (cf. assumption (H2) of Scornet et al. [2015b]). Indeed

$$
\begin{aligned}
J_n &= \sum_{\substack{1\leqslant j,m\leqslant n \\ j\neq m}} \mathbb{E}\left[w_j^{\diamond}\left(\mathbf{x}\right) w_m^{\diamond}\left(\mathbf{x}\right) Z_j^{\diamond} Z_m^{\diamond}\right] \\
&= \sum_{\substack{1\leqslant j,m\leqslant n \\ j\neq m}} \mathbb{E}\left[\mathbb{E}\left[w_j^{\diamond}\left(\mathbf{x}\right) w_m^{\diamond}\left(\mathbf{x}\right) Z_j^{\diamond} Z_m^{\diamond}\middle| \Theta_1, \ldots, \Theta_k, \mathcal{D}_n, \mathbf{X}^{\diamond 1}, \ldots, \mathbf{X}^{\diamond n}, Y^{\diamond j}\right]\right] \\
&= \sum_{\substack{1\leqslant j,m\leqslant n \\ j\neq m}} \mathbb{E}\left[w_j^{\diamond}\left(\mathbf{x}\right) w_m^{\diamond}\left(\mathbf{x}\right) Z_j^{\diamond}\mathbb{E}\left[Z_m^{\diamond}\middle| \mathbf{X}^{\diamond m}\right]\right] \text{ using that } w_j^{\diamond}\left(\mathbf{x}\right), w_m^{\diamond}\left(\mathbf{x}\right) \text{ and} \\
&\quad Z_j^{\diamond} \text{ are } \Theta_1, \ldots, \Theta_k, \mathcal{D}_n, \left(\mathbf{X}^{\diamond j}\right)_{j=1,\ldots,n}, Y^{\diamond j} \text{ measurable.} \\
&= 0 \text{ because } \mathbb{E}\left[Z_m^{\diamond}\middle| \mathbf{X}^{\diamond m}\right] = 0 .
\end{aligned}
$$

Finally,

$$\forall \varepsilon > 0, \quad \mathbb{P}\left(|W_n| \geqslant \varepsilon\right) \leqslant \frac{\mathbb{E}\left[(W_n)^2\right]}{\varepsilon^2} = \frac{I_n}{\varepsilon^2} \ .$$

Hence, since $\sum_{n\geq 1} I_{n^2} < \infty$, Borel–Cantelli Lemma gives

$$\forall \varepsilon > 0, \quad \mathbb{P}\left(\limsup_{n\to\infty} \left\{|W_{n^2}| \geqslant \varepsilon\right\}\right) = 0 \ ,$$

which implies that $W_{n^2} \xrightarrow[n\to\infty]{a.s.} 0$.

★ Let us now show that $(W_n)_{n\geqslant 1}$ converges almost surely to 0.

Let $p = p(n) = \lfloor \sqrt{n} \rfloor$, we have $W_n - W_{p^2} = \sum_{j=p^2+1}^{n} w_j^\diamond(\mathbf{x}) Z_j^\diamond$. Fix $\varepsilon > 0$ and consider again $\lambda = \dfrac{\mathbb{E}\left[N^b\left(A_n(\Theta)\right)\right]}{4}$,

$$\begin{aligned}
\mathbb{P}\left(|W_n - W_{p^2}| \geqslant \varepsilon\right) &\leqslant \mathbb{P}\left(\sum_{j=p^2+1}^{n} w_j^\diamond(\mathbf{x}) \left|Z_j^\diamond\right| \geqslant \varepsilon\right) \\
&\leqslant \mathbb{P}\left(\frac{2\sqrt{n}}{\min\limits_{\ell=1,\dots,k} N_n^\diamond\left(\mathbf{x}; \Theta_\ell, \mathbf{X}^{\diamond 1}, \dots, \mathbf{X}^{\diamond n}, \mathcal{D}_n\right)} \geqslant \varepsilon\right) \\
&\leqslant \mathbb{P}\left(\frac{2\sqrt{n}}{\min\limits_{\ell=1,\dots,k} N_n^\diamond\left(\mathbf{x}; \Theta_\ell, \mathbf{X}^{\diamond 1}, \dots, \mathbf{X}^{\diamond n}, \mathcal{D}_n\right)} \geqslant \varepsilon, \ \forall \ell \setminus N^\diamond\left(A_n(\Theta_\ell)\right) > \lambda\right) \\
&\quad + \mathbb{P}\left(\frac{2\sqrt{n}}{\min\limits_{\ell=1,\dots,k} N_n^\diamond\left(\mathbf{x}; \Theta_\ell, \mathbf{X}^{\diamond 1}, \dots, \mathbf{X}^{\diamond n}, \mathcal{D}_n\right)} \geqslant \varepsilon, \ \exists \ell \setminus N^\diamond\left(A_n(\Theta_\ell)\right) \leqslant \lambda\right) \ .
\end{aligned}$$

The first right-hand term is zero because $\mathbb{E}\left[N^b\left(A_n(\Theta)\right)\right] \geqslant \dfrac{8\sqrt{n}}{\varepsilon}$ for n large enough thanks to Assumption 4.4.2 item 3 and therefore:

$$\begin{aligned}
\mathbb{P}\left(|W_n - W_{p^2}| \geqslant \varepsilon\right) &\leqslant \mathbb{P}\left(\exists \ell \setminus N^\diamond\left(A_n(\Theta_\ell)\right) \leqslant \lambda\right) \\
&\leqslant k\mathbb{P}\left(N^\diamond\left(A_n(\Theta)\right) \leqslant \lambda\right) \\
&\leqslant k\left[\mathbb{P}\left(N^\diamond\left(A_n(\Theta)\right) \leqslant \lambda, \ \left|N^b\left(A_n(\Theta)\right) - N^\diamond\left(A_n(\Theta)\right)\right| \leqslant \lambda\right)\right. \\
&\quad \left. + \mathbb{P}\left(N^\diamond\left(A_n(\Theta)\right) \leqslant \lambda, \ \left|N^b\left(A_n(\Theta)\right) - N^\diamond\left(A_n(\Theta)\right)\right| > \lambda\right)\right] \\
&\leqslant k\left[\mathbb{P}\left(N^b\left(A_n(\Theta)\right) \leqslant 2\lambda\right) + \mathbb{P}\left(\left|N^b\left(A_n(\Theta)\right) - N^\diamond\left(A_n(\Theta)\right)\right| > \lambda\right)\right] \ .
\end{aligned}$$

Again, thanks to the Bienaymé-Tchebychev's inequality, we have

$$\begin{aligned}\mathbb{P}\left(N^b\left(A_n\left(\Theta\right)\right)\leqslant 2\lambda\right) &\leqslant 4\frac{\mathrm{Var}\left(N^b\left(A_n\left(\Theta\right)\right)\right)}{\left(\mathbb{E}\left[N^b\left(A_n\left(\Theta\right)\right)\right]\right)^2}\\ &\leqslant 4\left(\mathrm{CV}\left(N^b\left(A_n\left(\Theta\right)\right)\right)\right)^2 .\end{aligned}$$

Finally, thanks to Lemma 4.5.3 and Assumption 4.4.2 items 1. and 3., we have for n large enough

$$\begin{aligned}\mathbb{P}\left(\left|W_n - W_{p^2}\right|\geqslant\varepsilon\right) &\leqslant 4k\left(\mathrm{CV}\left(N^b\left(A_n\left(\Theta\right)\right)\right)\right)^2 + k\mathbb{P}\left(\left|N^b\left(A_n\left(\Theta\right)\right) - N^\diamond\left(A_n\left(\Theta\right)\right)\right| > \lambda\right)\\ &\leqslant \frac{4CM^2}{n\left(\ln\left(n\right)\right)^\gamma} + 24Cn^\alpha(n+1)^{2d}\exp\left[-\frac{K^2\left(\ln\left(n\right)\right)^{2\beta}}{4608}\right] .\end{aligned}$$

Hence, using Borel–Cantelli Lemma

$$\forall\varepsilon>0,\quad \mathbb{P}\left(\limsup_{n\to\infty}\left\{\left|W_n - W_{p^2}\right|\geqslant\varepsilon\right\}\right) = 0 ,$$

which implies that the random variable $W_n - W_{p^2} \xrightarrow[n\to\infty]{a.s.} 0$.

From this, we deduce that $\left(W_n\right)_{n\geqslant 1}$ goes to 0 a.s.

3 Finally, we show that $\left(V_n\right)_{n\geqslant 1}$ goes to 0 a.s.

$$\begin{aligned}\left|V_n\right| &= \left|\sum_{j=1}^n w_j^\diamond\left(\mathbf{x}\right)\left(F\left(y\middle|\,\mathbf{X}^{\diamond j}\right) - F\left(y\middle|\,\mathbf{x}\right)\right)\right|\\ &\leqslant \frac{1}{k}\sum_{\ell=1}^k\left(\sum_{j=1}^n \frac{\mathbb{1}_{\{\mathbf{X}^{\diamond j}\in A_n(\mathbf{x};\Theta_\ell,\mathcal{D}_n)\}}}{N_n^\diamond\left(\mathbf{x};\Theta_\ell,\mathbf{X}^{\diamond 1},\ldots,\mathbf{X}^{\diamond n},\mathcal{D}_n\right)}\left|F\left(y\middle|\,\mathbf{X}^{\diamond j}\right) - F\left(y\middle|\,\mathbf{x}\right)\right|\right)\\ &\leqslant \frac{1}{k}\sum_{\ell=1}^k\left(\sum_{j=1}^n \frac{\mathbb{1}_{\{\mathbf{X}^{\diamond j}\in A_n(\mathbf{x};\Theta_\ell,\mathcal{D}_n)\}}}{N_n^\diamond\left(\mathbf{x};\Theta_\ell,\mathbf{X}^{\diamond 1},\ldots,\mathbf{X}^{\diamond n},\mathcal{D}_n\right)}\sup_{\mathbf{z}\in A_n(\mathbf{x};\Theta_\ell,\mathcal{D}_n)}\left|F\left(y\middle|\,\mathbf{z}\right) - F\left(y\middle|\,\mathbf{x}\right)\right|\right)\\ &\leqslant \frac{1}{k}\sum_{\ell=1}^k\sup_{\mathbf{z}\in A_n(\mathbf{x};\Theta_\ell,\mathcal{D}_n)}\left|F\left(y\middle|\,\mathbf{z}\right) - F\left(y\middle|\,\mathbf{x}\right)\right|\end{aligned}$$

Hence, by using Assumption 4.4.1, we have

$$\left|\sum_{j=1}^n w_j^\diamond\left(\mathbf{x}\right)\left(F\left(y\middle|\,\mathbf{X}^{\diamond j}\right) - F\left(y\middle|\,\mathbf{x}\right)\right)\right| \xrightarrow[n\to\infty]{a.s.} 0 .$$

This allows us to conclude that

$$\forall\mathbf{x}\in\mathcal{X},\forall y\in\mathbb{R},\quad F_{k,n}^\diamond\left(y\middle|\,\mathbf{X}=\mathbf{x}\right)\xrightarrow[n\to\infty]{a.s.} F\left(y\middle|\,\mathbf{X}=\mathbf{x}\right) .$$

■

Let us now turn to the proof of Lemma 4.5.2 which shows that the dummy estimator $F^{\diamond}_{k,n}$ is close to the interesting one $F^{b}_{k,n}$.

Proof of Lemma 4.5.2.

Let $\mathbf{x} \in \mathcal{X}$ and $y \in \mathbb{R}$, we have that

$$\begin{aligned}
&\left|F^{\diamond}_{k,n}\left(y|\,\mathbf{X}=\mathbf{x}\right) - F^{b}_{k,n}\left(y|\,\mathbf{X}=\mathbf{x}\right)\right| \\
&= \left|\frac{1}{k}\sum_{\ell=1}^{k}\left(\sum_{j=1}^{n}\frac{\mathbb{1}_{\{\mathbf{X}^{\diamond j}\in A_n(\mathbf{x};\Theta_\ell,\mathcal{D}_n)\}}\mathbb{1}_{\{Y^{\diamond j}\leqslant y\}}}{N^{\diamond}_n\left(\mathbf{x};\Theta_\ell,\mathbf{X}^{\diamond 1},\ldots,\mathbf{X}^{\diamond n},\mathcal{D}_n\right)} - \sum_{j=1}^{n}\frac{B_j\left(\Theta^1_\ell,\mathcal{D}_n\right)\mathbb{1}_{\{\mathbf{X}^{j}\in A_n(\mathbf{x};\Theta_\ell,\mathcal{D}_n)\}}\mathbb{1}_{\{Y^{j}\leqslant y\}}}{N^{b}_n\left(\mathbf{x};\Theta_\ell,\mathcal{D}_n\right)}\right)\right| \\
&= \left|\frac{1}{k}\sum_{\ell=1}^{k}\left(\frac{\#\left\{j\leqslant J^{\diamond}\ /\ \mathbf{X}^{\diamond(j)}\in A_n\left(\Theta_\ell\right)\right\}}{N^{\diamond}\left(A_n\left(\Theta_\ell\right)\right)} - \frac{\sum_{j\in\mathcal{S}}B_j\left(\Theta^1_\ell,\mathcal{D}_n\right)}{N^{b}\left(A_n\left(\Theta_\ell\right)\right)}\right)\right| \text{ with } \mathcal{S} = \left\{j\leqslant J\ /\ \mathbf{X}^{(j)}\in A_n\left(\Theta_\ell\right)\right\}\ ,
\end{aligned}$$

where we denote $A_n\left(\Theta_\ell\right) = A_n\left(\mathbf{x};\Theta_\ell,\mathcal{D}_n\right), N^{\diamond}\left(A_n\left(\Theta_\ell\right)\right) = N^{\diamond}_n\left(\mathbf{x};\Theta_\ell,\mathbf{X}^{\diamond 1},\ldots,\mathbf{X}^{\diamond n},\mathcal{D}_n\right)$ and $N^{b}\left(A_n\left(\Theta_\ell\right)\right) = N^{b}_n\left(\mathbf{x};\Theta_\ell,\mathcal{D}_n\right)$. $J, J^{\diamond}$ are such that $Y^{\diamond(J^{\diamond})} \leqslant y < Y^{\diamond(J^{\diamond}+1)}$ and $Y^{(J)} \leqslant y < Y^{(J+1)}$, with $Y^{\diamond(j)}$ (resp. $Y^{(j)}$) the order statistics of $(Y^{\diamond 1},\ldots,Y^{\diamond n})$ (resp. $(Y^1,\ldots,Y^n)$) and the $\mathbf{X}^{\diamond(j)}$ (resp. $\mathbf{X}^{(j)}$) the corresponding $\mathbf{X}^{\diamond p}$'s (resp. $\mathbf{X}^{p}$'s).

Let us consider for some $\ell \in [\![1,k]\!]$,

$$G = \frac{\#\left\{j\leqslant J^{\diamond}\ /\ \mathbf{X}^{\diamond(j)}\in A_n\left(\Theta_\ell\right)\right\}}{N^{\diamond}\left(A_n\left(\Theta_\ell\right)\right)} - \frac{\sum_{j\in\mathcal{S}}B_j\left(\Theta^1_\ell,\mathcal{D}_n\right)}{N^{b}\left(A_n\left(\Theta_\ell\right)\right)} \overset{\text{def}}{=} \frac{N^{\diamond}_{J^{\diamond}}\left(A_n\left(\Theta_\ell\right)\right)}{N^{\diamond}\left(A_n\left(\Theta_\ell\right)\right)} - \frac{N_J\left(A_n\left(\Theta_\ell\right)\right)}{N^{b}\left(A_n\left(\Theta_\ell\right)\right)}\ .$$

We have,

$$\begin{aligned}
|G| &\leqslant \frac{\left|N^{\diamond}\left(A_n\left(\Theta_\ell\right)\right) - N^{b}\left(A_n\left(\Theta_\ell\right)\right)\right|}{N^{b}\left(A_n\left(\Theta_\ell\right)\right)} + \frac{\left|N^{\diamond}_{J^{\diamond}}\left(A_n\left(\Theta_\ell\right)\right) - N_J\left(A_n\left(\Theta_\ell\right)\right)\right|}{N^{b}\left(A_n\left(\Theta_\ell\right)\right)} \\
&\overset{\text{def}}{=} |G_1| + |G_2|
\end{aligned}$$

We continue the proof below in the case where Assumption 4.4.2 item 3. is satisfied. The case where item 2. is verified is done easier following the same lines. Let $\varepsilon > 0$.

1. We are now going to show the almost everywhere convergence to 0 for each term G_1 and G_2. Let us start with G_1.

$$\begin{aligned}
\mathbb{P}\left(|G_1| > \varepsilon\right) &= \mathbb{P}\left(\frac{\left|N^{\diamond}\left(A_n\left(\Theta_\ell\right)\right) - N^{b}\left(A_n\left(\Theta_\ell\right)\right)\right|}{N^{b}\left(A_n\left(\Theta_\ell\right)\right)} > \varepsilon\right) \\
&= \mathbb{P}\left(\left|N^{\diamond}\left(A_n\left(\Theta_\ell\right)\right) - N^{b}\left(A_n\left(\Theta_\ell\right)\right)\right| > \varepsilon N^{b}\left(A_n\left(\Theta_\ell\right)\right), N^{b}\left(A_n\left(\Theta_\ell\right)\right) > \lambda\right) \\
&\quad + \mathbb{P}\left(\left|N^{\diamond}\left(A_n\left(\Theta_\ell\right)\right) - N^{b}\left(A_n\left(\Theta_\ell\right)\right)\right| > \varepsilon N^{b}\left(A_n\left(\Theta_\ell\right)\right), N^{b}\left(A_n\left(\Theta_\ell\right)\right) \leqslant \lambda\right)
\end{aligned}$$

$$\text{where } \lambda = \frac{\mathbb{E}\left[N^b\left(A_n\left(\Theta\right)\right)\right]}{2}$$

$$\leqslant \mathbb{P}\left(\left|N^{\diamond}\left(A_n\left(\Theta_\ell\right)\right) - N^b\left(A_n\left(\Theta_\ell\right)\right)\right| > \varepsilon\lambda\right) + \mathbb{P}\left(N^b\left(A_n\left(\Theta_\ell\right)\right) \leqslant \lambda\right)$$

Again, thanks to the Bienaymé-Tchebychev's inequality,

$$\begin{aligned}\mathbb{P}\left(N^b\left(A_n\left(\Theta_\ell\right)\right) \leqslant \lambda\right) &\leqslant 4\frac{\mathrm{Var}\left(N^b\left(A_n\left(\Theta_\ell\right)\right)\right)}{\left(\mathbb{E}\left[N^b\left(A_n\left(\Theta_\ell\right)\right)\right]\right)^2} \\ &\leqslant 4\left(\mathrm{CV}\left(N^b\left(A_n\left(\Theta\right)\right)\right)\right)^2 \ . \end{aligned} \tag{4.11}$$

Now, using Lemma 4.5.3 and Assumption 4.4.2, we get

$$\begin{aligned}\mathbb{P}\left(|G_1| > \varepsilon\right) &\leqslant \mathbb{P}\left(\left|N^{\diamond}\left(A_n\left(\Theta_\ell\right)\right) - N^b\left(A_n\left(\Theta_\ell\right)\right)\right| > \varepsilon\lambda\right) + 4\left(\mathrm{CV}\left(N^b\left(A_n\left(\Theta\right)\right)\right)\right)^2 \\ &\leqslant 24(n+1)^{2d}\exp\left[-\frac{\varepsilon^2K^2\left(\ln\left(n\right)\right)^{2\beta}}{1152}\right] + \frac{4M^2}{n^{\alpha+1}\left(\ln\left(n\right)\right)^{\gamma}} \ .\end{aligned}$$

Then, thanks to Borel–Cantelli Lemma

$$\forall\varepsilon > 0, \quad \mathbb{P}\left(\limsup_{n\to\infty}\left\{|G_1| > \varepsilon\right\}\right) = 0 \ ,$$

which implies $G_1 \xrightarrow[n\to\infty]{a.s.} 0$.

2 Now, consider the G_2 term.

$$\begin{aligned}\mathbb{P}\left(|G_2| > \varepsilon\right) &= \mathbb{P}\left(\frac{\left|N^{\diamond}_{J^{\diamond}}\left(A_n\left(\Theta_\ell\right)\right) - N_J\left(A_n\left(\Theta_\ell\right)\right)\right|}{N^b\left(A_n\left(\Theta_\ell\right)\right)} > \varepsilon\right) \\ &= \mathbb{P}\left(\left|N^{\diamond}_{J^{\diamond}}\left(A_n\left(\Theta_\ell\right)\right) - N_J\left(A_n\left(\Theta_\ell\right)\right)\right| > \varepsilon N^b\left(A_n\left(\Theta_\ell\right)\right), N^b\left(A_n\left(\Theta_\ell\right)\right) > \lambda\right) \\ &\quad + \mathbb{P}\left(\left|N^{\diamond}_{J^{\diamond}}\left(A_n\left(\Theta_\ell\right)\right) - N_J\left(A_n\left(\Theta_\ell\right)\right)\right| > \varepsilon N^b\left(A_n\left(\Theta_\ell\right)\right), N^b\left(A_n\left(\Theta_\ell\right)\right) \leqslant \lambda\right)\end{aligned}$$

$$\text{where } \lambda = \frac{\mathbb{E}\left[N^b\left(A_n\left(\Theta\right)\right)\right]}{2}$$

$$\leqslant \mathbb{P}\left(\left|N^{\diamond}_{J^{\diamond}}\left(A_n\left(\Theta_\ell\right)\right) - N_J\left(A_n\left(\Theta_\ell\right)\right)\right| > \varepsilon\lambda\right) + \mathbb{P}\left(N^b\left(A_n\left(\Theta_\ell\right)\right) \leqslant \lambda\right)$$

We are going to bound the first term by using again the Vapnik-Chervonenkis theory.

By considering the class $\mathcal{B} = \left\{ \prod_{i=1}^{d} [a_i, b_i] \times]-\infty, y] : a_i, b_i \in \overline{\mathbb{R}} \right\}$, we have

$$
\begin{aligned}
&\mathbb{P}\left(\left|N_{J^\diamond}^{\diamond}\left(A_n\left(\Theta_\ell\right)\right) - N_J\left(A_n\left(\Theta_\ell\right)\right)\right| > \varepsilon\lambda\right) \\
&= \mathbb{P}\left(\frac{\left|N_{J^\diamond}^{\diamond}\left(A_n\left(\Theta_\ell\right)\right) - N_J\left(A_n\left(\Theta_\ell\right)\right)\right|}{n} > \frac{\varepsilon\lambda}{n}\right) \\
&\leqslant \mathbb{P}\left(\left|\frac{N_{J^\diamond}^{\diamond}\left(A_n\left(\Theta_\ell\right)\right)}{n} - \mathbb{P}_{\mathbf{X},Y}\left((\mathbf{X},Y) \in A_n\left(\Theta_\ell\right) \times]-\infty, y]\right)\right| > \frac{\varepsilon\lambda}{3n}\right) \\
&\quad + \mathbb{P}\left(\left|\frac{1}{n}\sum_{j=1}^{n} \mathbb{1}_{\{(\mathbf{X}^j, Y^j) \in A_n(\Theta_\ell) \times]-\infty, y]\}} - \mathbb{P}_{\mathbf{X},Y}\left((\mathbf{X},Y) \in A_n\left(\Theta_\ell\right) \times]-\infty, y]\right)\right| > \frac{\varepsilon\lambda}{3n}\right) \\
&\quad + \mathbb{P}\left(\left|\frac{N_J\left(A_n\left(\Theta_\ell\right)\right)}{n} - \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}_{\{(\mathbf{X}^j, Y^j) \in A_n(\Theta_\ell) \times]-\infty, y]\}}\right| > \frac{\varepsilon\lambda}{3n}\right) \\
&\leqslant \mathbb{P}\left(\sup_{A \in \mathcal{B}}\left|\frac{1}{n}\sum_{j=1}^{n} \mathbb{1}_{\{(\mathbf{X}^{\diamond j}, Y^{\diamond j}) \in A\}} - \mathbb{P}_{\mathbf{X},Y}\left((\mathbf{X},Y) \in A\right)\right| > \frac{\varepsilon\lambda}{3n}\right) \qquad (4.12) \\
&\quad + \mathbb{P}\left(\sup_{A \in \mathcal{B}}\left|\frac{1}{n}\sum_{j=1}^{n} \mathbb{1}_{\{(\mathbf{X}^j, Y^j) \in A\}} - \mathbb{P}_{\mathbf{X},Y}\left((\mathbf{X},Y) \in A\right)\right| > \frac{\varepsilon\lambda}{3n}\right) \\
&\quad + \mathbb{P}\left(\sup_{A \in \mathcal{B}}\left|\frac{1}{n}\sum_{j=1}^{n} B_j\left(\Theta_\ell^1, \mathcal{D}_n\right) \mathbb{1}_{\{(\mathbf{X}^j, Y^j) \in A\}} - \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}_{\{(\mathbf{X}^j, Y^j) \in A\}}\right| > \frac{\varepsilon\lambda}{3n}\right) .
\end{aligned}
$$

Let us make some comments on Vapnik-Chervonenkis dimension of the class $\mathcal{B}$. If we had the class $\left\{ \prod_{i=1}^{d} [a_i, b_i] \times]-\infty, c] : a_i, b_i \in \overline{\mathbb{R}}, c \in \mathbb{R} \right\}$, it could be shown by calculations similar to those from Theorem 13.8 in Devroye et al. [2013] that the Vapnik-Chervonenkis dimension is $2d + 1$. But in our case, the element c is fixed as y, then all the possibilities to break the points are related to the elements a_i, b_i, which thus gives us a Vapnik-Chervonenkis dimension equals to $2d$. Therefore, the first two right-hand terms are handled thanks to a direct application of the Theorem of Vapnik and Chervonenkis [1971] over the class $\mathcal{B}$. The latter deserves special attention.

The third term is treated as in the proof of Lemma 4.5.3. We use as before, the representation of the bootstrap component with the random variables $Z_\ell^1, \ldots, Z_\ell^n$. We apply Vapnik-Chervonenkis' Theorem under the conditional distribution given $\mathcal{D}_n$ and get

$$
\mathbb{P}\left(\sup_{A \in \mathcal{B}}\left|\frac{1}{n}\sum_{j=1}^{n} \mathbb{1}_{\left\{Z_\ell^j \in A\right\}} - \mathbb{P}\left(Z \in A | \mathcal{D}_n\right)\right| > \frac{\varepsilon\lambda}{3n} \middle| \mathcal{D}_n\right) \leqslant 8(n+1)^{2d} e^{-\varepsilon^2 \cdot \lambda^2/288n} .
$$

Therefore,

$$
\mathbb{P}\left(\sup_{A \in \mathcal{B}}\left|\frac{1}{n}\sum_{j=1}^{n} \mathbb{1}_{\left\{Z_\ell^j \in A\right\}} - \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}_{\{(\mathbf{X}^j, Y^j) \in A\}}\right| > \frac{\varepsilon\lambda}{3n}\right) \leqslant 8(n+1)^{2d} e^{-\varepsilon^2\lambda^2/288n} .
$$

Hence, at last

$$\mathbb{P}\left(\left|N^{\diamond}_{J^{\diamond}}\left(A_n\left(\Theta_\ell\right)\right) - N_J\left(A_n\left(\Theta_\ell\right)\right)\right| > \varepsilon\lambda\right) \leqslant 24(n+1)^{2d} e^{-\varepsilon^2\lambda^2/288n} \ .$$

As a consequence, Equation (4.11) and Assumption 4.4.2 lead to

$$\begin{aligned}\mathbb{P}\left(|G_2| > \varepsilon\right) &\leqslant \mathbb{P}\left(\left|N^{\diamond}_{J^{\diamond}}\left(A_n\left(\Theta_\ell\right)\right) - N_J\left(A_n\left(\Theta_\ell\right)\right)\right| > \varepsilon\lambda\right) + 4\left(\mathrm{CV}\left(N\left(A_n\left(\Theta\right)\right)\right)\right)^2 \\ &\leqslant 24(n+1)^{2d}\exp\left[-\frac{\varepsilon^2 K^2\left(\ln\left(n\right)\right)^{2\beta}}{1152}\right] + \frac{4M^2}{n^{\alpha+1}\left(\ln\left(n\right)\right)^{\gamma}} \ .\end{aligned}$$

Thanks to Borel–Cantelli Lemma, we get

$$\forall \varepsilon > 0, \quad \mathbb{P}\left(\limsup_{n\to\infty}\left\{|G_2| > \varepsilon\right\}\right) = 0 \ ,$$

which implies that $G_2 \xrightarrow[n\to\infty]{a.s.} 0$.

We conclude that G goes to 0 for all ℓ, thus,

$$\forall \mathbf{x} \in \mathcal{X}, \forall y \in \mathbb{R}, \quad \left|F^{\diamond}_{k,n}\left(y|\,\mathbf{X} = \mathbf{x}\right) - F^{b}_{k,n}\left(y|\,\mathbf{X} = \mathbf{x}\right)\right| \xrightarrow[n\to\infty]{a.s.} 0 \ .$$

In the case where Assumption 4.4.2 item 2. is verified, it exists $K > 0$ such that

$$N^b\left(A_n\left(\Theta_\ell\right)\right) \geqslant K\sqrt{n}\left(\ln\left(n\right)\right)^{\beta} \ \ a.s.$$

So that $\mathbb{P}(|G_1| > \varepsilon)$ and $\mathbb{P}(|G_2| > \varepsilon)$ are bounded above respectively by

- $\mathbb{P}\left(\left|N^b\left(A_n\left(\Theta_\ell\right)\right) - N^{\diamond}\left(A_n\left(\Theta_\ell\right)\right)\right| > \varepsilon K\sqrt{n}\left(\ln\left(n\right)\right)^{\beta}\right)$,
- and $\mathbb{P}\left(\left|N^{\diamond}_{J^{\diamond}}\left(A_n\left(\Theta_\ell\right)\right) - N_J\left(A_n\left(\Theta_\ell\right)\right)\right| > \varepsilon K\sqrt{n}\left(\ln\left(n\right)\right)^{\beta}\right)$.

A simple application of Lemma 4.5.3 and an adaptation of it to $N_J\left(A_n\left(\Theta_\ell\right)\right)$ show that G_1 and G_2 go to 0 a.s. ■

This conclude the proof of Theorem 4.4.1, we now sketch the proof of Theorem 4.4.2 which is a bit simpler.

4.5.2 Proof of Theorem 4.4.2

The different steps are similar to those of the proof of Theorem 4.4.1 and the dummy estimator $F^{\diamond}_{k,n}\left(y|\,\mathbf{X} = \mathbf{x}\right)$ introduced in the proof of Theorem 4.4.1 will be reused.

Let $\mathbf{x} \in \mathcal{X}$ and $y \in \mathbb{R}$, we have

$$\begin{aligned}\left|F^{o}_{k,n}\left(y|\,\mathbf{X} = \mathbf{x}\right) - F\left(y|\,\mathbf{X} = \mathbf{x}\right)\right| \leqslant &\left|F^{\diamond}_{k,n}\left(y|\,\mathbf{X} = \mathbf{x}\right) - F\left(y|\,\mathbf{X} = \mathbf{x}\right)\right| \\ &+ \left|F^{\diamond}_{k,n}\left(y|\,\mathbf{X} = \mathbf{x}\right) - F^{o}_{k,n}\left(y|\,\mathbf{X} = \mathbf{x}\right)\right| \ .\end{aligned}$$

As in the proof of Theorem 4.4.1, the first right-hand term is handled thanks to Proposition 4.5.1 which gives

$$\forall \mathbf{x} \in \mathcal{X}, \forall y \in \mathbb{R}, \quad F^{\diamond}_{k,n}\left(y | \mathbf{X} = \mathbf{x}\right) \xrightarrow[n\to\infty]{a.s.} F\left(y | \mathbf{X} = \mathbf{x}\right) \ .$$

The convergence of the second right-hand term is handled into the following Lemma 4.5.4.

Lemma 4.5.4.
Consider a random forest which satisfies Assumption 4.4.2. Then,

$$\forall \mathbf{x} \in \mathcal{X}, \forall y \in \mathbb{R}, \quad \left|F^{\diamond}_{k,n}\left(y | \mathbf{X} = \mathbf{x}\right) - F^{o}_{k,n}\left(y | \mathbf{X} = \mathbf{x}\right)\right| \xrightarrow[n\to\infty]{a.s.} 0$$

This allows us to conclude that

$$\forall \mathbf{x} \in \mathcal{X}, \forall y \in \mathbb{R}, \quad F^{o}_{k,n}\left(y | \mathbf{X} = \mathbf{x}\right) \xrightarrow[n\to\infty]{a.s.} F\left(y | \mathbf{X} = \mathbf{x}\right) \ .$$

As in the proof of Theorem 4.4.1, the almost sure uniform convergence relative to y of the estimator is achieved using Dini's second theorem, which concludes the proof.

■

Proof of Lemma 4.5.4.

The proof is done by following the same steps as the proof of Lemma 4.5.2 and with the same notations. Let $\mathbf{x} \in \mathcal{X}$ and $y \in \mathbb{R}$, we have

$$\begin{aligned}
&\left|F^{\diamond}_{k,n}\left(y | \mathbf{X} = \mathbf{x}\right) - F^{o}_{k,n}\left(y | \mathbf{X} = \mathbf{x}\right)\right| \\
&= \left|\frac{1}{k}\sum_{\ell=1}^{k}\left(\frac{\#\left\{j \leqslant J^{\diamond} \ / \ \mathbf{X}^{\diamond(j)} \in A_n\left(\Theta_\ell\right)\right\}}{N^{\diamond}\left(A_n\left(\Theta_\ell\right)\right)} - \frac{\#\left\{j \leqslant J \ / \ \mathbf{X}^{(j)} \in A_n\left(\Theta_\ell\right)\right\}}{N^{o}\left(A_n\left(\Theta_\ell\right)\right)}\right)\right| \ .
\end{aligned}$$

For any $\ell \in [\![1, k]\!]$,

$$G = \frac{\#\left\{j \leqslant J^{\diamond} \ / \ \mathbf{X}^{\diamond(j)} \in A_n\left(\Theta_\ell\right)\right\}}{N^{\diamond}\left(A_n\left(\Theta_\ell\right)\right)} - \frac{\#\left\{j \leqslant J \ / \ \mathbf{X}^{(j)} \in A_n\left(\Theta_\ell\right)\right\}}{N^{o}\left(A_n\left(\Theta_\ell\right)\right)} \overset{\text{def}}{=} \frac{N^{\diamond}_{J^{\diamond}}\left(A_n\left(\Theta_\ell\right)\right)}{N^{\diamond}\left(A_n\left(\Theta_\ell\right)\right)} - \frac{N_J\left(A_n\left(\Theta_\ell\right)\right)}{N^{o}\left(A_n\left(\Theta_\ell\right)\right)} \ .$$

We have,

$$\begin{aligned}
|G| &\leqslant \frac{\left|N^{\diamond}\left(A_n\left(\Theta_\ell\right)\right) - N^{o}\left(A_n\left(\Theta_\ell\right)\right)\right|}{N^{o}\left(A_n\left(\Theta_\ell\right)\right)} + \frac{\left|N^{\diamond}_{J^{\diamond}}\left(A_n\left(\Theta_\ell\right)\right) - N_J\left(A_n\left(\Theta_\ell\right)\right)\right|}{N^{o}\left(A_n\left(\Theta_\ell\right)\right)} \\
&\overset{\text{def}}{=} |G_1| + |G_2|
\end{aligned}$$

We prove that G_1 and G_2 go to 0 a.s. in the case where Assumption 4.4.2 item 3. is verified. The case where item 2. is satisfied is done easier following the same lines as in the proof of Lemma 4.5.2. Let $\varepsilon > 0$.

1 Let us start by proving the a.s. convergence to 0 of G_1.

$$
\begin{aligned}
&\mathbb{P}\left(|G_1| > \varepsilon\right) \\
&= \mathbb{P}\left(\frac{|N^{\diamond}\left(A_n\left(\Theta_\ell\right)\right) - N^{o}\left(A_n\left(\Theta_\ell\right)\right)|}{N^{o}\left(A_n\left(\Theta_\ell\right)\right)} > \varepsilon\right) \\
&= \mathbb{P}\left(|N^{\diamond}\left(A_n\left(\Theta_\ell\right)\right) - N^{o}\left(A_n\left(\Theta_\ell\right)\right)| > \varepsilon N^{o}\left(A_n\left(\Theta_\ell\right)\right), \exists \ell \setminus \left|N^{b}\left(A_n\left(\Theta_\ell\right)\right) - N^{o}\left(A_n\left(\Theta_\ell\right)\right)\right| > \lambda\right) \\
&\quad + \mathbb{P}\left(|N^{\diamond}\left(A_n\left(\Theta_\ell\right)\right) - N^{o}\left(A_n\left(\Theta_\ell\right)\right)| > \varepsilon N^{o}\left(A_n\left(\Theta_\ell\right)\right), \forall \ell \setminus \left|N^{b}\left(A_n\left(\Theta_\ell\right)\right) - N^{o}\left(A_n\left(\Theta_\ell\right)\right)\right| \leqslant \lambda\right)
\end{aligned}
$$

$$
\text{with } \lambda = \frac{\mathbb{E}\left[N^{b}\left(A_n\left(\Theta\right)\right)\right]}{4}
$$

$$
\begin{aligned}
&\leqslant k\mathbb{P}\left(\left|N^{b}\left(A_n\left(\Theta\right)\right) - N^{o}\left(A_n\left(\Theta\right)\right)\right| > \lambda\right) \\
&\quad + \mathbb{P}\left(|N^{\diamond}\left(A_n\left(\Theta_\ell\right)\right) - N^{o}\left(A_n\left(\Theta_\ell\right)\right)| > \varepsilon\left(N^{b}\left(A_n\left(\Theta_\ell\right)\right) - \lambda\right), N^{b}\left(A_n\left(\Theta_\ell\right)\right) > \delta\right) \\
&\quad + \mathbb{P}\left(|N^{\diamond}\left(A_n\left(\Theta_\ell\right)\right) - N^{o}\left(A_n\left(\Theta_\ell\right)\right)| > \varepsilon\left(N^{b}\left(A_n\left(\Theta_\ell\right)\right) - \lambda\right), N^{b}\left(A_n\left(\Theta_\ell\right)\right) \leqslant \delta\right)
\end{aligned}
$$

$$
\text{with } \delta = \frac{\mathbb{E}\left[N^{b}\left(A_n\left(\Theta\right)\right)\right]}{2}
$$

$$
\begin{aligned}
&\leqslant k\mathbb{P}\left(\left|N^{b}\left(A_n\left(\Theta\right)\right) - N^{o}\left(A_n\left(\Theta\right)\right)\right| > \lambda\right) + \mathbb{P}\left(|N^{\diamond}\left(A_n\left(\Theta_\ell\right)\right) - N^{o}\left(A_n\left(\Theta_\ell\right)\right)| > \varepsilon\lambda\right) \\
&\quad + \mathbb{P}\left(N^{b}\left(A_n\left(\Theta_\ell\right)\right) \leqslant \delta\right)
\end{aligned}
$$

The first two right-hand terms will be bounded by using the Vapnik-Chervonenkis' theory with the class $\mathcal{B} = \left\{\prod_{i=1}^{d}\left[a_i, b_i\right] : a_i, b_i \in \overline{\mathbb{R}}\right\}$. Let us start with the first right hand-term as follows

$$
\mathbb{P}\left(\left|N^{b}\left(A_n\left(\Theta\right)\right) - N^{o}\left(A_n\left(\Theta\right)\right)\right| > \lambda\right) \leqslant \mathbb{P}\left(\sup_{A\in\mathcal{B}}\left|\frac{1}{n}\sum_{j=1}^{n} B_j\left(\Theta^1, \mathcal{D}_n\right)\mathbb{1}_{\{\mathbf{X}^j\in A\}} - \frac{1}{n}\sum_{j=1}^{n}\mathbb{1}_{\{\mathbf{X}^j\in A\}}\right| > \frac{\lambda}{n}\right).
$$

This term is handled as in the proof of Lemma 4.5.3 by rewriting the bootstrap component thanks to the variables selected with replacement from the set $\mathcal{D}_n = \{(\mathbf{X}^1, Y^1), \ldots, (\mathbf{X}^n, Y^n)\}$ instead of the random vector $\left(B_j\left(\Theta^1, \mathcal{D}_n\right)\right)_{j=1,\ldots,n}$.

$$
\begin{aligned}
&\mathbb{P}\left(\sup_{A\in\mathcal{B}}\left|\frac{1}{n}\sum_{j=1}^{n} B_j\left(\Theta^1, \mathcal{D}_n\right)\mathbb{1}_{\{\mathbf{X}^j\in A\}} - \frac{1}{n}\sum_{j=1}^{n}\mathbb{1}_{\{\mathbf{X}^j\in A\}}\right| > \frac{\lambda}{n}\right) \\
&= \mathbb{E}\left[\mathbb{P}\left(\sup_{A\in\mathcal{B}}\left|\frac{1}{n}\sum_{j=1}^{n}\mathbb{1}_{\left\{Z_1^j\in A\right\}} - \mathbb{P}\left(Z_1 \in A|\, \mathcal{D}_n\right)\right| > \frac{\lambda}{n}\middle|\, \mathcal{D}_n\right)\right]
\end{aligned}
$$

By applying Vapnik-Chervonenkis' Theorem under the conditional distribution given $\mathcal{D}_n$, we get

$$\mathbb{P}\left(\sup_{A\in\mathcal{B}}\left|\frac{1}{n}\sum_{j=1}^{n}\mathbb{1}_{\left\{Z_1^j\in A\right\}}-\mathbb{P}\left(Z_1\in A|\,\mathcal{D}_n\right)\right|>\frac{\lambda}{n}\middle|\,\mathcal{D}_n\right)\leqslant 8(n+1)^{2d}e^{-\lambda^2/32n}\ .$$

Therefore,

$$\mathbb{P}\left(\sup_{A\in\mathcal{B}}\left|\frac{1}{n}\sum_{j=1}^{n}\mathbb{1}_{\left\{Z_1^j\in A\right\}}-\frac{1}{n}\sum_{j=1}^{n}\mathbb{1}_{\{\mathbf{X}^j\in A\}}\right|>\frac{\lambda}{n}\right)\leqslant 8(n+1)^{2d}e^{-\lambda^2/32n}\ .$$

Finally, we get the overall upper bound

$$\mathbb{P}\left(\left|N^b\left(A_n\left(\Theta\right)\right)-N^o\left(A_n\left(\Theta\right)\right)\right|>\lambda\right)\leqslant 8(n+1)^{2d}e^{-\lambda^2/32n}\ .$$

Regarding the second right hand-term, we have

$$\begin{aligned}\mathbb{P}\left(\left|N^\diamond\left(A_n\left(\Theta_\ell\right)\right)-N^o\left(A_n\left(\Theta_\ell\right)\right)\right|>\varepsilon\lambda\right)&=\mathbb{P}\left(\frac{\left|N^\diamond\left(A_n\left(\Theta_\ell\right)\right)-N^o\left(A_n\left(\Theta_\ell\right)\right)\right|}{n}>\frac{\varepsilon\lambda}{n}\right)\\&\leqslant\mathbb{P}\left(\sup_{A\in\mathcal{B}}\left|\frac{1}{n}\sum_{j=1}^{n}\mathbb{1}_{\{\mathbf{X}^{\diamond j}\in A\}}-\mathbb{P}_{\mathbf{X}}\left(\mathbf{X}\in A\right)\right|>\frac{\varepsilon\lambda}{2n}\right)\\&\quad+\mathbb{P}\left(\sup_{A\in\mathcal{B}}\left|\frac{1}{n}\sum_{j=1}^{n}\mathbb{1}_{\{\mathbf{X}^{j}\in A\}}-\mathbb{P}_{\mathbf{X}}\left(\mathbf{X}\in A\right)\right|>\frac{\varepsilon\lambda}{2n}\right)\\&\leqslant 16(n+1)^{2d}e^{-\varepsilon^2\lambda^2/128n}\ .\end{aligned}$$

Finally, using Equation (4.11) and Assumption 4.4.2, we have

$$\begin{aligned}\mathbb{P}\left(|G_1|>\varepsilon\right)&\leqslant 8Cn^{\alpha}(n+1)^{2d}\exp\left[-\frac{K^2\left(\ln\left(n\right)\right)^{2\beta}}{512}\right]\\&\quad+16(n+1)^{2d}\exp\left[-\frac{\varepsilon^2K^2\left(\ln\left(n\right)\right)^{2\beta}}{2048}\right]+\frac{4M^2}{n^{\alpha+1}\left(\ln\left(n\right)\right)^{\gamma}}\ .\end{aligned}$$

Then, thanks to Borel–Cantelli Lemma: $G_1\xrightarrow[n\to\infty]{a.s.}0$.

2 Now, consider the G_2 term,

$$\begin{aligned}&\mathbb{P}\left(|G_2|>\varepsilon\right)\\&=\mathbb{P}\left(\frac{\left|N_{J^\diamond}^{\diamond}\left(A_n\left(\Theta_\ell\right)\right)-N_J\left(A_n\left(\Theta_\ell\right)\right)\right|}{N^o\left(A_n\left(\Theta_\ell\right)\right)}>\varepsilon\right)\\&=\mathbb{P}\left(\left|N_{J^\diamond}^{\diamond}\left(A_n\left(\Theta_\ell\right)\right)-N_J\left(A_n\left(\Theta_\ell\right)\right)\right|>\varepsilon N^o\left(A_n\left(\Theta_\ell\right)\right),\exists\ell\setminus\left|N^b\left(A_n\left(\Theta_\ell\right)\right)-N^o\left(A_n\left(\Theta_\ell\right)\right)\right|>\lambda\right)\\&\quad+\mathbb{P}\left(\left|N_{J^\diamond}^{\diamond}\left(A_n\left(\Theta_\ell\right)\right)-N_J\left(A_n\left(\Theta_\ell\right)\right)\right|>\varepsilon N^o\left(A_n\left(\Theta_\ell\right)\right),\forall\ell\setminus\left|N^b\left(A_n\left(\Theta_\ell\right)\right)-N^o\left(A_n\left(\Theta_\ell\right)\right)\right|\leqslant\lambda\right)\end{aligned}$$

$$\text{where } \lambda = \frac{\mathbb{E}\left[N^b\left(A_n\left(\Theta\right)\right)\right]}{4}$$

$$\begin{aligned}
&\leqslant k\mathbb{P}\left(\left|N^b\left(A_n\left(\Theta\right)\right) - N^o\left(A_n\left(\Theta\right)\right)\right| > \lambda\right) \\
&\quad + \mathbb{P}\left(\left|N^{\diamond}_{J^{\diamond}}\left(A_n\left(\Theta_\ell\right)\right) - N_J\left(A_n\left(\Theta_\ell\right)\right)\right| > \varepsilon\left(N^b\left(A_n\left(\Theta_\ell\right)\right) - \lambda\right), N^b\left(A_n\left(\Theta_\ell\right)\right) > \delta\right) \\
&\quad + \mathbb{P}\left(\left|N^{\diamond}_{J^{\diamond}}\left(A_n\left(\Theta_\ell\right)\right) - N_J\left(A_n\left(\Theta_\ell\right)\right)\right| > \varepsilon\left(N^b\left(A_n\left(\Theta_\ell\right)\right) - \lambda\right), N^b\left(A_n\left(\Theta_\ell\right)\right) \leqslant \delta\right)
\end{aligned}$$

$$\text{where } \delta = \frac{\mathbb{E}\left[N^b\left(A_n\left(\Theta\right)\right)\right]}{2}$$

$$\begin{aligned}
&\leqslant k\mathbb{P}\left(\left|N^b\left(A_n\left(\Theta\right)\right) - N^o\left(A_n\left(\Theta\right)\right)\right| > \lambda\right) + \mathbb{P}\left(\left|N^{\diamond}_{J^{\diamond}}\left(A_n\left(\Theta_\ell\right)\right) - N_J\left(A_n\left(\Theta_\ell\right)\right)\right| > \varepsilon\lambda\right) \\
&\quad + \mathbb{P}\left(N^b\left(A_n\left(\Theta_\ell\right)\right) \leqslant \delta\right)
\end{aligned}$$

The middle term is treated as in Equation (4.12), we get

$$\mathbb{P}\left(\left|N^{\diamond}_{J^{\diamond}}\left(A_n\left(\Theta_\ell\right)\right) - N_J\left(A_n\left(\Theta_\ell\right)\right)\right| > \varepsilon\lambda\right) \leqslant 16(n+1)^{2d} e^{-\varepsilon^2\lambda^2/128n} \ .$$

As a consequence, Equation (4.11) and Assumption 4.4.2 give

$$\begin{aligned}
\mathbb{P}\left(|G_2| > \varepsilon\right) &\leqslant 8Cn^{\alpha}(n+1)^{2d}\exp\left[-\frac{K^2\left(\ln\left(n\right)\right)^{2\beta}}{512}\right] \\
&\quad + 16(n+1)^{2d}\exp\left[-\frac{\varepsilon^2K^2\left(\ln\left(n\right)\right)^{2\beta}}{2048}\right] + \frac{4M^2}{n^{\alpha+1}\left(\ln\left(n\right)\right)^{\gamma}} \ .
\end{aligned}$$

Thanks to Borel–Cantelli Lemma, we get $G_2 \xrightarrow[n\to\infty]{a.s.} 0$.

We conclude that G goes to 0 a.s. for all ℓ, thus

$$\forall \mathbf{x} \in \mathcal{X}, \forall y \in \mathbb{R}, \quad \left|F^{\diamond}_{k,n}\left(y|\,\mathbf{X} = \mathbf{x}\right) - F^{o}_{k,n}\left(y|\,\mathbf{X} = \mathbf{x}\right)\right| \xrightarrow[n\to\infty]{a.s.} 0 \ .$$

■

In order to illustrate the theoretical results, we provide a numerical example.

4.6 Numerical example

The convergence of the estimators, introduced in Section 4.3, is illustrated on the following toy example

$$Y = X_1 + X_2 + X_3 + \varepsilon \ , \tag{4.13}$$

where $\mathbf{X} = (X_1, X_2, X_3)$ are three independent random variables with $X_1 \sim GPD(1.5, 0.25)$ (a Generalised Pareto Distribution), $X_2 \sim \mathcal{LN}(1.1, 0.6)$ (a Log Normal Distribution), $X_3 \sim \Gamma(2, 0.6)$ (a Gamma Distribution) and ε is an independent centered Gaussian noise with variance $\sigma^2 = 4$.

The accuracy of the conditional distribution function estimators will be evaluated first, then that of the conditional quantile estimators.

4.6.1 Conditional distribution function

Let us start by assessing the performance of the C_CDF estimators using the Kolmogorov-Smirnov distance recalled below

$$KS(\mathbf{x}) = \max_y \left| \widehat{F}\left(y|\,\mathbf{x}\right) - F\left(y|\,\mathbf{x}\right) \right| \ ,$$

with $\widehat{F}\left(y|\,\mathbf{x}\right)$ being either $F^b_{k,n}\left(y|\,\mathbf{x}\right)$ or $F^o_{k,n}\left(y|\,\mathbf{x}\right)$ here. $F\left(y|\,\mathbf{x}\right)$ has the following expression in our example,

$$F\left(y|\,\mathbf{X} = \mathbf{x}\right) = \Phi\left(\frac{y - (x_1 + x_2 + x_3)}{\sigma}\right) \ ,$$

with Φ, the distribution function of the standard normal distribution $\mathcal{N}(0, 1)$.

For two randomly chosen points $\mathbf{x}$, the C_CDF estimates are built with a sample of size $n = 10^4$, a forest grown with $n_{trees} = 500$ and the minimum number of samples required to be at a leaf node set to $min\_samples\_leaf = \lfloor \sqrt{n} \cdot \log(n)^{1.5}/250 \rfloor$ for each tree. These experiments are replicated $s = 500$ times. Figure 4.1 below shows the C_CDF approximations computed with the estimator using the original dataset on the left-hand side and those of the estimator calculated with the bootstrap samples on the right-hand side. On each graph, the orange plain line is the true C_CDF, while the blue ones represent the 95% quantiles of the replications. Figure 4.1 therefore allows to display and compare approximated curves visually.

From a quantitative perspective, the quality of the estimators is measured at each point $\mathbf{x}$ using the following average Kolmogorov-Smirnov distance

$$\overline{KS}(\mathbf{x}) = \frac{1}{s} \sum_{j=1}^{s} KS_j(\mathbf{x}) \ .$$

According to the numerical results displayed in Figure 4.1, estimators perform well for points that are well represented in the training sample but, the performance decreases for extreme points. In order to reflect the overall performance of the estimators, we define in the sequel an averaged version of the previous metric and compute it with $p = 5 \times 10^4$ randomly chosen points $\mathbf{x}$

$$M\_\overline{KS} := \frac{1}{p} \sum_{j=1}^{p} \overline{KS}(\mathbf{x}^j) \ .$$

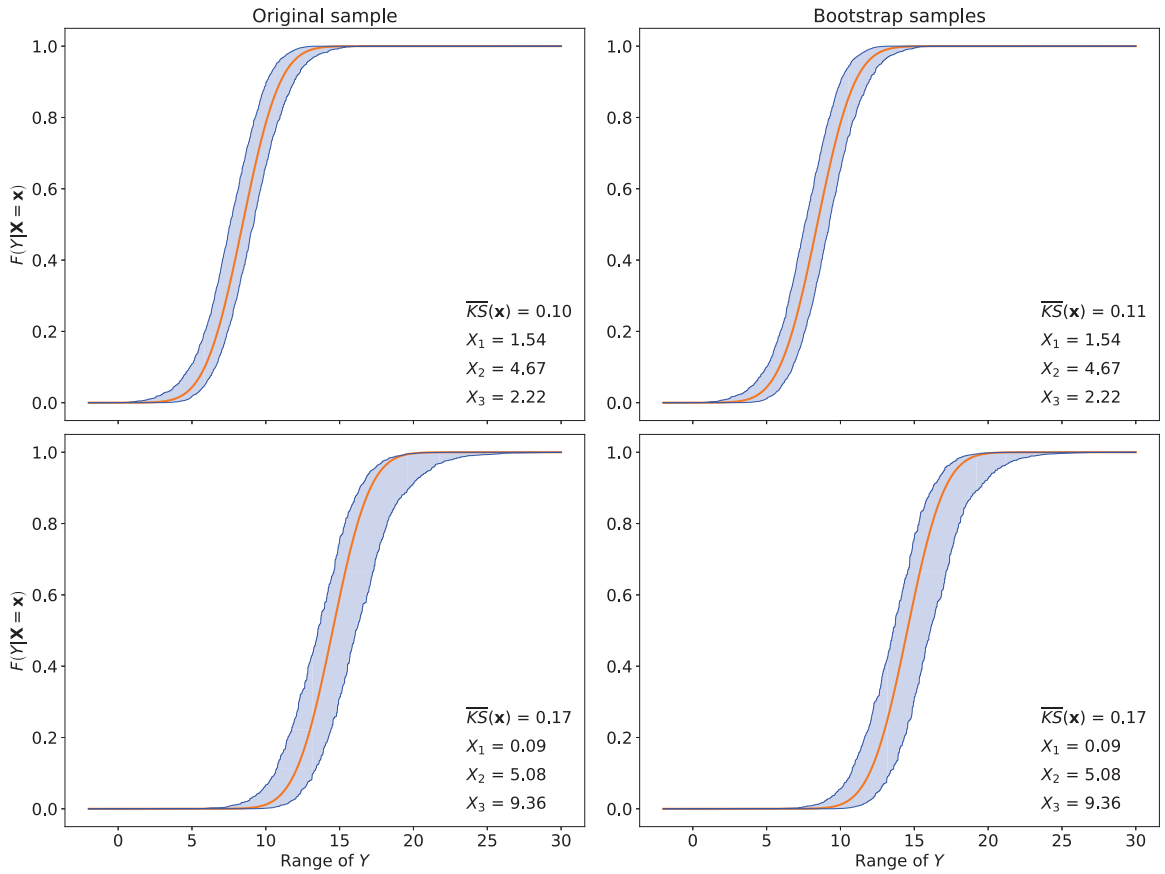


Figure 4.1 *Estimation of the conditional distribution function for two different values* $\mathbf{x}$ *by using the original sample (on the left side) and the bootstrap samples (on the right side). On each graph, the orange line is the true value along with the* 95% *confidence bands in blue.*

We get $M\_\overline{KS} = 0.1344$ for the estimator $F^b_{k,n}(y|\,\mathbf{x})$ and $M\_\overline{KS} = 0.1295$ for $F^o_{k,n}(y|\,\mathbf{x})$. Thus, it seems that both estimators have a good accuracy for estimating the C_CDF of most points $\mathbf{x}$.

Let us now assess the performance of the conditional quantile estimators.

4.6.2 Conditional quantiles

The analytic value of the α-quantile conditionally to $\mathbf{x} = (x_1, x_2, x_3)$ is easy to calculate,

$$q^\alpha(Y|\,\mathbf{x}) = x_1 + x_2 + x_3 + \sigma \times z_\alpha \ ,$$

with z_α, α-quantile of the standard normal distribution $\mathcal{N}(0,1)$.

Figure 4.2 shows for two specific points $\mathbf{x}$ and for several levels α ranging from 0.1 to 0.9, the distribution of the estimators of the conditional quantiles computed with the original dataset on the left-hand side and with the bootstrap samples on the right-hand side. The estimates have been calculated with the following setting: a sample of size $n = 10^4$, a forest grown with $n_{trees} = 500$ and the minimum number of samples required to be at a leaf node set to $min\_samples\_leaf = \lfloor \sqrt{n} \cdot \log(n)^{1.5}/250 \rfloor$. In order to assess the quality of the estimators at these points, the following indicators are computed by repeating the experiment $s = 500$ times.

$$
\begin{aligned}
RMSE\left(\mathbf{x}\right) &= \sqrt{\frac{1}{s}\sum_{j=1}^{s}\left(\hat{q}_j^{\alpha}\left(Y|\,\mathbf{x}\right) - q^{\alpha}\left(Y|\,\mathbf{x}\right)\right)^2}\,, \\
Bias\left(\mathbf{x}\right) &= \left|\frac{1}{s}\sum_{j=1}^{s}\hat{q}_j^{\alpha}\left(Y|\,\mathbf{x}\right) - q^{\alpha}\left(Y|\,\mathbf{x}\right)\right|\,, \\
Variance\left(\mathbf{x}\right) &= \frac{1}{s}\sum_{j=1}^{s}\left(\hat{q}_j^{\alpha}\left(Y|\,\mathbf{x}\right) - \frac{1}{s}\sum_{j=1}^{s}\hat{q}_j^{\alpha}\left(Y|\,\mathbf{x}\right)\right)^2\,,
\end{aligned}
$$

where $\hat{q}_j^{\alpha}\left(\mathbf{x}\right)$ is the estimator on the j's sample, $j = 1, \ldots, s$.

Based on the graphs obtained in Figure 4.2, it seems difficult to say if one estimator is better than the other. It appears that the performance of the two estimators differs a bit depending on the observation $\mathbf{x}$.

In order to get global measures of both estimators, we define an averaged version of the previous indicators computed with $p = 5 \times 10^4$ randomly chosen points $\mathbf{x}$ according to the following formulas

$$
\begin{aligned}
M\_RMSE &:= \frac{1}{p}\sum_{j=1}^{p} RMSE\left(\mathbf{x}^j\right)\,, \\
M\_Bias &:= \frac{1}{p}\sum_{j=1}^{p} Biais\left(\mathbf{x}^j\right)\,, \\
M\_Variance &:= \frac{1}{p}\sum_{j=1}^{p} Variance\left(\mathbf{x}^j\right)\,.
\end{aligned}
$$

By using the same setting as previously for the estimators, the numerical results for these three measures are listed in Table 4.1 for several levels α. First of all, both estimators have an equivalent RMSE, whereas the original dataset based estimator is the one with the smallest variance, which may seem surprising. Indeed, random forest method is an ensemble learning method that begins with bagging (the bootstrapped aggregation of regression tree predictions) in order to reduce the variance of the prediction function. Thus, for the particular case of the conditional quantile approximation, it is observed the opposite phenomenon on our example. In contrast, the bootstrap samples based estimator has the smallest bias for all α. The bias-variance tradeoff of both methods could be handled by tuning the hyperparameters of the random forest

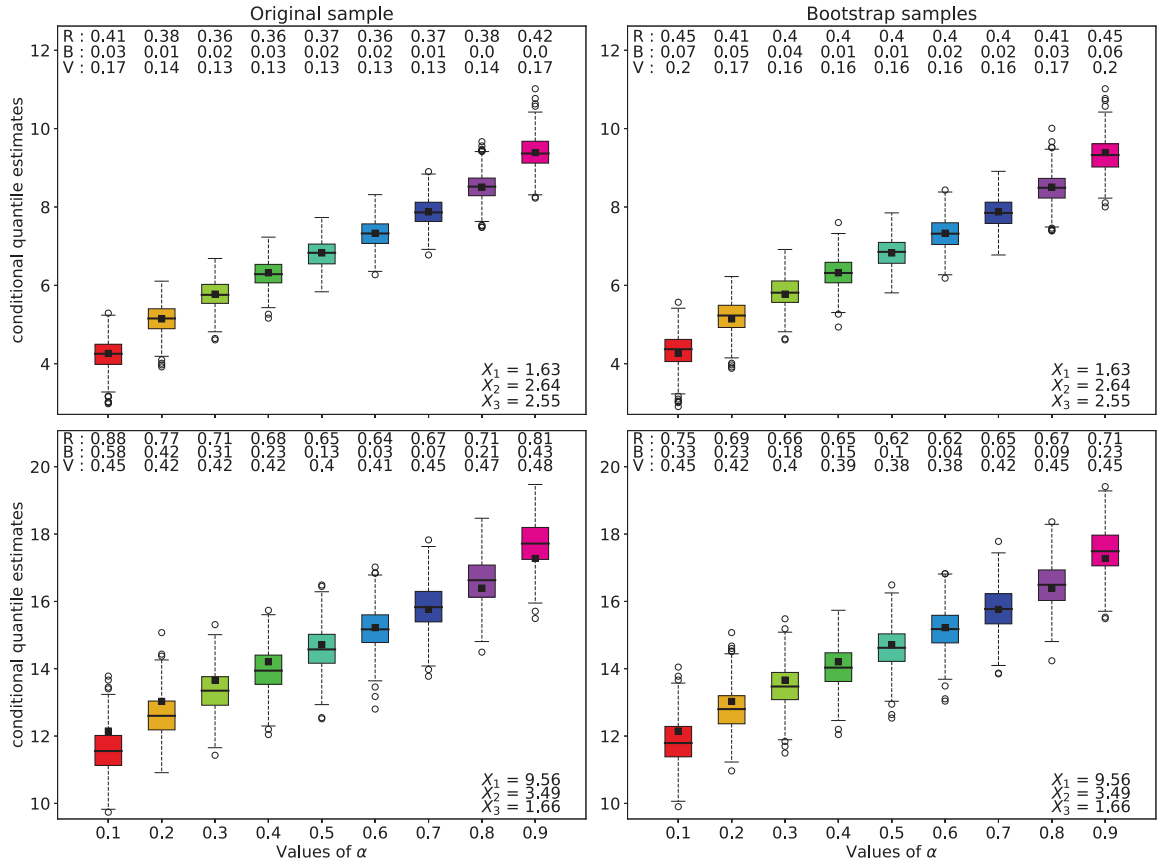


Figure 4.2 *Distribution of the conditional quantile approximations computed for three different values* $\mathbf{x}$ *by using the original sample (on the left side) and the bootstrap samples (on the right side). On each graph, the values above the boxplots are* `R` *for* $RMSE(\mathbf{x})$, `B` *for* $Bias(\mathbf{x})$ *and* `V` *for* $Variance(\mathbf{x})$.

such as *min_samples_leaf*, which would also allow to improve their accuracy. Finally, it has to be noted that the performance of the two estimators depends on the level α.

It is worth mentioning that Athey et al. [2019] propose in their paper another approach to estimate the conditional quantiles. The main differences between their method and ours are that it uses a splitting scheme tailored to the particular task at hand as well as honest and regular trees in the sense of Wager and Athey [2018]. As recalled in the introduction, honest trees randomly split the sample in half before constructing a tree, the first half is used when performing the splitting and the second half to populate the tree's leaf nodes. Honesty may hurt performance when working with very small datasets. Indeed, if the original sample is already small, honesty further cuts this one in half, so there may no longer be enough information to choose high-quality splits. Hence, it would be interesting to compare the performance of this method with ours at a later stage.

	Original sample			Bootstrap samples		
	M_RMSE	M_Bias	$M\_Variance$	M_RMSE	M_Bias	$M\_Variance$
$\alpha = 0.1$	0.6382	0.2382	0.2826	0.6410	0.1926	0.3115
$\alpha = 0.2$	0.5868	0.2011	0.2565	0.6008	0.1669	0.2846
$\alpha = 0.3$	0.5640	0.1791	0.2519	0.5837	0.1490	0.2791
$\alpha = 0.4$	0.5521	0.1638	0.2544	0.5748	0.1351	0.2808
$\alpha = 0.5$	0.5470	0.1530	0.2615	0.5714	0.1274	0.2874
$\alpha = 0.6$	0.5489	0.1482	0.2758	0.5736	0.1276	0.3010
$\alpha = 0.7$	0.5602	0.1530	0.3053	0.5836	0.1360	0.3298
$\alpha = 0.8$	0.5901	0.1766	0.3659	0.6085	0.1562	0.3890
$\alpha = 0.9$	0.6786	0.2443	0.6526	0.6842	0.2074	0.6837

Table 4.1 *Results of the averaged* $RMSE(\mathbf{x})$, $Bias(\mathbf{x})$ *and* $Variance(\mathbf{x})$ *computed over* $p = 5 \times 10^4$ *observations of* $\mathbf{x}$.

4.7 Conclusion

This article proposes two conditional distribution functions and conditional quantiles approximations based on random forests. The former is a natural generalisation of the random forest estimator of the regression function making use of the bootstrap samples, while the latter is based on a variant using only the original dataset.

The consistency of the bootstrap samples based estimator is shown under realistic assumptions and constitutes the major contribution of this paper. Indeed, this is the first consistency result handling the bootstrap component in a random forest method whereas it is usually replaced by subsampling. As for the second estimator, the consistency proof established in Meinshausen [2006] for a simplified random forest model is extended to a realistic one by taking into account all the randomness used in the procedure. The two estimators have close performances on our toy example. A specific interest of the bootsrap estimation is that the Out-Of-Bag samples could be used for cross-validation and / or back-testing procedures.

The estimators developed in this paper rest on trees grown with the CART-split criterion. But the assumptions providing the consistency results are detached from the split procedure used. Thus, the theoretical tools developed here could be useful for a large class of methods by just changing the splitting scheme. An ambitious additional work would be to develop a theoretical analysis for obtaining convergence rates and also to construct confidence intervals.

Acknowledgments

We are grateful to Andrés Cuberos, Ecaterina Nisipasu, Mathieu Poulin and Przemyslaw Sloma from SCOR for their valuable comments and support. We are also much indebted to Roland Denis and Benoit Fabrèges for intensive support on computational

aspects.

Chapter 5

Random Forest-based QOSA index estimation

This chapter will be submitted for publication to a journal in the near future.

Abstract
The standard quantitative methods of the Global Sensitivity Analysis of a numerical model, $Y = \eta(\mathbf{X})$ with d random inputs $\mathbf{X} = (X_1, \ldots, X_d)$, consist in quantifying the contributions of each of its input parameters in the variability of its output such as Sobol' indices or Shapley effects. These are useful tools of Sensitivity Analysis if one is interested in a particular characteristic of the Y distribution: the mean $\mathbb{E}[Y]$ of the numerical model. Indeed, they allow to quantify which variables have the greatest influence on the mean by using variance as a measure of distance. However, if we consider another characteristic of the Y distribution as quantity of interest, for example, a quantile of order alpha, it seems very intuitive to expect an extreme quantile to be sensitive to very different variables than the average. As a result, when the quantity of interest is the quantile of order alpha of the Y distribution, adapted indices named QOSA (Quantile Oriented Sensitivity Analysis) based on a contrast function were developed in order to determine the most influential variables. Estimators have been proposed to estimate the QOSA indices but these remain cumbersome to handle and are based on kernel estimation methods (problem with optimal bandwidth selection). We then propose new estimation methods based on the random forest allowing to estimate efficiently the QOSA indices.

Contents

5.1 Introduction

Nowadays, numerical models are ubiquitous in various fields, such as aerospace, economy, environment or insurance, they allow to approximate the behavior of physical phenomenon. Their main advantage is that they replace expensive, or even unachievable, real-life experiments and thus provide knowledge about the natural system. The extremely faithful representation of reality, made possible thanks to the increase in computing power, also explains this widespread use. However, this accuracy is often synonymous of complexity, ultimately leading to a difficult interpretation of the model. Besides, model inputs are usually uncertain due to a lack of information or the random nature of the factor, which means that the resulting output can be regarded as random. It is then important to assess the impact of this uncertainty on the model output. Global Sensitivity Analysis (GSA) methods solve these issues by *studying how the uncertainty in the output of a model can be apportioned to different sources of uncertainty in the model inputs* [Saltelli et al., 2004]. Hence, GSA allows to investigate input-ouput relationships by identifying the inputs that strongly influence the model response. Conversely, it may be of interest to see that although some inputs may not be very well established, they do not significantly contribute to output uncertainty.

Variance-based approaches are well-established and widely used for GSA. Among them, the sensitivity indices developed by Sobol [1993] are very popular. This last method stands on the assumption that the inputs are independent. Under this hypothesis, the overall variance of a scalar output can be split down into different partial variances using the so-called Hoeffding [1948] decomposition. Then, the first-order Sobol' index quantifies the individual contribution of an input to the output variance while the total Sobol' index [Jansen et al., 1994; Homma and Saltelli, 1996] measures the marginal and interaction effects. However, even if they are extremely popular and informative measures, variance-based approaches suffer from theoretical limitation. Indeed, by definition, they study only the impact of the inputs on the expectation of the output by considering the variance as distance measure. Since the expectation of

the output distribution is not always the quantity of interest, this measure turns out to be unsuitable in many cases studies.

A new class of sensitivity indices, generalizing the first-order Sobol' index to other quantities of interest than the expectation, has been introduced in Fort et al. [2016]. These indices called Goal Oriented Sensitivity Analysis (GOSA) compare the minimum of a specific contrast function to its conditional counterpart when one of the inputs is fixed. The unconditional minimum being reached by the quantity of interest (for example a quantile). A list of cost or constrast functions, allowing to quantify how an input distribution affects a particular feature of the output distribution, can be found in Fort et al. [2016].

In this paper, we focus on Quantile Oriented Sensitivity Analysis (QOSA) measuring the impact of the inputs on the α-quantile of the output distribution. Browne et al. [2017]; Maume-Deschamps and Niang [2018] introduced a statistical estimator of the first-order QOSA index based on a kernel approach. Lastly, Kala [2019] defined the second and higher order QOSA indices as well as a variance-like decomposition for quantiles in the case of independent inputs.
Despite these recents works, the question of the effective estimation of the first-order QOSA index remains open. Indeed, it turns out to be difficult to compute in practice because it requires an accurate estimate of either the conditional quantile of the output given an input, or the minimum of a conditional expectation of the output given an input. Kala [2019] handles this feature with a brute force Monte-Carlo approach. As a matter of fact, for each value of an input, realizations of the other inputs are generated conditionally to the fixed value. Therefore, in this approach, the dependency structure of inputs has to be known, which is not always the case. Besides, the computational cost is too high to consider its use in an industrial context when dealing with costly models. Browne et al. [2017]; Maume-Deschamps and Niang [2018] developed kernel-based estimators to avoid this double-loop issue. But, when using a small dataset, their performance is highly dependent of the bandwidth parameter. Browne et al. [2017] proposed a cumbersome algorithm for setting an efficient bandwidth that is not straighforward to implement in practice. As for the estimator of Maume-Deschamps and Niang [2018], a large dataset is needed in order to have a low estimation error, as no algorithm of bandwidth parameter selection is established.

To overcome these issues, we explore the random forest algorithm introduced by Breiman [2001] in order to estimate the conditional distribution of the output given an input. The main contribution of this paper is to provide different estimation strategies of the first-order QOSA index based on this method.

The paper is organized as follows. We recall in Section 5.2 the definition of the first-order QOSA index and initiate the estimation process. Section 5.3 presents the random forest algorithm and several estimators based on this method are described in Section 5.4. The entire process is summarized in Section 5.5 and the performance of the estimators is investigated in Section 5.6 on simulated data. Finally, a conclusion is given in Section 5.7.

5.2 Estimation of the QOSA index

Let us consider the input-output system where $\mathbf{X} = (X_1, \ldots, X_d) \in \mathbb{R}^d$ is a random vector of d independent inputs and $Y = f(\mathbf{X})$ is the output random variable of a measurable deterministic function $f : \mathbb{R}^d \to \mathbb{R}$ which can be a mathematical function or a computational code. Then, given a level $\alpha \in]0,1[$, Fort et al. [2016] introduced the first-order Quantile Oriented Sensitivity Analysis (QOSA) index, related to the input X_i, as

$$S_i^\alpha = \frac{\min_{\theta\in\mathbb{R}} \mathbb{E}\left[\psi_\alpha(Y,\theta)\right] - \mathbb{E}\left[\min_{\theta\in\mathbb{R}} \mathbb{E}\left[\psi_\alpha(Y,\theta)|\, X_i\right]\right]}{\min_{\theta\in\mathbb{R}} \mathbb{E}\left[\psi_\alpha(Y,\theta)\right]}\,, \tag{5.1}$$

with the contrast function $\psi_\alpha : (y,\theta) \mapsto (y-\theta)\left(\alpha - \mathbb{1}_{\{y\leqslant\theta\}}\right)$. This function, also called *pinball loss* or *check function* in the literature is the cornerstone of the quantile regression [Koenker and Hallock, 2001]. Quantile and conditional quantile are related to this loss function as follows

$$q^\alpha(Y) = \arg\min_{\theta\in\mathbb{R}} \mathbb{E}\left[\psi_\alpha(Y,\theta)\right] \quad \text{and} \quad q^\alpha(Y|\, X_i) = \arg\min_{\theta\in\mathbb{R}} \mathbb{E}\left[\psi_\alpha(Y,\theta)|\, X_i\right]\,,$$

where $q^\alpha(Y)$ is the α-quantile of Y and $q^\alpha(Y|\, X_i)$, the α-quantile of Y given X_i. Thus, the index S_i^α can be rewritten in the following way,

$$S_i^\alpha = 1 - \frac{\mathbb{E}\left[\min_{\theta\in\mathbb{R}} \mathbb{E}\left[\psi_\alpha(Y,\theta)|\, X_i\right]\right]}{\min_{\theta\in\mathbb{R}} \mathbb{E}\left[\psi_\alpha(Y,\theta)\right]} = 1 - \frac{\mathbb{E}\left[\psi_\alpha(Y, q^\alpha(Y|\, X_i))\right]}{\mathbb{E}\left[\psi_\alpha(Y, q^\alpha(Y))\right]} = 1 - \frac{O}{P}\,, \tag{5.2}$$

where O refers to $\mathbb{E}\left[\min_{\theta\in\mathbb{R}} \mathbb{E}\left[\psi_\alpha(Y,\theta)|\, X_i\right]\right] = \mathbb{E}\left[\psi_\alpha(Y, q^\alpha(Y|\, X_i))\right]$ and P, to $\min_{\theta\in\mathbb{R}} \mathbb{E}\left[\psi_\alpha(Y,\theta)\right] = \mathbb{E}\left[\psi_\alpha(Y, q^\alpha(Y))\right]$.

Hence, as stated in Browne et al. [2017], the index S_i^α compares the mean distance between Y and its conditional quantile to the mean distance between Y and its quantile, where the pinball loss function ψ_α is the considered distance. This index has some basic properties requested for a reasonable sensitivity index such as $0 \leqslant S_i^\alpha \leqslant 1$, $S_i^\alpha = 0$ if Y is independent of X_i and $S_i^\alpha = 1$ if Y is X_i measurable.

It should be mentioned that Kucherenko et al. [2019] proposed new indices to assess the impact of inputs on the α-quantile of the output distribution. They directly quantify the mean distance between quantiles $q^\alpha(Y)$ and $q^\alpha(Y|\, X_i)$ rather than the mean distance between average contrast functions like in the first-order QOSA index. Different estimation strategies are investigated in their paper (brute force Monte Carlo and double-loop reordering approach). But a major limitation is that a large sample size is required to get an accurate computation of the index (samples of size 2^{18} are used in their paper).

Let us now initiate the estimation procedure for the first-order QOSA index S_i^α, associated to a specific input X_i and a level α.

We consider an i.i.d n-sample $\mathcal{D}_n^\diamond = \left(\mathbf{X}^{\diamond j}, Y^{\diamond j}\right)_{j=1,\dots,n}$ such that $Y^{\diamond j} = f\left(\mathbf{X}^{\diamond j}\right), j = 1,\dots,n$. Then, a first natural estimator of the P term of the QOSA index based on the quantity $\mathbb{E}\left[\psi_\alpha\left(Y, q^\alpha\left(Y\right)\right)\right]$ is proposed

$$\widehat{P}_1 = \frac{1}{n}\sum_{j=1}^{n} \psi_\alpha\left(Y^{\diamond j}, \widehat{q}^\alpha(Y)\right) \ , \tag{5.3}$$

with $\widehat{q}^\alpha(Y)$, the classical empirical estimator for $q^\alpha\left(Y\right)$ obtained from $\mathcal{D}_n^\diamond$.
By using the other quantity $\min_{\theta\in\mathbb{R}} \mathbb{E}\left[\psi_\alpha\left(Y,\theta\right)\right]$, the P term can be estimated as follows

$$\widehat{P}_2 = \min_{\theta\in\mathbb{R}} \frac{1}{n}\sum_{j=1}^{n} \psi_\alpha\left(Y^{\diamond j}, \theta\right) \ , \tag{5.4}$$

where the minimum is reached for one of the elements of $\left(Y^{\diamond j}\right)_{j=1,\dots,n}$. As the function to minimize is decreasing then increasing, this estimator therefore requires to compute $\frac{1}{n}\sum_{j=1}^{n} \psi_\alpha\left(Y^{\diamond j}, Y^{\diamond(k)}\right)$, $k = 1,\dots,n$, until it increases, with $Y^{\diamond(k)}$ the order statistics of $(Y^{\diamond 1},\dots,Y^{\diamond n})$. This process is much more time-consuming than the first estimator where we just need to compute the quantile and then plug it. Thus, in the sequel, we are going to use the $\widehat{P}_1$ estimator.

The O term of the QOSA index is trickier to estimate because a good approximation of the conditional distribution of Y given X_i is necessary. Both existing estimators of the QOSA index currently provided in Browne et al. [2017]; Maume-Deschamps and Niang [2018] handle this feature thanks to kernel-based methods. But in practice, with these methods, we are faced with determining the optimal bandwidth parameter or using large sample sizes in order to have a sufficiently low estimation error when employing a non optimal bandwidth. Thus, when dealing with costly computational models, a precise enough estimation of these indices can be difficult to achieve or even unfeasible.

We propose in this paper to address these issues by using the random forest method for estimating the conditional distribution. Therefore, several statistical estimators for the O term of the first-order QOSA index will be defined in Section 5.4. Let us first recall the random forest algorithm.

5.3 Random forests

Random forests are ensemble learning methods, first introduced by Breiman [2001], which can be used in classification or regression problems. We only focus on their use for regression task and assume to be given a training sample $\mathcal{D}_n = \left(\mathbf{X}^j, Y^j\right)_{j=1,\dots,n}$ of i.i.d random variables distributed as the prototype pair $(\mathbf{X}, Y)$.

Breiman's forest growns a collection of k regression trees based on the CART procedure described in Breiman et al. [1984]. Building several different trees from a

single dataset requires to randomize the tree building process. Randomness injected in each tree is denoted by Θ_ℓ where $(\Theta_\ell)_{\ell=1,\dots,k}$ are independent random variables distributed as Θ (independent of $\mathcal{D}_n$). $\Theta = (\Theta_1, \Theta_2)$ contains indices of observations selected to build the tree and indices of splitting candidate directions in each cell.

In more detail, the ℓ-th tree is built using a bootstrap sample $\mathcal{D}_n^\star(\Theta_\ell)$ from the original dataset. Only these observations are used to construct the tree and to make the tree prediction. Once the observations have been selected, the algorithm forms a recursive partitioning of the input space. In each cell, a number $max\_features$ of variables is selected uniformly at random among all inputs. Then, the best split is chosen as the one optimizing the CART splitting criterion only along the $max\_features$ preselected directions. This process is repeated in each cell. A stopping criterion, often implemented, is that a split point at any depth will only be considered if it leaves at least $min\_samples\_leaf$ samples in each of the left and right child nodes. After tree partition has been completed, the prediction of the ℓ-th tree denoted by $m_n^b(\mathbf{x}; \Theta_\ell, \mathcal{D}_n)$ at a new point $\mathbf{x}$ is computed by averaging the $N_n^b(\mathbf{x}; \Theta_\ell, \mathcal{D}_n)$ observations falling into the cell $A_n(\mathbf{x}; \Theta_\ell, \mathcal{D}_n)$ of the new point.
Hence, the random forest prediction is the average of the k predicted values:

$$m_{k,n}^b(\mathbf{x}; \Theta_1, \dots, \Theta_k, \mathcal{D}_n) = \frac{1}{k}\sum_{\ell=1}^{k} m_n^b(\mathbf{x}; \Theta_\ell, \mathcal{D}_n) = \frac{1}{k}\sum_{\ell=1}^{k}\left(\sum_{j\in\mathcal{D}_n^\star(\Theta_\ell)} \frac{\mathbb{1}_{\{\mathbf{X}^j\in A_n(\mathbf{x};\Theta_\ell,\mathcal{D}_n)\}}}{N_n^b(\mathbf{x}; \Theta_\ell, \mathcal{D}_n)} Y^j\right) . \tag{5.5}$$

By defining the random variable $B_j(\Theta_\ell^1, \mathcal{D}_n)$ as the number of times that the observation $(\mathbf{X}^j, Y^j)$ has been used from the original dataset for the ℓ-th tree construction, the conditional mean estimator in Equation (5.5) is rewritten as follows

$$m_{k,n}^b(\mathbf{x}; \Theta_1, \dots, \Theta_k, \mathcal{D}_n) = \sum_{j=1}^{n} w_{n,j}^b(\mathbf{x}; \Theta_1, \dots, \Theta_k, \mathcal{D}_n)\, Y^j , \tag{5.6}$$

where the weights $w_{n,j}^b(\mathbf{x}; \Theta_1, \dots, \Theta_k, \mathcal{D}_n)$ are defined by

$$w_{n,j}^b(\mathbf{x}; \Theta_1, \dots, \Theta_k, \mathcal{D}_n) = \frac{1}{k}\sum_{\ell=1}^{k} \frac{B_j(\Theta_\ell^1, \mathcal{D}_n)\, \mathbb{1}_{\{\mathbf{X}^j\in A_n(\mathbf{x};\Theta_\ell,\mathcal{D}_n)\}}}{N_n^b(\mathbf{x}; \Theta_\ell, \mathcal{D}_n)} . \tag{5.7}$$

A variant of the Equation (5.6) provides another estimation of the conditional mean. Trees are still grown as in the standard random forest algorithm being based on the bootstrap samples but, for the tree prediction, the original dataset $\mathcal{D}_n$ is used instead of the bootstrap sample $\mathcal{D}_n^\star(\Theta_\ell)$ associated to the ℓ-th tree and we get

$$m_{k,n}^o(\mathbf{x}; \Theta_1, \dots, \Theta_k, \mathcal{D}_n) = \sum_{j=1}^{n} w_{n,j}^o(\mathbf{x}; \Theta_1, \dots, \Theta_k, \mathcal{D}_n)\, Y^j , \tag{5.8}$$

where the weights $w_{n,j}^o(\mathbf{x}; \Theta_1, \dots, \Theta_k, \mathcal{D}_n)$ are defined by

$$w_{n,j}^o(\mathbf{x}; \Theta_1, \dots, \Theta_k, \mathcal{D}_n) = \frac{1}{k}\sum_{\ell=1}^{k} \frac{\mathbb{1}_{\{\mathbf{X}^j\in A_n(\mathbf{x};\Theta_\ell,\mathcal{D}_n)\}}}{N_n^o(\mathbf{x}; \Theta_\ell, \mathcal{D}_n)} . \tag{5.9}$$

It has to be noted that contrary to Equation (5.7) where $N_n^b(\mathbf{x};\Theta_\ell,\mathcal{D}_n)$ refers to the number of elements of $\mathcal{D}_n^\star(\Theta_\ell)$ falling into $A_n(\mathbf{x};\Theta_\ell,\mathcal{D}_n)$, in Equation (5.9), $N_n^o(\mathbf{x};\Theta_\ell,\mathcal{D}_n)$ is the number of elements of $\mathcal{D}_n$ that fall into $A_n(\mathbf{x};\Theta_\ell,\mathcal{D}_n)$.

Thus, both weighted approaches using, either the bootstrap samples (Equation (5.6)) or the original dataset (Equation (5.8)), allow to see the random forest method as a local averaging estimate [Lin and Jeon, 2006; Scornet, 2016c] and will be at the heart of the strategies proposed for estimating the O term of the QOSA index. In the following, to lighten notation we will omit the dependence to Θ and $\mathcal{D}_n$ in the weights.

5.4 Estimation of the O term of the QOSA index

By using the random forest method aforementioned, ten estimators of the O term may be defined. The first four rely on the expression $\mathbb{E}[\psi_\alpha(Y,q^\alpha(Y|X_i))]$ and the others on $\mathbb{E}\left[\min_{\theta\in\mathbb{R}}\mathbb{E}[\psi_\alpha(Y,\theta)|X_i]\right]$.

5.4.1 Quantile-based O term estimators

In this section, the estimations of the O term of the QOSA index are based on the quantity $\mathbb{E}[\psi_\alpha(Y,q^\alpha(Y|X_i))]$. Using two training samples denoted by $\mathcal{D}_n^\diamond = (\mathbf{X}^{\diamond j},Y^{\diamond j})_{j=1,\dots,n}$ and $\mathcal{D}_n=(\mathbf{X}^j,Y^j)_{j=1,\dots,n}$, we define

$$\widehat{R}_i=\frac{1}{n}\sum_{j=1}^{n}\psi_\alpha\left(Y^{\diamond j},\widehat{q}^\alpha\left(Y|X_i=X_i^{\diamond j}\right)\right)\ ,$$

where the sample $\mathcal{D}_n$ is used to get $\widehat{q}^\alpha(Y|X_i=x_i)$, an estimator of the conditional quantile $q^\alpha(Y|X_i=x_i)$. It is obtained thanks to two approaches based on the random forests, described in the sequel.

5.4.1.1 Quantile estimation with a weighted approach

The method proposed in this subsection rests on a random forest built with the observations $\mathcal{D}_n^i=\left(X_i^j,Y^j\right)_{j=1,\dots,n}$ from $\mathcal{D}_n$, i.e., by explaining Y with only X_i. Then, the estimator of the Conditional Cumulative Distribution Function (C_CDF) introduced in Elie-Dit-Cosaque and Maume-Deschamps [2020] and recalled below is used to estimate the conditional quantile,

$$F_{k,n}^b(y|X_i=x_i)=\sum_{j=1}^{n}w_{n,j}^b(x_i)\,\mathbb{1}_{\{Y^j\leqslant y\}}\ , \tag{5.10}$$

where the $w_{n,j}^{b}(x_i)$'s are defined in Equation (5.7).

Hence, given a level $\alpha \in [0,1]$, the conditional quantile estimator $\widehat{q}^{\alpha}(Y|X_i = x_i)$ is defined as follows

$$\widehat{q}^{\alpha}(Y|X_i = x_i) = \inf\left\{Y^p,\ p = 1,\dots,n : F_{k,n}^{b}(Y^p|X_i = x_i) \geqslant \alpha\right\} .$$

As a result, the estimator of $\mathbb{E}[\psi_{\alpha}(Y, q^{\alpha}(Y|X_i))]$ based on this method is denoted $\widehat{R}_i^{1,b}$.

Another estimator of the C_CDF can be achieved by just replacing the weights $w_{n,j}^{b}(x_i)$ based on the bootstrap samples of the forest by those using the original dataset $w_{n,j}^{o}(x_i)$ provided in Equation (5.9). That gives the following estimator which has been proposed in Meinshausen [2006],

$$F_{k,n}^{o}(y|X_i = x_i) = \sum_{j=1}^{n} w_{n,j}^{o}(x_i)\,\mathbb{1}_{\{Y^j \leqslant y\}} .$$

The conditional quantiles are then estimated by plugging $F_{k,n}^{o}(y|X_i = x_i)$ instead of $F(Y|X_i = x_i)$. Accordingly, the associated estimator of $\mathbb{E}[\psi_{\alpha}(Y, q^{\alpha}(Y|X_i))]$ based on these weights is denoted $\widehat{R}_i^{1,o}$.

5.4.1.2 Quantile estimation within a leaf

At first, we grow a set of k trees indexed by $\ell = 1,\dots,k$ using the sample $\mathcal{D}_n^i$ and for each tree, let $\mathcal{L}_{\ell}^{b}(x_i)$ be the set of the observations of the bootstrap sample $\mathcal{D}_n^{\star}(\Theta_{\ell})$ used for the ℓ-th tree construction falling into the same leaf node as the observation x_i.

For the ℓ-th tree, the estimator $\widehat{q}_{\ell}^{b,\alpha}(Y|X_i = x_i)$ of $q^{\alpha}(Y|X_i = x_i)$ is obtained with the observations in $\mathcal{L}_{\ell}^{b}(x_i)$ as follows

$$\widehat{q}_{\ell}^{b,\alpha}(Y|X_i = x_i) = \inf\left\{Y^p,\ p = 1,\dots,|\mathcal{L}_{\ell}^{b}(x_i)| : \sum_{j \in \mathcal{L}_{\ell}^{b}(x_i)} \frac{\mathbb{1}_{\{Y^j \leqslant Y^p\}}}{|\mathcal{L}_{\ell}^{b}(x_i)|} \geqslant \alpha\right\} .$$

The values from the k randomized trees are then agregated to obtain the following random forest estimate

$$\widehat{q}^{\alpha}(Y|X_i = x_i) = \frac{1}{k}\sum_{\ell=1}^{k} \widehat{q}_{\ell}^{b,\alpha}(Y|X_i = x_i) .$$

As for the conditional mean estimate defined in Section 5.3 or for the C_CDF approximation introduced in Subsection 5.4.1.1, we can provide a variant using the original sample. Thus, once the forest constructed with the bootstrap samples, let

$\mathcal{L}_\ell^o(x_i)$ be the set of the original observations from $\mathcal{D}_n^i$ falling in the same leaf node as x_i for the ℓ-th tree and define

$$\widehat{q}^{\alpha}(Y|X_i = x_i) = \frac{1}{k}\sum_{\ell=1}^{k} \widehat{q}_\ell^{o,\alpha}(Y|X_i = x_i) \ ,$$

where $\widehat{q}_\ell^{o,\alpha}(Y|X_i = x_i)$ is computed with the observations in $\mathcal{L}_\ell^o(x_i)$ for each tree,

$$\widehat{q}_\ell^{o,\alpha}(Y|X_i = x_i) = \inf\left\{ Y^p,\ p = 1,\dots,|\mathcal{L}_\ell^o(x_i)| : \sum_{j\in\mathcal{L}_\ell^o(x_i)} \frac{\mathbb{1}_{\{Y^j \leqslant Y^p\}}}{|\mathcal{L}_\ell^o(x_i)|} \geqslant \alpha \right\} \ .$$

Thus, these two methods allow us to propose the following estimator $\widehat{R}_i^{2,b}$ (resp. $\widehat{R}_i^{2,o}$) of $\mathbb{E}[\psi_\alpha(Y, q^\alpha(Y|X_i))]$ using the bootstrap samples (resp. the original sample).

5.4.2 Minimum-based O term estimators

The estimators developped in Subsection 5.4.1, based on $\mathbb{E}[\psi_\alpha(Y, q^\alpha(Y|X_i))]$, require to approximate the conditional quantile and then plug it to estimate the O term. As mentioned before, the model f could be time-consuming. Therefore, they may be inappropriate as two training samples are necessary. Hence, we propose in this part to develop estimators of the O term taking advantage from the expression $\mathbb{E}\left[\min_{\theta\in\mathbb{R}} \mathbb{E}[\psi_\alpha(Y,\theta)|X_i]\right]$ for which we just need to find the minimum instead of plugging the quantile.

5.4.2.1 Minimum estimation with a weighted approach

First of all, a random forest is built with the observations $\mathcal{D}_n^i$. Then, by considering an additional sample $(\mathbf{X}^{\diamond j})_{j=1,\dots,n}$ independent of $\mathcal{D}_n$, the O term may be estimated as follows

$$\widehat{Q}_i^{1,b} = \frac{1}{n}\sum_{m=1}^{n} \min_{p=1,\dots,n} \sum_{j=1}^{n} w_{n,j}^b(X_i^{\diamond m})\,\psi_\alpha\left(Y^j, Y^p\right) \ .$$

Let us notice that the conditional expectation $\mathbb{E}[\psi_\alpha(Y,\theta)|X_i = x_i]$ is estimated with $\sum_{j=1}^{n} w_{n,j}^b(x_i)\,\psi_\alpha(Y^j,\theta)$ whose the minimum is reached for θ equals one of the elements of $(Y^j)_{j=1,\dots,n}$.

The estimator provided above is still valid by just replacing the weigths $w_{n,j}^b(x_i)$ by the other version $w_{n,j}^o(x_i)$ presented in Equation (5.9) using the original dataset. Hence, we get another estimator of the O term denoted by $\widehat{Q}_i^{1,o}$.

5.4.2.2 Minimum estimation within a leaf

In this subsection, we are going to take advantage of the tree structure in order to propose a new estimator. To begin with, let us consider that a random forest is built with the observations $\mathcal{D}_n^i$.

Then, the key point is that an additional sample is no longer required in order to process the outer expectation of the O term. Indeed, for the ℓ-th tree, the observations falling into in its m-th leaf node approximate the conditional distribution of Y given a certain point $X_i = x_i$, which allows to estimate the minimum of the conditional expectation $\min_{\theta\in\mathbb{R}} \mathbb{E}\left[\psi_\alpha\left(Y,\theta\right)\middle| X_i = x_i\right]$. Furthermore, by noting that leaves are distributed according to the distribution of X_i, we can therefore make the average over them to deal with the outer expectation. Thus, by noting N_{leaves}^{ℓ} as the number of leaves of the ℓ-th tree and $\mathcal{L}_{\ell,m}^{b}$ as the set of the observations of the bootstrap sample $\mathcal{D}_n^{i\star}\left(\Theta_\ell\right)$ used for the ℓ-th tree construction falling into the m-th leaf node of this one, then we can define the following tree estimate for the O term

$$\frac{1}{N_{leaves}^{\ell}} \sum_{m=1}^{N_{leaves}^{\ell}} \left(\min_{p\in\mathcal{L}_{\ell,m}^{b}} \sum_{j\in\mathcal{L}_{\ell,m}^{b}} \frac{\psi_\alpha\left(Y^j, Y^p\right)}{|\mathcal{L}_{\ell,m}^{b}|} \right) .$$

The approximations of the k randomized trees are then averaged to obtain the following random forest estimate

$$\widehat{Q}_i^{2,b} = \frac{1}{k}\sum_{\ell=1}^{k} \left[\frac{1}{N_{leaves}^{\ell}} \sum_{m=1}^{N_{leaves}^{\ell}} \left(\min_{p\in\mathcal{L}_{\ell,m}^{b}} \sum_{j\in\mathcal{L}_{\ell,m}^{b}} \frac{\psi_\alpha\left(Y^j, Y^p\right)}{|\mathcal{L}_{\ell,m}^{b}|} \right) \right] .$$

As before, the entire procedure described above is still valid by just replacing $\mathcal{L}_{\ell,m}^{b}$ by $\mathcal{L}_{\ell,m}^{o}$ which is the set of the original training examples falling in the m-th leaf node of the ℓ-th tree. Thanks to this change, we get another estimator of the O term denoted by $\widehat{Q}_i^{2,o}$. It should be noted that looking for the minimum in the leaves directly implies that they are sufficiently sampled for the method to be valid.

5.4.2.3 Minimum estimation with a weighted approach and complete trees

In Subsections 5.4.2.1 and 5.4.2.2, the conditional distribution of Y given X_i is obtained from trees grown with $\mathcal{D}_n^i$. Instead of using this approach, we propose in this part to build a forest with complete trees, i.e. grown with all the model's inputs and then adjust the weights to recover the conditional expectation $\mathbb{E}\left[\psi_\alpha\left(Y,\theta\right)\middle| X_i\right]$.

Thus, as noticed, a full random forest is first constructed with the whole dataset $\mathcal{D}_n$. Then, by using an additional sample $\left(\mathbf{X}^{\diamond j}\right)_{j=1,\ldots,n}$ independent of $\mathcal{D}_n$, the conditional

expectation $\mathbb{E}\left[\psi_\alpha\left(Y,\theta\right)\middle|\, X_i = x_i\right]$ is estimated as follows

$$\begin{aligned}\mathbb{E}\left[\psi_\alpha\left(Y,\theta\right)\middle|\, X_i = x_i\right] &= \mathbb{E}\left[\mathbb{E}\left[\psi_\alpha\left(Y,\theta\right)\middle|\, X_1,\ldots,x_i,\ldots,X_d\right]\middle|\, X_i = x_i\right] \\ &\approx \frac{1}{n}\sum_{l=1}^{n}\left(\sum_{j=1}^{n} w_{n,j}^{b}\left(\left(\mathbf{X}_1^{\diamond\ell},\ldots,\mathbf{X}_{i-1}^{\diamond\ell},x_i,\mathbf{X}_{i+1}^{\diamond\ell},\ldots,\mathbf{X}_d^{\diamond\ell}\right)\right)\right) \\ &\approx \sum_{j=1}^{n} w_{n,j}^{b,i}\left(x_i\right)\psi_\alpha\left(Y^j,\theta\right)\ ,\end{aligned}$$

where the suitable weights $w_{n,j}^{b,i}\left(x_i\right)$ are defined by

$$w_{n,j}^{b,i}\left(x_i\right) = \frac{1}{n}\sum_{\ell=1}^{n} w_{n,j}^{b}\left(\left(\mathbf{X}_{-i}^{\diamond\ell},x_i\right)\right)\ . \tag{5.11}$$

The notation $\mathbf{X}_{-i}$ indicates the set of all variables except X_i and we note that the conditional expectation given $X_i = x_i$ is recovered by averaging over the components $\mathbf{X}_{-i}$. Thus, having independent inputs is very convenient. Otherwise, it would be necessary to know the dependency structure in order to generate the observations $\left(\mathbf{X}_{-i}^{\diamond l}\right)_{l=1,\ldots,n}$ for each new point $X_i = x_i$, which would make this estimator very cumbersome.

In addition to being used to recover the conditional expectation given $X_i = x_i$, the sample $\left(\mathbf{X}^{\diamond j}\right)_{j=1,\ldots,n}$ is also used to estimate the outer expectation and we finally obtain the following estimator for the O term

$$\widehat{Q}_i^{3,b} = \frac{1}{n}\sum_{m=1}^{n}\min_{p=1,\ldots,n}\sum_{j=1}^{n} w_{n,j}^{b,i}\left(X_i^{\diamond m}\right)\psi_\alpha\left(Y^j,Y^p\right)\ .$$

By using the weights $w_{n,j}^{o}\left(\mathbf{x}\right)$ instead of $w_{n,j}^{b}\left(\mathbf{x}\right)$, we may define the estimator $\widehat{Q}_i^{3,o}$.

5.5 Overall estimation procedure

After defining the respective estimators for each term of the first-order QOSA index in Sections 5.2 and 5.4, the overall estimators are set in the following. In order to improve their accuracy, different strategies are also presented to tune hyperparameters of the random forest.

5.5.1 Issues with the leaf size

When using a random forest method for a regression task, a prediction is generally obtained by using the default values proposed in the packages for the $max\_features$ and $min\_samples\_leaf$ hyperparameters. There are some empirical studies on the

impact of these hyperparameters such as Díaz-Uriarte and De Andres [2006]; Scornet [2017]; Duroux and Scornet [2018] but no theoretical garantuee to support the default values.

Concerning the estimation methods of the O term of the QOSA index proposed in Section 5.4, except for $\widehat{Q}_i^{3,b}$ and $\widehat{Q}_i^{3,o}$, it turns out that the values of the hyperparameters must be chosen carefully.
First of all, as a forest explaining Y by X_i is built for each model's input, the $max\_features$ hyperparameter has no impact in our procedures because it equals 1. Regarding the $min\_samples\_leaf$ hyperparameter, its impact on the quality of the estimators is investigated through the following toy example

$$Y = X_1 - X_2 \ , \tag{5.12}$$

with $X_1, X_2 \sim \mathcal{E}(1)$. This standard example is commonly used in Sensitivity Analysis literature to assess the quality of QOSA index estimators such as in Fort et al. [2016]; Browne et al. [2017]; Maume-Deschamps and Niang [2018].

To illustrate the influence of this hyperparameter, we present in Figure 5.1 the boxplot of $\widehat{R}_1^{1,o}$ made with 100 values for different leaf sizes. For each value of $min\_samples\_leaf$, an estimation $\widehat{R}_1^{1,o}$ is computed using two samples of size $n = 10^4$ and a forest grown with $n_{trees} = 500$. Then, the boxplots are compared with the analytical value given below and represented with the dotted orange line on each graph in Figure 5.1:

$$\mathbb{E}\left[\psi_\alpha\left(Y, q^\alpha\left(Y|\, X_1\right)\right)\right] = e^{-q^{1-\alpha}(X_2)}\left(1 + q^{1-\alpha}\left(X_2\right)\right) - \alpha \ .$$

Based on the results obtained in Figure 5.1, we see that for each level α, the performance of $\widehat{R}_1^{1,o}$ depends highly on the choice of the $min\_samples\_leaf$ hyperparameter. Indeed, with the grid proposed for the values of $min\_samples\_leaf$, the optimum value seems to be 258 for $\alpha = 0.1$, 83 for $\alpha = 0.3$, 47 for $\alpha = 0.7$ and 27 for $\alpha = 0.9$.
This issue about the leaf size is only highlighted for $\widehat{R}_i^{1,o}$ but is also encountered for both methods, stated in Subsection 5.4.1, computing the conditional quantile with either the bootstrap samples or the original sample.

By using the same setting as in Figure 5.1, the distribution of $\widehat{Q}_1^{1,o}$ is presented in Figure 5.2 in order to assess the impact of the $min\_samples\_leaf$ hyperparameter for a method where the minimum is estimated instead of plugging the quantile. The quality of $\widehat{Q}_1^{1,o}$ also seems to depend on the leaf size and the optimum value, allowing to well estimate $\mathbb{E}\left[\min_{\theta\in\mathbb{R}} \mathbb{E}\left[\psi_\alpha\left(Y,\theta\right)|\, X_1\right]\right]$ for each level α, is the same as in Figure 5.1.
As before, this concern about the leaf size was only emphasized for $\widehat{Q}_i^{1,o}$ but is also encountered for both methods, detailed in Subsections 5.4.2.1 and 5.4.2.2, approximating the minimum with either the bootstrap samples or the original sample.

For the methods $\widehat{Q}_i^{3,b}$ and $\widehat{Q}_i^{3,o}$, based on complete trees, it seems that the tuning of the leaf size is less important as observed in Figure 5.3. Indeed, whatever the α level, the best results are observed for almost fully developed trees.

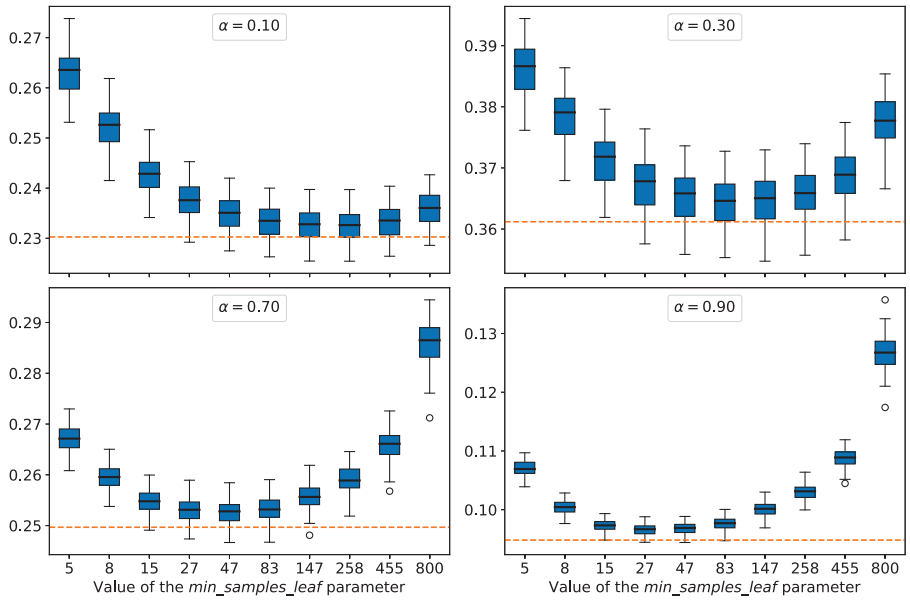


Figure 5.1 *For several levels α: distribution of $\widehat{R}_1^{1,o}$, the estimation of the O term associated to the variable X_1 for different leaf sizes. The dotted orange line represents the true value on each plot.*

Thus, for all other estimators of the O term proposed in Section 5.4, a method giving us the optimal value of the leaf size for each level α is required to properly estimate the first-order QOSA index.

5.5.2 Tuning the leaf size

In order to tune the leaf size of our estimators, two methods are presented in this part. They lead to significatively improve the efficiency of the estimation. The first one rests on a classical cross-validation procedure and the second one uses the Out-Of-Bag samples.

5.5.2.1 Cross-validation procedure

The estimators of the O term developed in Subsection 5.4.1 are part of the conditional quantile estimation problem. Indeed, in a regression scheme, the conditional mean minimizes the expected squared error loss, while the conditional quantile $q^\alpha(Y|X_i = x_i)$ minimizes the following expected loss

$$q^\alpha(Y|X_i) = \underset{h:\mathbb{R}\to\mathbb{R}}{\arg\min}\, \mathbb{E}\left[\psi_\alpha(Y, h(X_i))\right] \ .$$

Thus, estimators of $\mathbb{E}\left[\psi_\alpha(Y, q^\alpha(Y|X_i))\right]$ established in Subsection 5.4.1 allow to assess the quality of the approximation of the true conditional quantile function. The smaller

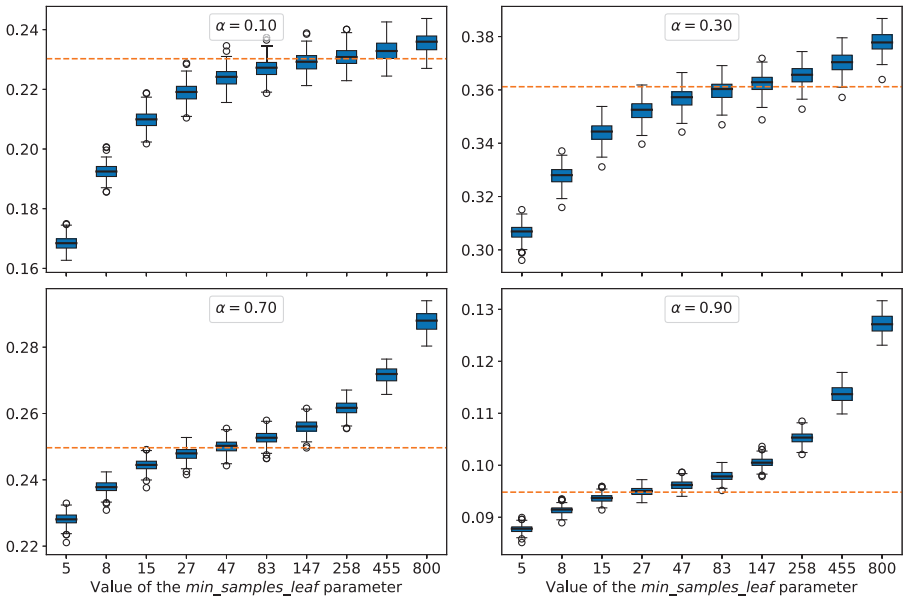


Figure 5.2 *For several levels α: distribution of $\widehat{Q}_1^{1,o}$, the estimation of the O term associated to the variable X_1 for different leaf sizes. The dotted orange line represents the true value on each plot.*

they are, the better the estimate of the conditional quantile function is. That is verified in Figure 5.1 and explains why we have this convex shape depending on the leaf size. As a matter of fact, when the value of the *min_samples_leaf* hyperparameter is incorrectly chosen, the approximation of the true conditional quantile function is wrong and so, this of $\mathbb{E}\left[\psi_\alpha\left(Y, q^\alpha\left(Y|\, X_i\right)\right)\right]$ too. At the opposite, if the hyperparameter is well chosen, the results are better.

Hence, in order to estimate well the conditional quantile function $q^\alpha\left(Y|\, X_i\right)$ and therefore, $\mathbb{E}\left[\psi_\alpha\left(Y, q^\alpha\left(Y|\, X_i\right)\right)\right]$ (which is our goal), the optimum value of the leaf size will be chosen within a predefined grid containing potential values as being the one minimizing the empirical generalization error computed with a K-fold cross-validation procedure. A detailed description of this process is given in Algorithm 5 with $\widehat{R}_i^{1,o}$ for instance. The principle is the same for all estimators defined in Subsection 5.4.1.

It has to be noted that the number of folds K should be chosen carefully. Indeed, a lower value of K results in a more biased estimation of the generalization error, and hence undesirable. In contrast, a larger value of K is less biased, but can suffer from large variability. The choice of K is usually 5 or 10, but there is no formal rule.

Regarding the minimum-based estimators, using a similar approach with a K-fold cross-validation procedure is unsuitable due to the behavior of these ones depending on the leaf size (cf. Figure 5.2). Consequently, we propose to get the optimal value with one of the estimators plugging the quantile, in conjunction with the cross-validation process detailed in Algorithm 5. Once done, the estimator based on the minimum is

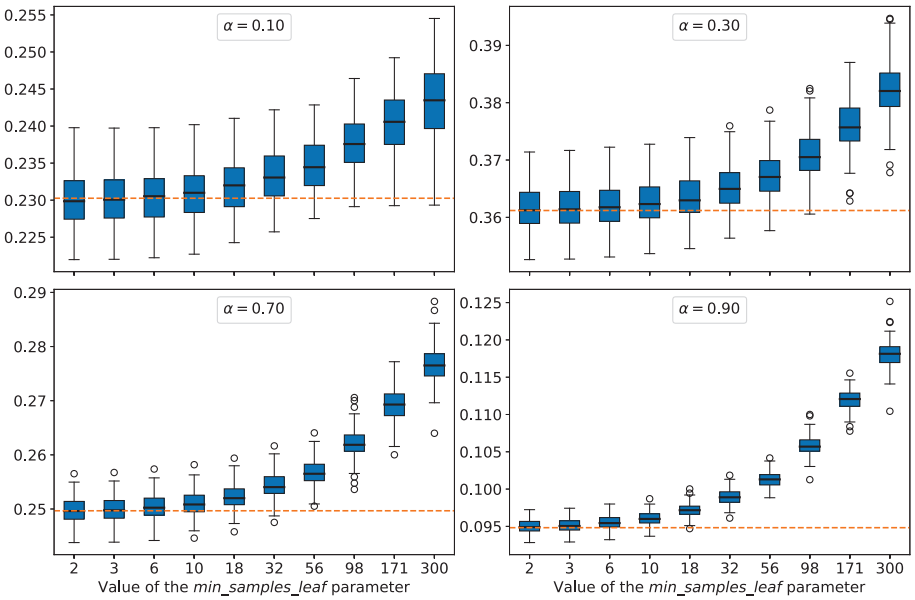


Figure 5.3 *For several levels α: distribution of $\widehat{Q}_1^{3,o}$, the estimation of the O term associated to the variable X_1 for different leaf sizes. The dotted orange line represents the true value on each plot.*

computed with the optimal value obtained.

5.5.2.2 Out-Of-Bag quantile error

The estimators $\widehat{R}_i^{2,b}$ and $\widehat{R}_i^{2,o}$ detailed in Subsection 5.4.1.2 deserve special attention. Indeed, another much less cumbersome approach than cross-validation can be used to tune the leaf size. It is based on an adaptation to our context of the widespread "Out-Of-Bag" (OOB) error [Breiman, 1996b] in regression and classification to estimate the generalization error.

First, define OOB quantile error for $\widehat{R}_i^{2,b}$.
Let us fix an observation (X_i^j, Y^j) from $\mathcal{D}_n^i$ and consider the set of trees built with the bootstrap samples not containing this observation, i.e. for which this one is "Out-Of-Bag". We then aggregate only the predictions of these trees to make our prediction $\widehat{q}_{oob}^{b,\alpha}\left(Y|\, X_i = X_i^j\right)$ of $q^\alpha\left(Y|\, X_i = X_i^j\right)$. After this operation carried out for all the data in $\mathcal{D}_n^i$, we calculate the error related to the approximation of the true conditional quantile function, i.e. the empirical generalization error

$$\widehat{OOB}_i^b = \frac{1}{n}\sum_{j=1}^{n} \psi_\alpha\left(Y^j, \widehat{q}_{oob}^{b,\alpha}\left(Y|\, X_i = X_i^j\right)\right) \ . \tag{5.13}$$

Now, turn to the estimator $\widehat{R}_i^{2,o}$.

Algorithme 5 : K-fold cross-validation procedure explained with $\widehat{R}_i^{1,o}$

Input :

- Datasets: $\mathcal{D}_n^{\diamond i} = \left(X_i^{\diamond j}, Y^{\diamond j}\right)_{j=1,\dots,n}$ from $\mathcal{D}_n^{\diamond}$ and $\mathcal{D}_n^i = \left(X_i^j, Y^j\right)_{j=1,\dots,n}$ from $\mathcal{D}_n$
- Number of trees: $k \in \mathbb{N}^{\star}$
- The order where estimating $\mathbb{E}\left[\psi_\alpha\left(Y, q^\alpha\left(Y|\, X_i\right)\right)\right] : \alpha \in]0,1[$
- Grid where looking for the best parameter: *grid_min_samples_leaf*
- Number of folds: $K \in \{2, \dots, n\}$

Output : Estimated value of $\mathbb{E}\left[\psi_\alpha\left(Y, q^\alpha\left(Y|\, X_i\right)\right)\right]$ at the α-level with $\widehat{R}_i^{1,o}$

begin Cross-validation procedure
  Randomly split the dataset $\mathcal{D}_n^i$ into K folds.
  foreach $\ell \in$ *grid_min_samples_leaf* **do**
    foreach *fold* **do**
      Take the current fold as a test set.
      Take the remaining groups as a training set.
      Fit a random forest model on the training set with the current ℓ as *min_samples_leaf* hyperparameter.
      Evaluate the conditional quantiles at the observations X_i in the test dataset and then compute $\widehat{R}_i^{1,o}$ on the test set.
      Retain the estimation obtained.
    end
    Summarize the quality related to the current ℓ by averaging the K estimated values and save the mean.
  end
end
Select as optimal value ℓ_{opt} for the *min_samples_leaf* hyperparameter, this one with the smallest mean.
Fit a random forest model on the complete dataset $\mathcal{D}_n^i$ by fixing the *min_samples_leaf* hyperparameter to ℓ_{opt}.
Compute $\widehat{R}_i^{1,o}$ with $\mathcal{D}_n^{\diamond i}$.

Again, let us fix an observation (X_i^j, Y^j) from $\mathcal{D}_n^i$ and consider $\mathcal{I}^j$ as the set of trees built with the bootstrap samples not containing this observation. For the ℓ-th tree in $\mathcal{I}^j$, the conditional quantile given $X_i = X_i^j$ is estimated with the set of the observations $\bar{\mathcal{L}}_\ell^o\left(X_i^j\right) = \mathcal{L}_\ell^o\left(X_i^j\right) \backslash (X_i^j, Y^j)$. The predictions from each tree in $\mathcal{I}^j$ are subsequently aggregated to make the prediction $\widehat{q}_{oob}^{o,\alpha}\left(Y|\, X_i = X_i^j\right)$ of $q^\alpha\left(Y|\, X_i = X_i^j\right)$. Once done for all the data in $\mathcal{D}_n^i$, the following empirical generalization error is computed

$$\widehat{OOB_i}^o = \frac{1}{n}\sum_{j=1}^{n} \psi_\alpha\left(Y^j, \widehat{q}_{oob}^{o,\alpha}\left(Y|\, X_i = X_i^j\right)\right) \ . \tag{5.14}$$

The advantage of this method, compared to cross-validation techniques, is that it

does not require cutting out the training sample $\mathcal{D}_n^i$ and takes place during the forest construction process.

Thus, given the dataset $\mathcal{D}_n^i$ and a grid containing potential values of the *min_samples_leaf* hyperparameter, a random forest is built for each one and the OOB quantile error associated is computed. Then, the optimal hyperparameter is chosen as this one with the smallest error.

5.5.3 Full estimation procedure

Now, we have all the components in order to set the estimators of the first-order QOSA index S_i^α. These are separated in two classes according to the estimation method adopted for the O term. First of all, with the methods plugging the quantile, we define

$$\widehat{S}_i^\alpha = 1 - \frac{\widehat{R}_i}{\widehat{P}_1} \text{ with } \widehat{R}_i \in \left\{\widehat{R}_i^{1,b}, \widehat{R}_i^{1,o}, \widehat{R}_i^{2,b}, \widehat{R}_i^{2,o}\right\} . \tag{5.15}$$

The whole procedure integrating the cross-validation process for these methods is detailed in Algorithm 6.

On the other hand, regarding the methods based on the minimum to compute the O term, we set

$$\widehat{S}_i^\alpha = 1 - \frac{\widehat{Q}_i}{\widehat{P}_1} \text{ with } \widehat{Q}_i \in \left\{\widehat{Q}_i^{1,b}, \widehat{Q}_i^{1,o}, \widehat{Q}_i^{2,b}, \widehat{Q}_i^{2,o}, \widehat{Q}_i^{3,b}, \widehat{Q}_i^{3,o}\right\} . \tag{5.16}$$

The estimation process based on the minimum is formalized in Algorithms 7, 8 and 9. For the sake of clarity, they are all gathered in Appendix 5.8.1 . Algorithm 7 (resp. 9) estimating the QOSA index with $\widehat{Q}_i^{1,b}$ or $\widehat{Q}_i^{1,o}$ (resp. $\widehat{Q}_i^{3,b}$ or $\widehat{Q}_i^{3,o}$), needs a full training sample $\mathcal{D}_n$ as well as a partial one $(\mathbf{X}^{\diamond j})_{j=1,\dots,n}$. While estimating the QOSA index with $\widehat{Q}_i^{2,b}$ or $\widehat{Q}_i^{2,o}$ only requires one training sample $\mathcal{D}_n$. This is a major advantage over methods plugging the quantile that need two full training samples.

So far, no consistency result has been proved for $\widehat{S}_i^\alpha$. These various estimators are reviewed in the next section in order to establish their efficiency in practice. Moreover, all these algorithms are implemented within a `python` package named `qosa-indices` available at Elie-Dit-Cosaque [2020], it can be also freely downloaded on the PyPI website.

5.6 Numerical illustrations

Let us now carry out some simulations in order to investigate the influence of the number of trees on our estimators and compare the decrease of the estimation error of each one in function of the train sample-size. From these results, the performance of

the two best estimators as well as those based on kernel methods defined in Browne et al. [2017]; Maume-Deschamps and Niang [2018] is assessed. Then, their scalability is tested on a toy example.

5.6.1 Convergence with the number of trees and the train sample-size

We start by studying the impact of the number of trees on the performance of our estimators except for those using $\widehat{Q}_i^{3,b}$ and $\widehat{Q}_i^{3,o}$ because of the computational cost. This survey is carried out with the model introduced in Equation (5.12) and the following setting.

The estimators of the QOSA index are computed with samples of size $n = 10^4$. The leaf size is tuned over a grid with 20 numbers evenly spaced ranging from 5 to 300 by using a 3-fold cross-validation procedure for $\widehat{R}_i^{1,b}$ and $\widehat{R}_i^{1,o}$ while the strategy based on the OOB samples, developed in Subsection 5.5.2.2, is used for $\widehat{R}_i^{2,b}$ and $\widehat{R}_i^{2,o}$. Regarding the minimum based estimators, the optimal leaf size is obtained via $\widehat{R}_i^{1,o}$ during the 3-fold cross-validation process. Then, to assess the efficiency of our estimators, we repeat the experiment $s = 200$ times and compute the following metrics for each one

$$\begin{aligned} RMSE_i^\alpha &= \sqrt{\frac{1}{s}\sum_{j=1}^{s}\left(\widehat{S}_i^{\alpha,j} - S_i^\alpha\right)^2}\,, \\ Bias_i^\alpha &= \left|\frac{1}{s}\sum_{j=1}^{s}\widehat{S}_i^{\alpha,j} - S_i^\alpha\right|\,, \\ Variance_i^\alpha &= \frac{1}{s}\sum_{j=1}^{s}\left(\widehat{S}_i^{\alpha,j} - \frac{1}{s}\sum_{j=1}^{s}\widehat{S}_i^{\alpha,j}\right)^2\,, \end{aligned} \tag{5.17}$$

with S_i^α, the analytical values that were provided in Fort et al. [2016].

In Figure 5.4, for three levels α, we present the evolution of the different metrics related to the variable X_1 of our toy example in function of the number of trees ranging from 1 to 200 (in log scale). More precisely, sub-figures at the top of Figure 5.4 show the Root Mean Square Error (RMSE), in the middle, the bias and the variance at the bottom.

We observe that regardless of the level α, RMSE of our estimators is small. The number of trees seems to have no impact for those using $\widehat{Q}_i^{1,o}$ and $\widehat{Q}_i^{2,o}$ as the RMSE value is almost always the same. RMSE of the others decreases in function of the number of trees until it reaches a threshold starting at about 50 trees. Indeed, it is well known that from a certain number, increasing the number of trees becomes useless but results in higher calculation costs. However, we did not expect to have a stable estimation error with so few trees.
Besides, still from the RMSE curves, it first appears that the estimators using the original sample (plain lines) have a lower error compared to those using the bootstrap

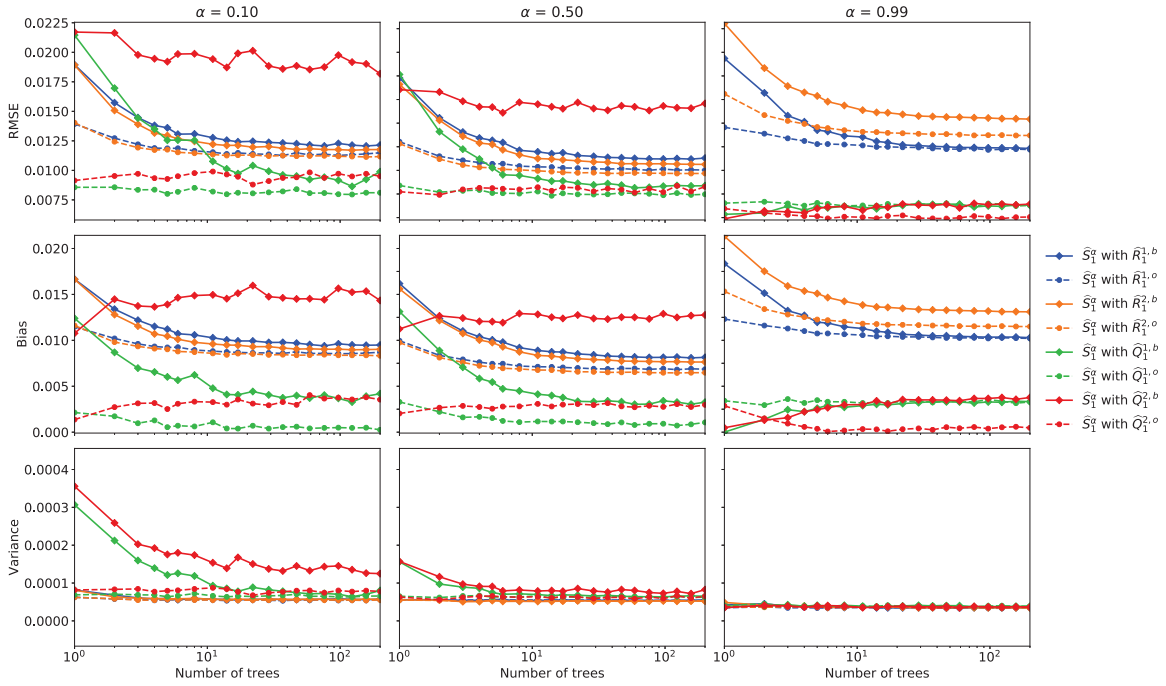


Figure 5.4 *Evolution of RMSE, bias and variance of the estimators associated with X_1 in function of the number of trees for three levels α.*

samples (dotted lines). On the other hand, the performance of the minimum based estimators (green and red lines) seems better than those based on the quantile (blue and orange lines). That might be explained by the additional error due to the estimation of the conditional quantile.
Variance of all estimators is close to 0 and the bias curves have the same behavior as RMSE curves. This means that bias is the main/only source of error in the RMSE. This bias could be reduced by taking a larger grid where looking for the optimal leaf size during the cross-validation or using another method to find the optimum.

Let us now compare the decrease of the estimation error in function of the train sample-size. As observed in Figure 5.4, take a very large number of trees is not required in order to have a stable estimation error. Thus, we take $n_{trees} = 100$ and the same setting as before for other parameters in the next study and observe the evolution of the metrics introduced in Equation (5.17) in function of the sample size.

Figure 5.5 presents RMSE, bias and variance of our estimators for different sample sizes. We observe that all the metrics associated with the various estimators converge to 0 at different rates. Indeed, the convergence rates of the metrics of the quantile-based estimators are slower than those based on the minimum.

Hence, from our experiments, it turns out that the minimum-based estimators give the best results. This is an interesting feature because they need less data than those

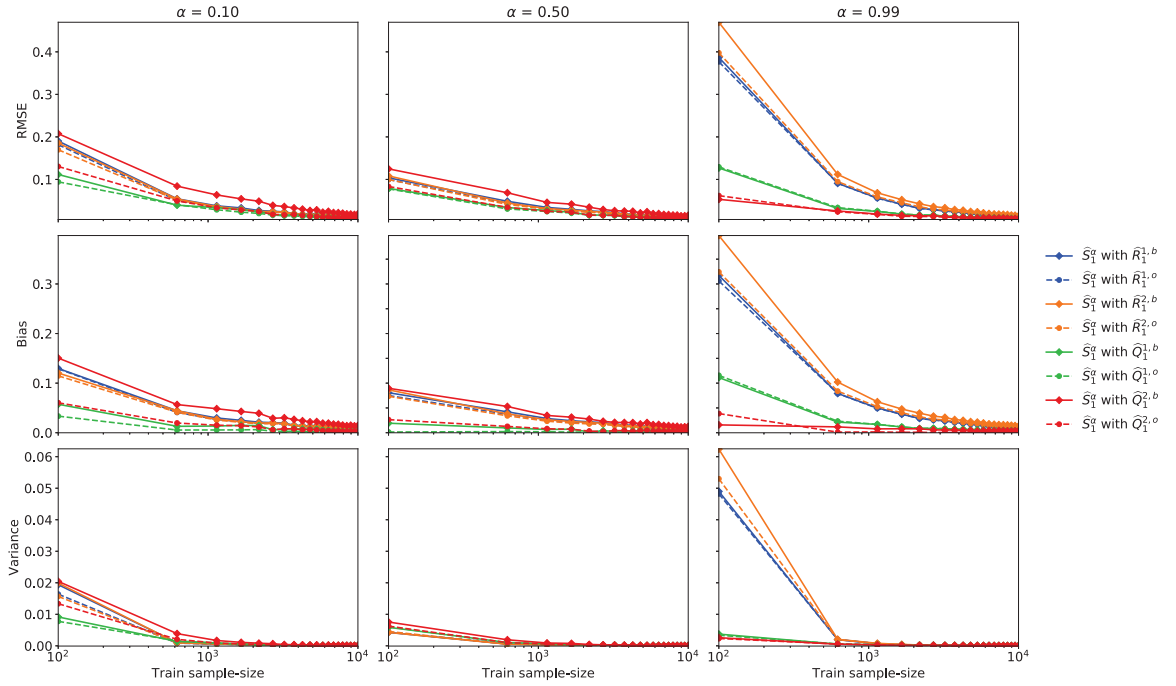


Figure 5.5 *Evolution of RMSE, bias and variance of the estimators associated with X_1 in function of the train sample-size for three levels α.*

plugging the quantile. Furthermore, few trees are necessary in order to reduce the estimation error. It therefore allows to get a good estimation of the indices with a reasonable computational cost.

5.6.2 Comparison with kernel methods

In this subsection, we compare on the toy example introduced in Equation (5.12):

- the kernel-based estimators proposed in Browne et al. [2017]; Maume-Deschamps and Niang [2018] denoted by $\check{S}_i^\alpha$ and $\widetilde{S}_i^\alpha$,
- the minimum-based QOSA index estimators building one forest for each input and using the original sample,
- and the minimum-based QOSA index estimators using a forest grown with trees fully developed.

The estimators of the QOSA indices are computed with samples of size $n = 10^4$. Forest methods are grown with $n_{trees} = 100$. The optimal leaf size for the minimum-based estimators building one forest for each input is obtained with $\widehat{R}_i^{1,o}$ during the 3-fold cross-validation process over a grid containing 20 numbers evenly spaced ranging from 5 to 300. Regarding the minimum-based estimators using a forest grown with

trees fully developed, the *min_samples_leaf* hyperparemeter equals 2.
In order to have comparable methods, a cross-validation procedure is also implemented for the kernel-based estimators to choose the optimal bandwidth parameter. It is selected within over a grid containing 20 potential values ranging from 0.001 to 1. Then, we assess the performance of the different estimators by computing their empirical root mean squared error with 100 experiments.

	$\widehat{S}_i^\alpha$ with $\widehat{Q}_i^{1,o}$		$\widehat{S}_i^\alpha$ with $\widehat{Q}_i^{2,o}$		$\widehat{S}_i^\alpha$ with $\widehat{Q}_i^{3,b}$		$\widehat{S}_i^\alpha$ with $\widehat{Q}_i^{3,o}$		$\check{S}_i^\alpha$		$\widetilde{S}_i^\alpha$	
	X_1	X_2	X_1	X_2	X_1	X_2	X_1	X_2	X_1	X_2	X_1	X_2
$\alpha = 0.1$	0.007	0.006	0.009	0.006	0.017	0.006	0.017	0.006	0.020	0.044	0.061	0.006
$\alpha = 0.25$	0.008	0.006	0.009	0.006	0.013	0.007	0.013	0.007	0.013	0.036	0.042	0.012
$\alpha = 0.5$	0.008	0.006	0.008	0.007	0.010	0.009	0.010	0.009	0.019	0.021	0.027	0.025
$\alpha = 0.75$	0.008	0.007	0.008	0.008	0.008	0.014	0.008	0.014	0.035	0.012	0.014	0.042
$\alpha = 0.99$	0.006	0.016	0.006	0.018	0.006	0.032	0.006	0.032	0.084	0.071	0.013	0.11
run time	1 hr		18 min 24 sec		10 hr 41 min		8 hr 18 min		1 hr 55 min		1 min 51 sec	

Table 5.1 *RMSE and run time for the toy example of the random forest based estimators:* $\widehat{S}_i^\alpha$ *computed with* $\widehat{Q}_i^{1,o}$, $\widehat{Q}_i^{2,o}$, $\widehat{Q}_i^{3,b}$ *and* $\widehat{Q}_i^{3,o}$ *as well as those based on kernel:* $\widetilde{S}_i^\alpha$ *and* $\check{S}_i^\alpha$.

Table 5.1 contains the empirical root mean square error of the different estimators associated to eah input as well as the overall run time requested to obtain them. About their performance, it seems that the random forest-based estimators are better than the kernel methods. Nevertheless, as regards the methods using $\widehat{Q}_i^{3,b}$ and $\widehat{Q}_i^{3,o}$, while they have a low error and do not need to tune the leaf size, their run time with the current implementation is too long to be used in practice. Accordingly, we recommend to compute the indices with $\widehat{Q}_i^{1,o}$ and $\widehat{Q}_i^{2,o}$ in order to get good estimations of the first-order QOSA indices in a reasonable time.

5.6.3 Scalability of the methods

The influence of the model's dimension d over the performance of the estimators using $\widehat{Q}_i^{1,o}$ and $\widehat{Q}_i^{2,o}$ is investigated in this subsection with the following additive Exponential framework

$$Y = \sum_{i=1}^{d} X_i \ . \tag{5.18}$$

Independent inputs $X_i, i = 1, \ldots, d$, follow an Exponential distribution $\mathcal{E}(\lambda_i)$ and the resulting output Y is a generalized Erlang distribution also called Hypoexponential distribution. By taking advantage of the other expression of the first-order QOSA index given in Maume-Deschamps and Niang [2018], we obtain the following semi closed-form analytical formula

$$S_i^\alpha = 1 - \frac{\alpha\mathbb{E}\left[Xs_{(-i)}\right] - \mathbb{E}\left[Xs_{(-i)}\mathbb{1}_{\left\{Xs_{(-i)} \leqslant q^\alpha\left(Xs_{(-i)}\right)\right\}}\right]}{\alpha\mathbb{E}\left[Y\right] - \mathbb{E}\left[Y\mathbb{1}_{\{Y \leqslant q^\alpha(Y)\}}\right]} \ , \tag{5.19}$$

with $Xs_{(-i)} = \sum_{j \neq i} X_j$ that also follows a Hypoexponential distribution. Knowing the cumulative distribution function of the Hypoexponential distribution, quantiles $q^{\alpha}(Y)$ and $q^{\alpha}\left(Xs_{(-i)}\right)$ are computed by numeric inversion and the analytical expression of the truncated expectations is derived from Marceau [2013].

For a specific dimension d, d values evenly spaced are selected from the interval $[0.3, 1.25]$ and then each one represents the λ_i parameter of an input X_i, $i = 1, \ldots, d$. QOSA index estimations are then computed with samples of size $n = 10^4$, a forest grown with $n_{trees} = 100$ and the setting defined hereafter. The leaf size is tuned with $\widehat{R}_i^{1,o}$ over a grid with 20 numbers evenly spaced ranging from 5 to 300 by using a 3-fold cross-validation. Each experiment is done 100 times in order to compute the RMSE defined in Equation (5.17) for each input, and then we take the weighted mean by the analytical values of the QOSA indices over all dimensions in order to get a global measure.

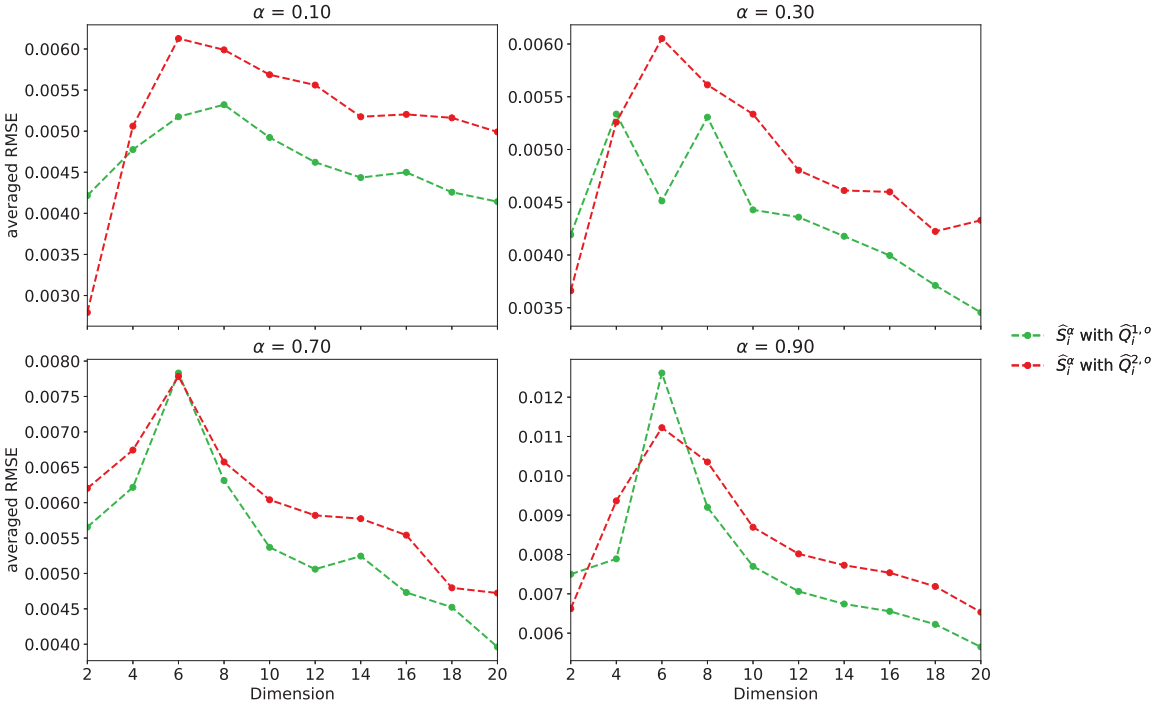


Figure 5.6 *Evolution of the averaged RMSE over all dimensions of the estimators calculated with* $\widehat{Q}_i^{1,o}$ *and* $\widehat{Q}_i^{2,o}$ *in function of the model dimension for four levels* α.

Figure 5.6 presents the weighted RMSE as a function of the increasing dimension of our model for several levels α. For each one, we observe that the error increases slowly at the beginning until the dimension 6 for both methods then decreases. This phenomenon is due to the chosen parametrization. Indeed, when increasing the dimension of the model, the respective impact of each input is reduced. Thus, from a certain dimension, all the analytical values of the first-order QOSA indices become

small and even close to 0 for some inputs. Our etimators properly capture this trend as they decrease by increasing the dimension. However, the estimator using $\widehat{Q}_i^{1,o}$ seems better than this one employing $\widehat{Q}_2^{1,o}$ as its error is lower.

5.7 Conclusion

In this paper, we introduced several estimators for the first-order QOSA index by using the random forest method. Some of them use the original sample while the others use the bootstrap samples generated during the forest construction. Both classes of estimator seem to be efficient even if we observe in our experiments that the methods using the original sample have a lower estimation error than those based the bootstrap ones. Thus, supplementary studies should be conducted to inquire into this difference. Furthermore, the performance of these methods is highly dependent on the leaf size. This parameter could be compared to the bandwidth parameter of kernel estimators as it controls the bias of the method. But, it turns out to be easier to calibrate and we propose two methods to do this.

It is also well known for the random forest methods that the number of trees k should be chosen large enough to reach the desired statistical precision and small enough to make the calculations feasible as the computational cost increases linearly with k as mentioned in Scornet [2017]. But, we have seen on our "toy example" that estimators proposed herein require few trees in order to have a low estimation error. This makes possible to estimate the indices correctly while maintaining a reasonable computation time.

Besides, we obtain in our application better results for our estimators when comparing with the kernel methods. A major advantage is that we have developed an estimator that requires only one training sample, whereas kernel methods require two training samples or a full one plus a partial. This feature is interesting when dealing with costly models. Another significant asset of our estimators is that their efficiency seems maintained when increasing the model dimension.

Despite these benefits, the proof for the estimators' consistency as well as the asymptotic analysis to establish the convergence rates and confidence intervals remains a major wish for the future. It is also important to remember that these indices do not have an analogue to the variance decomposition offered by Sobol indices through the theorem of Hoeffding [1948]. Thus, using the values of Shapley [1953] could be interesting to get condensed and easy-to-interpret indices with a good apportionment of the interaction and dependences contributions between the inputs involved.

5.8 Appendix

5.8.1 Algorithms for estimating the first-order QOSA index

Algorithme 6 : QOSA index estimators plugging the quantile

Input :

- Datasets: $\mathcal{D}_n^{\diamond} = \left(\mathbf{X}^{\diamond j}, Y^{\diamond j}\right)_{j=1,\dots,n}$ and $\mathcal{D}_n = \left(\mathbf{X}^{j}, Y^{j}\right)_{j=1,\dots,n}$
- Number of trees: $k \in \mathbb{N}^{\star}$
- Order where estimating the QOSA index : $\alpha \in]0,1[$
- Grid where looking for the best parameter: $grid\_min\_samples\_leaf$
- Number of folds: $K \in \{2,\dots,n\}$

Output : Estimated value of the QOSA index at the α-order $\widehat{S}_i^{\alpha}$ for all inputs.

Compute $\widehat{P}$ thanks to Equation (5.3).
foreach $i = 1,\dots,d$ **do**
  $\mathcal{D}_n^{\diamond i} = \left(X_i^{\diamond j}, Y^{\diamond j}\right)_{j=1,\dots,n}$ from $\mathcal{D}_n^{\diamond}$ and $\mathcal{D}_n^{i} = \left(X_i^{j}, Y^{j}\right)_{j=1,\dots,n}$ from $\mathcal{D}_n$
  Cross-validation as in Algorithm 5 with $\mathcal{D}_n^i$ to get the optimal leaf size ℓ_{opt}.
  Fit a random forest model with $\mathcal{D}_n^i$ by fixing the $min\_samples\_leaf$ hyperparameter to ℓ_{opt}.
  Compute the estimator $\widehat{R}_i$ with $\mathcal{D}_n^{\diamond i}$.
  Compute $\widehat{S}_i^{\alpha} = 1 - \widehat{R}_i/\widehat{P}$.
end

Algorithme 7 : QOSA index estimators with the weighted minimum approach

Input :

- Datasets: $\mathcal{D}_n = \left(\mathbf{X}^j, Y^j\right)_{j=1,\dots,n}$ and $\left(\mathbf{X}^{\diamond j}\right)_{j=1,\dots,n}$
- Number of trees: $k \in \mathbb{N}^\star$
- Order where estimating the QOSA index : $\alpha \in]0,1[$
- Grid where looking for the best parameter: *grid_min_samples_leaf*
- Number of folds: $K \in \{2,\dots,n\}$

Output : Estimated value of the QOSA index at the α-order $\widehat{S}_i^\alpha$ for all inputs.

Compute $\widehat{P}$ thanks to Equation (5.3).
foreach $i = 1,\dots,d$ **do**
  $\mathcal{D}_n^i = \left(X_i^j, Y^j\right)_{j=1,\dots,n}$ from $\mathcal{D}_n$ and $\left(X_i^{\diamond j}\right)_{j=1,\dots,n}$
  Cross-validation as in Algorithm 5 with $\mathcal{D}_n^i$ to get the optimal leaf size ℓ_{opt}.
  Fit a random forest model with $\mathcal{D}_n^i$ by fixing the *min_samples_leaf* hyperparameter to ℓ_{opt}.
  Compute the estimator $\widehat{Q}_i \in \left\{\widehat{Q}_i^{1,b}, \widehat{Q}_i^{1,o}\right\}$ with $\mathcal{D}_n^{\diamond i}$ and $\left(X_i^{\diamond j}\right)_{j=1,\dots,n}$.
  Compute $\widehat{S}_i^\alpha = 1 - \widehat{Q}_i/\widehat{P}$
end

Algorithme 8 : QOSA index estimators computing the minimum in leaves

Input :

- Datasets: $\mathcal{D}_n = \left(\mathbf{X}^j, Y^j\right)_{j=1,\dots,n}$
- Number of trees: $k \in \mathbb{N}^\star$
- Order where estimating the QOSA index : $\alpha \in]0,1[$
- Grid where looking for the best parameter: *grid_min_samples_leaf*
- Number of folds: $K \in \{2,\dots,n\}$

Output : Estimated value of the QOSA index at the α-order $\widehat{S}_i^\alpha$ for all inputs.

Compute $\widehat{P}$ thanks to Equation (5.3).
foreach $i = 1,\dots,d$ **do**
  $\mathcal{D}_n^i = \left(X_i^j, Y^j\right)_{j=1,\dots,n}$ from $\mathcal{D}_n$
  Cross-validation as in Algorithm 5 with $\mathcal{D}_n^i$ to get the optimal leaf size ℓ_{opt}.
  Fit a random forest model with $\mathcal{D}_n^i$ by fixing the *min_samples_leaf* hyperparameter to ℓ_{opt}.
  Compute the estimator $\widehat{Q}_i \in \left\{\widehat{Q}_i^{2,b}, \widehat{Q}_i^{2,o}\right\}$.
  Compute $\widehat{S}_i^\alpha = 1 - \widehat{Q}_i/\widehat{P}$
end

Algorithme 9 : QOSA index estimators with the weighted minimum and fully grown trees

Input :

- Datasets: $\mathcal{D}_n = \left(\mathbf{X}^j, Y^j\right)_{j=1,\dots,n}$ and $\left(\mathbf{X}^{\diamond j}\right)_{j=1,\dots,n}$
- Number of trees: $k \in \mathbb{N}^\star$
- Order where estimating the QOSA index : $\alpha \in]0,1[$
- Minimum number of samples required in a leaf node: $min\_samples\_leaf \in \{1, ..., n\}$

Output : Estimated value of the QOSA index at the α-order $\widehat{S}_i^\alpha$ for all inputs.

Compute $\widehat{P}$ thanks to Equation (5.3).
Fit a random forest model with $\mathcal{D}_n$ and the $min\_samples\_leaf$ hyperparameter.
foreach $i = 1, \dots, d$ **do**
    Compute the estimator $\widehat{Q}_i \in \left\{\widehat{Q}_i^{3,b}, \widehat{Q}_i^{3,o}\right\}$ with $\left(\mathbf{X}^{\diamond j}\right)_{j=1,\dots,n}$.
    Compute $\widehat{S}_i^\alpha = 1 - \widehat{Q}_i / \widehat{P}$
end

Chapitre 6

Conclusions et perspectives

Ce travail de thèse a été l'occasion d'aborder de nouvelles méthodes permettant de réaliser une quantification efficiente du risque de modèle structurel, et plus particulièrement le risque lié aux paramètres du modèle. Pour ce faire, nous avons premièrement porté une attention particulière aux outils d'analyse de sensibilité permettant d'évaluer précisément ce risque à la fois dans le cas de paramètres dépendants et indépendants. Ensuite, différents moyens de les calculer ont été explorés, tels que l'usage de méta-modèle ou d'algorithme de machine learning, e.g. les forêts aléatoires.

6.1 Conclusions

Un état de l'art des méthodes d'Analyse de Sensibilité Globale est réalisé dans le Chapitre 2. Tout d'abord, nous avons présenté les indices de sensibilité basés sur la variance dans le cas d'entrées indépendantes et dépendantes. Ensuite, des indices nommés GOSA (Goal Oriented Sensitivity Anlaysis) permettant de quantifier les effets principaux d'entrées indépendantes sur une caractéristique de la sortie autre que la moyenne, à l'aide d'une fonction de contraste appropriée, sont exposés. Nous nous sommes concentrés sur les indices QOSA (Quantile Oriented Sensitivity Analysis) pour lesquels la quantité d'intérêt considérée est un quantile d'ordre alpha de la sortie. Plusieurs propriétés ont été établies pour ces derniers, notamment que l'indice QOSA d'ordre un peut être supérieur à l'indice QOSA total pour un modèle non additif avec des entrées indépendantes. De plus, un travail préliminaire a mis en exergue que le même phénomène était observé pour tout type de modèle (i.e. additif et non additif) en présence de dépendance stochastique entre les entrées. Afin de surmonter ces limitations et d'avoir des indices donnant une interprétation claire de l'impact de chaque entrée, à la fois dans le cas indépendant et dépendant, sur une caractéristique spécifique de la sortie, nous avons défini des *indices de Shapley subordonnés à une fonction de contraste*. En utilisant en particulier la fonction de contraste liée au quantile, on a introduit des *indices de Shapley orientés quantile*. Ceux-ci permettent de

quantifier correctement l'influence des variables d'entrée en allouant équitablement les effets d'interactions et de dépendance comme soulignés sur les exemples analytiques considérés.

Dans le chapitre 3, nous avons étudié puis comparé les indices de Shapley à la stratégie développée par Mara et al. [2015] basée sur l'estimation de quatre indices de Sobol, *full* et *indépendants*, dans le cas de variables d'entrée dépendantes. Les indices de Shapley se sont révélés être une mesure d'importance intéressante. En effet, au lieu d'effectuer une interprétation combinée de quatre indices qui peut s'avérer complexe, ils donnent une appréciation condensée et précise de l'impact de chaque entrée. En plus de cette étude comparative, nous avons également mis en œuvre un échantillonnage bootstrap dans l'algorithme d'estimation des indices de Shapley établi par Song et al. [2016] afin d'obtenir des intervalles de confiance représentatifs de l'erreur Monte-Carlo. D'autre part, afin d'obtenir une faible erreur d'estimation, l'algorithme requiert un grand nombre d'évaluations du modèle, ce qui peut être problématique si ce dernier est coûteux en temps de calcul. Pour pallier cette difficulté, nous substituons le vrai modèle par un métamodèle, i.e. un modèle de krigeage et proposons un algorithme pour calculer l'erreur d'estimation globale issue de l'erreur du modèle de krigeage et de l'erreur d'échantillonnage Monte-Carlo.

Nous proposons, dans le Chapitre 4, deux méthodes d'estimation des fonctions de répartition conditionnelles et des quantiles conditionnels basées sur les forêts aléatoires. La première est une généralisation naturelle de l'estimateur forêt aléatoire (exploitant les échantillons bootstrap de chaque arbre) développée pour estimer la fonction de régression tandis que la seconde est basée sur une variante introduite par Meinshausen [2006] utilisant uniquement le jeu de données original une fois la forêt construite. La consistance des estimateurs proposés a été démontrée sous certaines conditions. Concernant la première méthode, il s'agit à notre connaissance du premier résultat traitant de la composante bootstrap pour une méthode de forêt aléatoire. Le deuxième résultat est, pour sa part, une extension de celui de Meinshausen [2006] qui a montré la consistance pour un modèle simplifié de forêt aléatoire. En effet, il considère dans sa preuve que les poids sont non aléatoires. Nous allons plus loin en montrant la consistance du vrai estimateur, i.e. avec des poids aléatoires.

Au sein du Chapitre 5, nous avons exploré l'utilisation d'algorithme de machine learning, en particulier la méthode des forêts aléatoires, afin de construire des estimateurs efficaces de l'indice QOSA d'ordre un. Plusieurs méthodes d'estimation basées sur les estimateurs des quantiles conditionnels développés dans le Chapitre 4 ainsi que de nouvelles élaborées au sein du chapitre sont proposées. Les estimateurs expérimentés présentent globalement de bonnes performances et s'avèrent exiger peu d'arbres pour obtenir une faible erreur d'estimation. Cela constitue un atout indéniable car il sera dès lors possible d'obtenir en pratique de bonnes estimations en un temps de calcul raisonnable. De plus, nous avons sélectionné les meilleurs estimateurs afin de les comparer à ceux établis par Browne et al. [2017]; Maume-Deschamps and Niang [2018]. Il en ressort de notre analyse comparative basée sur un exemple jouet que nos estimateurs sont plus performants tout en requérant moins de données.

6.2 Perspectives

Plusieurs perspectives à la fois théoriques et pratiques peuvent découler de ces travaux. Nous en proposons quelques pistes.

Dans le Chapitre 3, l'algorithme d'estimation des indices de Shapley établi par Song et al. [2016] est efficace mais extrêmement coûteux en raison notamment de l'estimation des variances conditionnelles. Une amélioration significative de ce dernier serait l'utilisation d'une méthode à noyau ou procédure similaire afin de réduire considérablement le nombre d'évaluations. Par ailleurs, le développement en polynômes de chaos de la réponse du modèle permet de calculer analytiquement les indices de Sobol à partir des coefficients polynomiaux [Crestaux et al., 2009]. Il serait intéressant d'étudier si une telle décomposition est possible pour les indices de Shapley.

Concernant le Chapitre 4, une hypothèse clé permettant de montrer la consistance des estimateurs est l'Hypothèse 4.4.1 rappelée ci-dessous.

Hypothèse 4.4.1.
Pour tout $\ell \in [\![1, k]\!]$, nous supposons que la variation de la fonction de répartition conditionnelle tend vers 0 dans chaque cellule :

$$\forall \mathbf{x} \in \mathcal{X}, \forall y \in \mathbb{R}, \quad \sup_{\mathbf{z} \in A_n(\mathbf{x}; \Theta_\ell, \mathcal{D}_n)} |F(y|\,\mathbf{z}) - F(y|\,\mathbf{x})| \xrightarrow[n\to\infty]{a.s.} 0 \ .$$

Bien que ce soit une hypothèse de "bon sens", une amélioration substantielle serait de pouvoir la supprimer. Cependant, cela nécessiterait de surmonter des difficultés techniques liées à la structure de partitionnement de l'algorithme CART. Un autre axe de recherche pourrait être l'étude de la vitesse de convergence ainsi que de la distribution asymptotique des différents estimateurs proposés.

Dans le Chapitre 5, malgré la mise en exergue des bonnes performances des différents estimateurs proposés, il serait intéressant, d'un point de vue théorique, de montrer la consistance de ceux-ci mais aussi d'effectuer une analyse asymptotique permettant d'obtenir les vitesses de convergence ainsi que les intervalles de confiance. D'autre part, d'un point de vue pratique, réutiliser certains outils développés dans ce chapitre afin d'établir un algorithme d'estimation efficace des nouveaux *indices de Shapley orientés quantile* serait une contribution significative mais compliquée à réaliser. En effet, l'implication de tous les sous-ensembles des entrées dans l'indice rendrait potentiellement l'algorithme très coûteux en grande dimension. Une première piste pourrait être d'adapter l'algorithme développé par Plischke et al. [2020], dans le cadre des indices de Shapley basés sur la variance, qui apporte des améliorations majeures aux implémentations actuelles de Song et al. [2016].

Enfin, comme rappelé dans l'introduction, les compagnies d'assurance ou de réassurance doivent disposer d'un capital nommé SCR afin de faire face aux risques qu'elles encourent dans leur activité. Celui-ci est déterminé en calculant le quantile à 99,5% de la variable aléatoire $\mathcal{S}$ représentative de la distribution de la valeur des fonds propres à

horizon un an. Les indices QOSA ou *indices de Shapley orientés quantile* peuvent être utilisés pour quantifier la contribution des différents paramètres au quantile $q^{99.5}(\mathcal{S})$. Toutefois, on pourrait aussi utiliser les *indices de Shapley subordonnés à une fonction de contraste*, avec en particulier la fonction de contraste associée à une probabilité de dépassement de seuil. Sous réserve que le quantile estimé $\widehat{q}^{99.5}(\mathcal{S})$ soit proche de la vraie valeur, cela permettrait d'évaluer directement la contribution des paramètres à la quantité $\mathbb{P}(\mathcal{S} \geqslant \widehat{q}^{99.5}(\mathcal{S}))$ et donc de déterminer les paramètres ayant le plus d'impact sur la probabilité qu'une compagnie se retrouve en situation de ruine économique.

Bibliographie

Amit, Y. and Geman, D. (1997). Shape quantization and recognition with randomized trees. *Neural computation*, 9(7) :1545–1588.

Arenal-Gutiérrez, E., Matrán, C., and Cuesta-Albertos, J. A. (1996). Unconditional glivenko-cantelli-type theorems and weak laws of large numbers for bootstrap. *Statistics & probability letters*, 26(4) :365–375.

Athey, S., Tibshirani, J., Wager, S., et al. (2019). Generalized random forests. *The Annals of Statistics*, 47(2) :1148–1178.

Barrieu, P. and Scandolo, G. (2015). Assessing financial model risk. *European Journal of Operational Research*, 242(2) :546–556.

Benoumechiara, N. and Elie-Dit-Cosaque, K. (2018). shapley-effects, a python package available at : `https://gitlab.com/CEMRACS17/shapley-effects/-/tree/dev`.

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.

Bezanson, J., Edelman, A., Karpinski, S., and Shah, V. B. (2017). Julia : A fresh approach to numerical computing. *SIAM review*, 59(1) :65–98.

Biau, G. (2012). Analysis of a random forests model. *Journal of Machine Learning Research*, 13(Apr) :1063–1095.

Biau, G. and Devroye, L. (2010). On the layered nearest neighbour estimate, the bagged nearest neighbour estimate and the random forest method in regression and classification. *Journal of Multivariate Analysis*, 101(10) :2499–2518.

Biau, G. and Scornet, E. (2016). A random forest guided tour. *Test*, 25(2) :197–227.

Borgonovo, E. (2006). Measuring uncertainty importance : investigation and comparison of alternative approaches. *Risk analysis*, 26(5) :1349–1361.

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 : An International Journal*, 31(3) :404–428.

Borgonovo, E. et al. (2017). Sensitivity analysis. *Number*, 251 :93–100.

Borgonovo, E. and Plischke, E. (2016). Sensitivity analysis : a review of recent advances. *European Journal of Operational Research*, 248(3) :869–887.

Breiman, L. (1996a). Bagging predictors. *Machine learning*, 24(2) :123–140.

Breiman, L. (1996b). Out-of-bag estimation.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1) :5–32.

Breiman, L. (2004). Consistency for a simple model of random forests.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). Classification and regression trees.

Browne, T., Fort, J.-C., Iooss, B., and Le Gratiet, L. (2017). Estimate of quantile-oriented sensitivity indices. Technical Report, 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., and Prieur, C. (2015). Generalized sobol sensitivity indices for dependent variables : numerical methods. *Journal of Statistical Computation and Simulation*, 85(7) :1306–1333.

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.

Crestaux, T., Le Maître, O., and Martinez, J.-M. (2009). Polynomial chaos expansion for sensitivity analysis. *Reliability Engineering & System Safety*, 94(7) :1161–1172.

CRO Forum (2017). Leading practices in model management. *CRO Forum*.

Da Veiga, S. (2015). Global sensitivity analysis with dependence measures. *Journal of Statistical Computation and Simulation*, 85(7) :1283–1305.

Davesne, C. (2015). Etude du risque de modèle dans le cadre d'un modèle interne. Master's thesis, SCOR SE, 5 avenue Kléber, 75795 Paris Cedex 16.

Davies, A. and Ghahramani, Z. (2014). The random forest kernel and other kernels for big data from random partitions. *arXiv preprint arXiv :1402.4293*.

Derennes, P. (2019). *Mesures de sensibilité de Borgonovo : estimation des indices d'ordre un et supérieur, et application à l'analyse de fiabilité*. PhD thesis, Université de Toulouse, Université Toulouse III-Paul Sabatier.

Devroye, L., Györfi, L., and Lugosi, G. (2013). *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media.

Díaz-Uriarte, R. and De Andres, S. A. (2006). Gene selection and classification of microarray data using random forest. *BMC bioinformatics*, 7(1) :3.

Dietterich, T. G. (2000). Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer.

Duroux, R. and Scornet, E. (2018). Impact of subsampling and tree depth on random forests. *ESAIM : Probability and Statistics*, 22 :96–128.

Efron, B. (1979). Bootstrap methods : Another look at the jackknife. *The Annals of Statistics*, 7 :1–26.

Efron, B. (1981). Nonparametric standard errors and confidence intervals. *canadian Journal of Statistics*, 9(2) :139–158.

Efron, B. and Stein, C. (1981). The jackknife estimate of variance. *The Annals of Statistics*, pages 586–596.

Elie-Dit-Cosaque, K. (2020). qosa-indices, a python package available at : `https://gitlab.com/qosa_index/qosa`.

Elie-Dit-Cosaque, K. and Maume-Deschamps, V. (2020). Random forest estimation of conditional distribution functions and conditional quantiles. *Preprint on HAL*.

Fabrège, B. and Maume-Deschamps, V. (2020). Conditional distribution forest : a julia package available at `https://github.com/bfabreges/conditionaldistributionforest.jl`.

Faivre, R., Iooss, B., Mahévas, S., Makowski, D., and Monod, H. (2016). *Analyse de sensibilité et exploration de modèles : application aux sciences de la nature et de l'environnement*. Editions Quae.

Fang, K.-T., Li, R., and Sudjianto, A. (2005). *Design and modeling for computer experiments*. CRC Press.

Fisher, R. and Mackenzie, W. (1923). Studies in crop variation : The manurial response of different potato varieties. *Journal of Agricultural Sciences*, 13 :311–320.

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.

Glasserman, P. and Xu, X. (2014). Robust risk measurement and model risk. *Quantitative Finance*, 14(1) :29–58.

Goehry, B. (2019). Random forests for time-dependent processes.

Györfi, L., Kohler, M., Krzyzak, A., and Walk, H. (2006). *A distribution-free theory of nonparametric regression*. Springer Science & Business Media.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized additive models*, volume 43. CRC press.

Helton, J. C. (1993). Uncertainty and sensitivity analysis techniques for use in performance assessment for radioactive waste disposal. *Reliability Engineering & System Safety*, 42(2-3) :327–367.

Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE transactions on pattern analysis and machine intelligence*, 20(8) :832–844.

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.

Iman, R. L. and Conover, W. J. (1979). The use of the rank transform in regression. *Technometrics*, 21(4) :499–509.

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. (2019). Shapley effects for sensitivity analysis with correlated inputs : comparisons with sobol'indices, numerical estimation and applications. *International Journal for Uncertainty Quantification*, 9(5).

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.

Jacques, J. (2005). *Contributions à l'analyse de sensibilité et à l'analyse discriminante généralisée*. PhD thesis, Université Joseph-Fourier - Grenoble I.

Janon, A. (2012). *Analyse de sensibilité et réduction de dimension. Application à l'océanographie*. PhD thesis, Université de Grenoble.

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., 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.

Kala, Z. (2019). Quantile-oriented global sensitivity analysis of design resistance. *Journal of Civil Engineering and Management*, 25(4) :297–305.

Kleijnen, J. P. and Helton, J. C. (1999). Statistical analyses of scatterplots to identify important factors in large-scale simulations, 1 : Review and comparison of techniques. *Reliability Engineering & System Safety*, 65(2) :147–185.

Koenker, R. and Bassett Jr, G. (1978). Regression quantiles. *Econometrica : journal of the Econometric Society*, pages 33–50.

Koenker, R. and Hallock, K. F. (2001). Quantile regression. *Journal of economic perspectives*, 15(4) :143–156.

Kucherenko, S., Song, S., and Wang, L. (2019). Quantile based global sensitivity measures. *Reliability Engineering & System Safety*, 185 :35–48.

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.

Lallement, T. (2014). Le risque de modèle. Master's thesis, SCOR SE, 5 avenue Kléber, 75795 Paris Cedex 16.

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.

Li, G., Rabitz, H., Yelvington, P. E., Oluwole, O. O., Bacon, F., Kolb, C. E., and Schoendorf, J. (2010). Global sensitivity analysis for systems with independent and/or correlated inputs. *The journal of physical chemistry A*, 114(19) :6022–6032.

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.

Lin, Y. and Jeon, Y. (2006). Random forests and adaptive nearest neighbors. *Journal of the American Statistical Association*, 101(474) :578–590.

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.

Marceau, E. (2013). *Modélisation et évaltuation quantitative des risques en actuariat*. Springer Berlin.

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.

Maume-Deschamps, V. and Niang, I. (2018). Estimation of quantile oriented sensitivity indices. *Statistics & Probability Letters*, 134 :122–127.

Maume-Deschamps, V., Rullière, D., and Usseglio-Carleve, A. (2017). Quantile predictions for elliptical random fields. *Journal of Multivariate Analysis*, 159 :1–17.

Meinshausen, N. (2006). Quantile regression forests. *Journal of Machine Learning Research*, 7(Jun) :983–999.

Meinshausen, N. (2019). Quantile regression forests, a r package available at `https://cran.r-project.org/package=quantregforest`.

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.

Morris, M. D. (1991). Factorial sampling plans for preliminary computational experiments. *Technometrics*, 33(2) :161–174.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability & Its Applications*, 9(1) :141–142.

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.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn : Machine learning in Python. *Journal of Machine Learning Research*, 12 :2825–2830.

Planchet, F. and Therond, P.-E. (2012). Risque de modele et détermination du capital économique dans le projet solvabilité 2. *International Review of Applied Financial Issues and Economics*, 3.

Plischke, E., Rabitti, G., and Borgonovo, E. (2020). Computing shapley effects for sensitivity analysis. *arXiv preprint arXiv :2002.12024*.

R Core Team (2019). *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rabitti, G. and Borgonovo, E. (2019). A shapley–owen index for interaction quantification. *SIAM/ASA Journal on Uncertainty Quantification*, 7(3) :1060–1075.

Rachdi, N. (2011). *Apprentissage statistique et computer experiments : approche quantitative du risque et des incertitudes en modélisation*. PhD thesis, Université de Toulouse, Université Toulouse III-Paul Sabatier.

Reserve, F. (2011). Supervisory guidance on model risk management. *Board of Governors of the Federal Reserve System, Office of the Comptroller of the Currency, SR Letter*, pages 11–7.

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. (2006). The critique of modelling and sensitivity analysis in the scientific discourse. an overview of good practices. *Transatlantic Uncertainty Colloquium (TAUC), Oct*.

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, M., et al. (2000). Sensitivity analysis. probability and statistics series. *John and Wiley & Sons, 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.

Scornet, E. (2016a). On the asymptotics of random forests. *Journal of Multivariate Analysis*, 146 :72–83.

Scornet, E. (2016b). Promenade en forêts aléatoires. *MATAPLI*, 111.

Scornet, E. (2016c). Random forests and kernel methods. *IEEE Transactions on Information Theory*, 62(3) :1485–1500.

Scornet, E. (2017). Tuning parameters in random forests. *ESAIM : Proceedings and Surveys*, 60 :144–162.

Scornet, E., Biau, G., and Vert, J.-P. (2015a). Supplementary materials for : Consistency of random forests. *arXiv*, 1510.

Scornet, E., Biau, G., Vert, J.-P., et al. (2015b). 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.

Sibbertsen, P., Stahl, G., and Luedtke, C. (2008). Measuring model risk. Technical report, Diskussionsbeitrag.

Sklar, M. (1959). *Fonctions de répartition à n dimensions et leurs marges*. Institut de Statistique de l'Université de Paris.

Sobol, I. M. (1993). Sensitivity estimates for nonlinear mathematical models. *Mathematical Modelling and Computational Experiments*, 1(4) :407–414.

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.

Stone, C. J. (1985). Additive regression and other nonparametric models. *The annals of Statistics*, pages 689–705.

Stone, C. J. (1994). The use of polynomial splines and their tensor products in multivariate function estimation. *The Annals of Statistics*, pages 118–171.

Vapnik, V. N. and Chervonenkis, A. Y. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2) :264–280.

Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523) :1228–1242.

Wager, S. and Walther, G. (2015). Adaptive concentration of regression trees, with application to random forests. *arXiv preprint arXiv :1503.06388*.

Watson, G. S. (1964). Smooth regression analysis. *Sankhyā : The Indian Journal of Statistics, Series A*, pages 359–372.

Winter, E. et al. (2002). The shapley value. *Handbook of game theory with economic applications*, 3(2) :2025–2054.

Développement de mesures d'incertitudes pour le risque de modèle dans des contextes incluant de la dépendance stochastique

Résumé. Cette thèse s'intéresse aux outils développés dans le domaine de l'Analyse de Sensibilité afin de réaliser une quantification efficiente du risque de modèle structurel, et plus particulièrement le risque lié aux paramètres du modèle. Dans un premier temps, nous adaptons l'algorithme d'estimation des indices de Shapley basés sur la variance pour obtenir des intervalles de confiance en plus de l'estimation mais aussi pour accélérer le calcul des indices. Cela est réalisé en substituant le modèle initial (probablement coûteux en temps de calcul) par un métamodèle de krigeage et nous proposons un algorithme pour prendre en compte l'erreur de métamodélisation dans le calcul des intervalles de confiance. Dans un second temps, deux méthodes d'estimation des fonctions de répartition conditionnelles et des quantiles conditionnels, basées sur les forêts aléatoires, pour lesquelles nous montrons la consistance sont présentées. Reposant sur ces nouvelles stratégies, plusieurs méthodes d'estimation efficaces des indices de sensibilité basés sur les quantiles (QOSA) sont également proposées. Enfin, une étude théorique de ces indices a été réalisée. Il s'avère que leur interprétation peut être délicate en dehors des modèles additifs dans le cas d'entrées indépendantes et pour tout type de modèle en présence de dépendance stochastique entre les entrées. Afin de surmonter ces limitations, nous proposons des *indices de Shapley subordonnés à une caractéristique de la sortie* et, en particulier, des *indices de Shapley orientés quantile*. Ces derniers semblent prometteurs car ils donnent une interprétation claire de l'impact de chaque entrée sur le quantile de la sortie, pour tout type de modèle, à la fois dans le cas d'entrées indépendantes et dépendantes.

Mots-clés : risque de modèle, analyse de sensibilité, forêts aléatoires, indices de Shapley orientés quantile

Development of uncertainty measures for model risk in contexts including stochastic dependence

Abstract. This thesis focuses on the tools developed in the field of Sensitivity Analysis in order to achieve an efficient quantification of the structural model risk, and more particularly the risk related to the model parameters. In a first step, we adapt the estimation algorithm of the variance-based Shapley effects to obtain confidence intervals in addition to the estimation but also to speed up the calculation of the indices. This is achieved by substituting the initial model (probably time-consuming) with a kriging metamodel and we propose an algorithm to take into account the metamodelling error in the calculation of the confidence intervals. In a second step, two methods for estimating conditional distribution functions and conditional quantiles, based on random forests, for which we show the consistency are presented. Based on these new strategies, several efficient estimation methods of the quantile-based sensitivity indices (QOSA) are also proposed. Finally, a theoretical study of these indices has been carried out. It turns out that their interpretation can be tricky outside of additive models in the case of independent inputs and for any type of model in the presence of stochastic dependence between inputs. To overcome these limitations, we propose *Goal-oriented Shapley effects* and, in particular, *Quantile-oriented Shapley effects*. The latter seem promising because they give a clear interpretation of the impact of each input on the output quantile, for any type of model, for both independent and dependent inputs.

Keywords: model risk, sensitivity analysis, random forests, Goal-oriented Shapley effects, Quantile-oriented Shapley effects

UNIVERSITÉ DE LYON