Development of interpretability methods for certifying machine learning models applied to critical systems

Marouane Il Idrissi

RÉPUBLIQUE FRANÇAISE
Liberté Égalité Fraternité

Université de Toulouse

Doctorat de l'Université de Toulouse

préparé à l'Université Toulouse III - Paul Sabatier

Développement de méthodes d'interprétabilité en apprentissage automatique pour la certification des intelligences artificielles reliées aux systèmes critiques

---

Thèse présentée et soutenue, le 4 mars 2024 par

Marouane IL IDRISSI

École doctorale
EDMITT - Ecole Doctorale Mathématiques, Informatique et Télécommunications de Toulouse

Spécialité
Mathématiques et Applications

Unité de recherche
IMT : Institut de Mathématiques de Toulouse

Thèse dirigée par
Jean-Michel LOUBES, Fabrice GAMBOA et Nicolas BOUSQUET

Composition du jury
M. Sébastien DA VEIGA, Président, Ecole Nationale de la Statistique et de l'Analyse de l'Information
Mme Marianne CLAUSEL, Rapporteure, Université de Lorraine
M. Elmar PLISCHKE, Rapporteur, Technische Universität Clausthal
M. Arthur CHARPENTIER, Examinateur, Université du Québec à Montréal
M. Jean-Michel LOUBES, Directeur de thèse, Université Toulouse III - Paul Sabatier
M. Fabrice GAMBOA, Co-directeur de thèse, Université Toulouse III - Paul Sabatier
M. Nicolas BOUSQUET, Co-directeur de thèse du monde socio-économique, EDF

Membres invités
M. Bertrand Iooss, EDF R&D

À Gaspard,
pour tout ce que tu as été,
tout ce que tu es,
et tout ce que tu seras.

REMERCIEMENTS

Il y a une phrase qui m'a beaucoup marquée durant ce doctorat :

"La richesse d'un doctorat ne réside pas dans les pages du manuscrit : le vrai résultat c'est le doctorant lui-même."

Ce que je tire principalement de cette expérience, je ne l'ai pas forcément trouvé dans les pages jaunies des vieux bouquins, mais plutôt dans les merveilleuses rencontres qui m'ont transportées jusqu'à l'ultime rédaction de ces quelques lignes. Commençons naturellement par *mes* quatre encadrants (a.k.a *Les Quatre Fantastiques*), qui m'ont soutenu et guidé pendant ces trois dernières années :

- Jean-Michel, merci de m'avoir fait confiance. Je te suis reconnaissant du fait que, malgré mon parcours non-linéaire, tu as accepté d'être mon directeur et donné cette chance de pouvoir faire mes preuves. Je tiens à te remercier pour ta patience, en particulier lors de nos semaines de travail à l'IMT. Tu m'as beaucoup appris, et par dessus tout, tu as su créer un environnement où j'étais suffisamment en confiance pour pouvoir proposer et explorer des idées (certes, souvent farfelues). Merci de m'avoir permis de passer outre ce satané syndrome de l'imposteur. Au-delà de ton soutien, je tiens aussi à te remercier pour tes intuitions fulgurantes et tes idées de génie, pour tes conseils sur la chose académique, et pour toutes les opportunités que tu as pu me dégoter.

- Fabrice, merci de m'avoir ouvert l'esprit. Tu m'as montré maintes fois que, pour atteindre nos rêves les plus fous, il suffit juste "d'une aprèm et d'un tableau blanc". Merci de m'avoir inculqué l'importance de "faire du foncier", et que des fois c'est mieux de comprendre ce qu'on veut faire avant de le faire. Malgré ton emploi du temps de ministre, tu as toujours su te rendre disponible, que ce soit des retours sur un papier en live lors d'un vol au-dessus de la Géorgie, que nos multiples réunions techniques autour d'un verre sur la place du Capitole. Ton enthousiasme et ta bonne humeur constants sont communicatifs, et ont souvent été le "petit coup de boost" qui m'a permis de surmonter les creux, et de naviguer les sommets. Merci de m'avoir suivi dans mes délires, de m'avoir guidé, et pour tous tes retours et conseils précieux. Travailler avec toi est un plaisir, et j'ai hâte de voir où le futur nous conduira.

- Bertrand, merci de m'avoir guidé. Tu as été l'instigateur et l'architecte de tellement de mes "premières fois" : ma première "vraie" expérience professionnelle, mon premier papier, mon premier entretien d'embauche, ma première conférence, ma première note de frais et j'en oublie très certainement ! Je tiens à te remercier pour tous tes conseils, que ce soit sur le boulot ou sur la vie en générale. C'était un plaisir de partager ta bonne humeur et ton humour pendant ces 4 dernières années. Et merci pour tes innombrables relectures, tes remarques pertinentes, et toute l'énergie que tu as investie pour que nos travaux soient présentés et représentés d'une manière digne de la velléité qu'on leur porte. Je te remercie pour tes encouragements et compliments, qui me vont droit au cœur. Tu sais jongler habilement entre théorie et pratique, en portant une même importance aux deux, un exemple rare qui inspire le chercheur que je tends à vouloir être. J'ai hâte de concrétiser toutes les idées que l'on a déjà entamées, et d'explorer les prochaines (si tu croyais t'être enfin débarrassé de mes papiers à rallonge, c'est râpé !).

- Nicolas, merci pour tout. Notre première rencontre a eu lieu dans cette petite salle de travail à l'ENSAI, où tu avais pris le temps de me pitcher ce combo stage/thèse entre deux soutenances. Merci de m'avoir fait confiance ce jour là, et merci d'avoir toujours répondu présent tout au long des 4 années qui ont suivi. Tu as vraiment fait en sorte de créer une atmosphère propice à ce que je puisse m'épanouir pendant cette thèse. Merci pour tes conseils et ta présence, et pour nos journées de travail qui partaient souvent en questionnement existentiels, mais qui débouchaient souvent sur tant de nouvelles pistes à explorer. Merci pour tes recommandations littéraires, d'avoir pris le temps de m'écouter quand j'étais perdu, et d'avoir toujours trouvé des solutions à tous les problèmes que l'on a rencontrés. Merci pour ton implication dans nos nombreux papiers (et désolé d'avoir saturé ton Overleaf), tes conseils de rédaction, et ta vision globale qui m'a permis de garder le cap. Finalement, merci d'avoir pris soin de moi, lorsque j'oubliais un peu trop de le faire. J'espère vraiment avoir été à la hauteur de tes espérances, et j'ai hâte de voir où nos prochaines tribulations vont nous mener !

Naturellement, après s'être attelés aux personnes qui m'ont suivi depuis le début de ce doctorat, passons à ceux qui m'ont permis de l'achever : le jury de ma soutenance.

- Elmar, thank you so much for your comments on the manuscript : they meaningfully improved it. You even sacrificed countless precious hours in German and French transit just to attend my

defense in-person, to which I am truly grateful. Even though we first met a short few days before the defense, your work has been very inspirational to mine during the course of my PhD. It truly is an honor to have you as one of my rapporteur.

- Marianne, merci infiniment pour votre relecture de mon manuscrit, et pour votre rapport sur mon manuscrit. Je suis très honoré de savoir que mes travaux ont pu intéresser une personne de votre stature. Encore merci d'avoir accepté d'être rapporteure de ma thèse, et j'espère que nos chemins se croiseront peut-être !

- Arthur, je tiens à te remercier pour avoir examiné mon manuscrit, et pour ton accueil plus que chaleureux à Montréal. Ce fut un vrai plaisir d'échanger avec toi ces derniers mois, et j'ai très hâte de pouvoir concrétiser nos idées avec Marie-Pier. Merci à toi pour la confiance que tu m'accordes déjà !

- Sébastien, merci d'avoir présidé mon Jury, et de m'avoir accueillit dans la communauté française de la UQ. J'ai énormément apprécié nos échanges, tant sur le point humain que sur le point technique, au fil des évènements auxquels on s'est croisés. Ton avis m'importe beaucoup, et je n'aurai jamais pu espérer meilleur président pour ma soutenance. Tu es une de ces rencontres qui marquent et qui forgent, et je m'estime chanceux d'avoir pu partager ces moments avec toi, et j'espère qu'ils se multiplieront !

Je tiens également à remercier les personnes avec qui j'ai eu la chance de pouvoir collaborer : Anaïs, Frédérique, Laura, Margot, Sophie et Vincent. Merci pour tout vos conseils et discussions qui m'ont permis de me forger en tant que chercheur. Ce fût un bonheur de se creuser les méninges, coder, rédiger et élaborer des stratégies pour répondre aux éternelles question reloues du reviewer #2 à vos côtés.

Merci à toutes les personnes que j'ai eu la chance de côtoyer à la R&D d'EDF. Vincent pour ton encadrement, ton écoute et tes "Chabridonnages au stylo rouge" qui m'ont tant appris. Sami pour m'avoir inculqué l'art de la CSC, tous tes conseils, et toutes nos sessions cafés à refaire le monde. Julien P. pour ton écoute, ta pratique de l'art mêmesque, et pour toutes ces barres de rire. Théo pour tes petites histoires, et ton soutien inconditionnel pendant nos pauses sur la terrasse du Bâtiment S. Pablo pour ta bonne humeur, tes suggestions et tes conseils : les PDEs ne me font plus aussi peur maintenant ! Charlotte pour ta confiance, tes conseils, ta disponbilité et ta capacité à tout rendre toujours plus sympa. Elias pour ton enthousiasme, nos chouettes moments de colocation Corse, et d'avoir partagé l'anarchie bordélique de nos derniers mois en tant que doctorants. Antoine pour tous tes conseils, et ces pauses café matinales. Sans oublier Alain, Alvàro, Angélique, Audrey, Aurélie, Azénor, Baalu, Brahim, Bruno C., Bruno M., Charlène, Cécile, Coline, Côme, David, Edgar, Emilie, Frank, François, Jérôme, Joseph, Josselin, Julien B., Jérôme, Laura, Louis, Léonardo, Madina, Manu, Margot, Mathieu, Merlin, Michael, Morgane, Paul, Pauline, Pierre-Yves, Salma, Sanaa, Sofiane, Soufiane, Vanessa, et tout ceux qui ont fait qu'aller au boulot était un plaisir.

Je tiens souhaite également remercier les personnes du RT-UQ, de la SFdS, et plus généralement les personnes qui m'ont accueillit à bras ouvert dans notre communauté, et m'ont fait me sentir à ma place : Adrien, Alejandro, Amandine, Antoine, Arthur, Athénaïs, Babacar, Baptiste, Brian, Charles, Charlie, Claire, Clément B., Clément G., Clément H., Cécile, Delphine, Faouzi, Gabriel, Gaël, Guillaume, Julie, Julien, Marie, Noé, Raphaël, Rudy, Salim et compagnie.

Au-delà du déroulement de la thèse, j'aimerai également souligner l'impact de certaines personnes, qui m'ont, d'une manière ou d'une autre, amené à entamer cette thèse. Commençons d'abord par l'équipe enseignante du B.U.T SD (feu D.U.T STID) de Niort, qui m'ont accueillit à bras ouvert à un moment clé de ma vie, m'ont insufflé leur passion pour la chose statistique et informatique, et par-dessus tout, m'ont fait comprendre que tout est atteignable, modulo un peu d'huile de coude. Vous m'avez planté l'idée qu'une thèse pouvait m'être envisageable, et pour ceci, je vous en suis reconnaissant. Je tiens également à remercier les professeurs de l'ENSAI, qui ont su faire germer cette idée, jusqu'à ce qu'elle se concrétise. En particulier, je tiens à remercier François Coquet pour ses conseils justes et avisés. Merci également à Salima El Kolei pour sa confiance et sa disponibilité sans faille pendant mon temps en tant qu'étudiant, mais aussi pour cette merveilleuse opportunité d'enseignement qui m'a tant appris : tu m'as tellement apporté au fil des années, et je t'en suis sincèrement redevable.

Pendant ces trois ans, j'ai également appris à quel point les amitiés sont importantes. Merci à tous les copains, qui ont suivi cette saga, de près ou de loin. Votre soutien m'a permis de tenir quand c'était difficile, de souffler, et de penser à autre chose et surtout. Merci à Lucas pour ces jolies sessions au sud du Maroc et pour nos rendez-vous brunch quotidiens avec ta Julie, à Antoine C. pour ton indé-

fectible présence et d'avoir accepté de nous lier, à Antoine P. pour ton aide mathématique précieuse et ces bonnes tranches de rigolade en compagnie de ton Héloïse, à Yannick pour tous ces moments de création musicale et tes maudites randos. Je remercie aussi Anaïs, Antoine R., Arthur, Asmynour, Aurelle, Auriane, Benoît, Bertrand & Agathe, Camille, Carla, David Lucille Alba et Lucas, Fabien, Gwénel & Juliette, Hugo, Imane, Laure, les Bastiens, Leyna, Lili, Lina, Marie, Maxime & Aliénor, Nikhil, Sarah & Pierre, Thomas, Valentin & Amina, Yann et tous les autres.

Je tiens à dédier ces dernières lignes à ma famille. A commencer par mes parents, dont le soutien indéfectible m'a permis de gravir la montagne universitaire. Merci d'avoir toujours été là, et merci pour tout ce que vous m'avez inculqué pour me permettre de m'épanouir. Je tiens également à remercier Anne-Lise, Christophe, Marie, Marie-Françoise, Jean-Louis et toute la tribu Fasolo-Martinez pour m'avoir accueilli à bras ouvert, et soutenu tout au long de cette thèse. Merci aussi à Jo & Vicky pour tous ces moments à Paris et à Bristol, et pour tout le temps que vous m'avez donné pour pouvoir ajouter la mention *"this article has been proof-read by a native speaker"* pour appuyer nos soumissions. Hind & Nico, merci pour toutes ces bouffées d'air iodées marseillaises, qui était d'autant plus nécessaires quand je suffoquais. Merci pour votre présence, vos conseils, votre soutien, tout votre amour, et merci de nous avoir permis de nous épanouir en tant que tonton et tata. Gaspard, merci d'apporter toute cette joie dans nos vies. Ton arrivée a apporté une nouvelle dimension à ma vie, je me délecte de chaque instant que je passe à tes côtés, et je tiens à te dédier cette thèse.

Enfin, je tiens à remercier la personne sans qui rien de tout ça n'aurait été possible. Léanne, merci pour ton amour et ton soutien. A mes yeux, cette thèse est autant la mienne que la tienne et rien n'est impossible à tes côtés. Je suis fier de toi et de tout ce que l'on a accompli tous les deux. Tu rends l'incertain excitant, et me donne envie de croquer la vie à pleine dents. Je t'aime et j'ai hâte de vivre pleinement tout ce que le futur nous réserve.

Contents

Remerciements i

Abbreviations and symbols viii

Abstracts ix

List of contributions x

General introduction xi

1 The need for interpretability 1
- 1.1 Context and motivations . . . 2
- 1.2 A mathematical framework for model interpretability . . . 6
- 1.3 Two main interpretability methods . . . 9
- 1.4 Illustrative use-cases . . . 10
- 1.5 Intention and content of the manuscript . . . 14

2 Decompositions of quantities of interest 17
- 2.1 Input influence assessment for interpretability . . . 19
- 2.2 Coalitional decompositions and influence measures . . . 19
- 2.3 Importance quantification and the variance as a QoI . . . 24
- 2.4 Partial conclusion . . . 28

3 Input-centric approach and cooperative games 29
- 3.1 Introduction . . . 31
- 3.2 Cooperative games and allocations . . . 31
- 3.3 Importance attribution with dependent inputs . . . 37
- 3.4 Illustration on use-cases . . . 44
- 3.5 The fundamental problem of the input-centric approach . . . 50

4 Model-centric approach and output decomposition 53
- 4.1 Introduction . . . 55
- 4.2 Preliminaries . . . 56
- 4.3 Coalitional output decomposition with dependent inputs . . . 62
- 4.4 Model-centric influence measures . . . 70
- 4.5 Analytical example: two Bernoulli inputs . . . 74
- 4.6 Conclusion . . . 75

5 Robustness to input perturbations 77
5.1 Assessing robustness by perturbing inputs . . . 79
5.2 Perturbing quantiles . . . 80
5.3 Wasserstein projections . . . 85
5.4 Computing the perturbed distributions . . . 87
5.5 Illustration on use-cases . . . 90
5.6 Discussion . . . 97

6 Conclusion and perspectives 99
6.1 Conclusion . . . 100
6.2 Perspectives . . . 101

References 103

A Measure and probability theory preliminaries 115

B Supplementary material for Chapter 2 117
B.1 Preliminaries on sets and orders . . . 118
B.2 Proofs . . . 120

C Supplementary material for Chapter 3 121
C.1 Estimators for the conditional elements . . . 122
C.2 Proofs . . . 123

D Supplementary material for Chapter 4 127
D.1 Some technical preliminaries . . . 128
D.2 Proof of Theorem 4.6 . . . 128
D.3 Proofs . . . 132
D.4 Analytical results . . . 135

E Supplementary material for Chapter 5 147
E.1 Some preliminaries . . . 148
E.2 Dependence modelling and copulas . . . 150
E.3 Computational details and code snippets . . . 150
E.4 Proofs . . . 152
E.5 Proof of Theorem 5.1 . . . 154

F Additional use-cases 161
F.1 A COVID-19 epidemiological model . . . 162
F.2 Ultrasonic non-destructive control of a weld . . . 166
F.3 Robot arm model . . . 169

G Résumé étendu 173
G.1 Contexte et motivation . . . 174
G.2 Analyse de sensibilité et interprétabilité post-hoc . . . 175
G.3 Un cadre mathématique pour l'interprétabilité des modèles . . . 178
G.4 Conundrums et méthodes d'interprétabilité . . . 180
G.5 Deux méthodes d'interprétabilité . . . 181
G.6 Articulation du manuscrit . . . 183

ABBREVIATIONS AND SYMBOLS

Abbreviation	Description	Reference
e.g.,	*Exempli gratia:* for example.	
i.e.,	*Id est:* that is.	
EDF	Électricité de France.	Page 2
EDF R&D	Research and development branch of Électricité de France.	Page 2
UQ	Uncertainty quantification.	Page 2
ML	Machine learning.	Page 2
SA	Sensitivity analysis.	Page 3
XAI	Explainable artificial intelligence.	Page 3
w.r.t.	With respect to.	Page 3
QoI	Quantity of interest.	Page 6
trunc.	Truncated.	Page 11
DoE	Design of experiment.	Page 13
poset	Partially ordered set.	Page 22
FANOVA	Functional analysis of variance.	Page 25
PME	Proportional marginal effects.	Page 40
PMVD	Proportional marginal variance decomposition.	Page 40
LMG	Lindeman-Merenda-Gold indices.	Page 40
GP	Gaussian process.	Page 48
HDMR	High-dimensional model representation.	Page 55
cdf	Cumulative distribution function.	Page 80
gqf	Generalized quantile function.	Page 80

Notation	Description
d	Positive integer representing the number of inputs.
D	$D = \{1, \ldots, d\}$, the set of input indices, or grand coalition of players.
$(\Omega, \mathcal{F}, \mathbb{P})$	Sample space.
$(E, \mathcal{E})$	Input space.
X	Random inputs.
P_X	Joint distribution of the random inputs.
σ_X	σ-algebra generated by the random inputs.
X_A	Subset of the random inputs.
σ_A	σ-algebra generated by the subset of the random inputs X_A.
P_{X_A}	Marginal distribution of a subset of the random inputs.
G	Black-box model.
$(Y, \mathcal{Y})$	Output space.
$G(X)$	Random output.
$\mathcal{G}_X$	Space of random outputs.
$(Q, \mathcal{Q})$	QoI space.
$\mathcal{P}_D$	Power-set of the set D, i.e., the set of subsets of D, including $\emptyset$.
ϕ	An influence measure, i.e., a function from $\mathcal{P}_D$ to Q.
v	A value measure or a value function.
ψ	An allocation.
$\mathcal{D}_v$	The Harsanyi dividend of a cooperative game with value function v.
Sh	Shapley effects.
S^T, S^{clos}	Total and closed Sobol' indices.
$\mathbb{L}^2(\sigma_X), \mathbb{L}^2(\sigma_A)$	Lebesgue space of square-integrable σ_X- (resp. σ_A-) measurable functions.
$\mathbb{E}_A[\cdot]$	Conditional expectation w.r.t. σ_A, i.e., orthogonal projection onto $\mathbb{L}^2(\sigma_A)$.
$\mathbb{M}_A[\cdot]$	Canonical oblique projection onto $\mathbb{L}^2(\sigma_A)$.
$\mathcal{D}$	Discrepancy between probability measures.
$\widetilde{X}$	Perturbed inputs.
$c_0(\cdot), c(\cdot)$	Dixmier's and Friedrichs' angles.

ABSTRACTS

Abstract (Français). Les algorithmes d'apprentissage automatique, qui ont énormément contribué à l'essor de l'intelligence artificielle (IA) moderne, ont démontré à maintes reprises leur haute performance pour la prévision de tâches complexes. Cependant, malgré le gain manifeste évident lié à l'utilisation de ces méthodes pour l'accélération et l'amélioration de la performance de tâches d'ingénierie variées (mise en relation d'informations collectées par des capteurs, détection de signaux rares, etc.), incluant en particulier la modélisation de systèmes critiques industriels (temps de calcul, valorisation de données récoltées, hybridation entre la physique et les données expérimentales), la modélisation par apprentissage automatique n'est toujours pas largement adoptée dans les pratiques d'ingénierie moderne. Les résultats empiriques des modèles appris sur certains jeux de données (benchmarks) ne suffisent pas à convaincre les instances de sûreté et de contrôle en charge des activités industrielles.

Cette thèse a pour but de développer des méthodes permettant la validation de l'usage de modèles boîtes-noires (dont les IA) par le biais de l'étude des incertitudes. Un formalisme mathématique global est proposé pour l'étude théorique des méthodes d'interprétabilité des modèles boîtes noires. Ce travail méthodologique permet de rapprocher deux domaines très proches : l'analyse de sensibilité (SA) des modèles numériques et l'interprétabilité post-hoc. Deux thématiques concrètes sont au cœur des travaux de cette thèse : la quantification d'influence et l'étude de robustesse face aux perturbations probabilistes. Une attention particulière est portée au cadre et aux propriétés théoriques des méthodes proposées dans le but d'offrir des outils convaincants allant au-delà des considérations empiriques. Des illustrations de leur utilisation, sur des cas d'études issues de problématiques réelles, permettent d'étayer la cohérence de leur utilisation en pratique.

La situation d'entrées dépendantes, c'est-à-dire lorsque les entrées du modèles boîte-noire ne sont pas supposées mutuellement indépendantes, prennent une place importante dans les travaux menés. Cette considération a permis la généralisation de méthodes existantes en SA et en intelligence artificielle explicable (XAI). Au-delà de leurs propriétés théoriques pertinentes, ces nouvelles méthodes sont davantage cohérentes avec les études pratiques, où les données récoltées sont souvent corrélées. En particulier, un stratagème de perturbation probabiliste conservant cette dépendance fondé sur des méthodes de transport optimal est proposé. De plus, une généralisation sous des hypothèses peu restrictives de la décomposition fonctionnelle d'Hoeffding est également démontrée. Elle permet d'étendre à un contexte non mutuellement indépendant une multitude de méthodes déjà existantes et utilisées en pratique. Les travaux présentés sont en lien étroit avec différents domaines mathématiques : statistiques, probabilités, combinatoire algébrique, optimisation, transport optimal, analyse fonctionnelle et théorie des jeux coopératifs. Plusieurs liens entre ces disciplines sont établis afin d'offrir une vision générale de l'étude d'interprétabilité des modèles boîtes-noires.

Abstract (English). Machine learning algorithms, which have significantly contributed to modern artificial intelligence (AI) advancement, have repeatedly demonstrated their performance in predicting complex tasks. However, despite the potential benefits of using these methods for modeling critical industrial systems (computation time, data value, hybridization between physics and experimental data), these algorithms have not yet been widely adopted in modern engineering practices. Empirical results on benchmark datasets do not seem sufficient to convince safety and control authorities responsible for industrial activities.

This thesis aims to develop methods for validating the use of black-box models (particularly those embedded in AI systems) through the study of uncertainties. A general and comprehensive mathematical formalism is proposed for the theoretical study of black-box model interpretability methods. This methodological work unifies two closely related research areas: sensitivity analysis (SA) of numerical models and post-hoc interpretability. Two central themes to this thesis are influence quantification and robustness to probabilistic perturbations. Special attention is paid to the framework and theoretical properties of the proposed methods to provide compelling tools that go beyond empirical considerations. Illustrations of their use on real-world problem cases support the consistency of their practical use.

The consideration of dependent inputs, i.e., when the inputs of the black-box models are not assumed to be mutually independent, plays a significant role in the research conducted. This consideration has allowed the generalization of existing methods in SA and explainable artificial intelligence (XAI). Beyond their relevant theoretical properties, these new methods are more consistent with practical studies, where collected data is often correlated. In particular, a probabilistic perturbation strategy that preserves this dependence based on optimal transport methods is proposed. Furthermore, a generalization under non-mutually independent assumptions of the Hoeffding functional decomposition is also demonstrated. It allows the extension of a multitude of existing methods used in practice. The presented work is closely related to various mathematical domains: statistics, probability, algebraic combinatorics, optimization, optimal transport, functional analysis, and cooperative game theory. Several connections between these disciplines are established to provide a global and comprehensive view of black-box model interpretability research.

LIST OF CONTRIBUTIONS

Published journal articles

M. Herin, M. Il Idrissi, V. Chabridon, and B. Iooss. Proportional marginal effects for global sensitivity analysis. *SIAM/ASA Journal on Uncertainty Quantification*, 2024. URL: https://hal.science/hal-03825935. In press

M. Il Idrissi, N. Bousquet, F. Gamboa, B. Iooss, and J. M. Loubes. On the coalitional decomposition of parameters of interest. *Comptes Rendus. Mathématique*, 361:1653–1662, 2023. DOI: 10.5802/crmath.521

M. Il Idrissi, V. Chabridon, and B. Iooss. Developments and applications of Shapley effects to reliability-oriented sensitivity analysis with correlated inputs. *Environmental Modelling and Software*, 143:105115, 2021. ISSN: 1364-8152. DOI: 10.1016/j.envsoft.2021.105115

Pre-prints

M. Il Idrissi, N. Bousquet, F. Gamboa, B. Iooss, and J. M. Loubes. Hoeffding decomposition of black-box models with dependent inputs. Preprint, 2023. URL: https://hal.science/hal-04233915

A. Foucault, M. Il Idrissi, B. Iooss, and S. Ancelet. Shapley effects and proportional marginal effects for global sensitivity analysis: application to computed tomography scan organ dose estimation. Preprint, 2023. URL: https://hal.science/hal-04114533

M. Il Idrissi, N. Bousquet, F. Gamboa, B. Iooss, and J. M. Loubes. Quantile-constrained Wasserstein projections for robust interpretability of numerical and machine learning models. Preprint, 2023. URL: https://hal.science/hal-03784768

L. Clouvel, B. Iooss, V. Chabridon, M. Il Idrissi, and F. Robin. A review on variance-based importance measures in the linear regression context. Preprint, 2023. URL: https://hal.science/hal-04102053

Conferences (papers, abstracts, posters, talks)

M. Il Idrissi, N. Bousquet, F. Gamboa, B. Iooss, and J. M. Loubes. Hoeffding decomposition, revisited. In *SIAM Conference on Uncertainty Quantification 2024*, Trieste, Italy, 2024. URL: https://www.siam.org/conferences/cm/conference/uq24

M. Il Idrissi, N. Bousquet, F. Gamboa, B. Iooss, and J. M. Loubes. Cooperative game theory and importance quantification. In *23rd European Young Statisticians Meeting of the Bernoulli Society*, Ljubljana, Slovenia, 2023. URL: https://sites.google.com/view/eysm2023

M. Il Idrissi, N. Bousquet, F. Gamboa, B. Iooss, and J. M. Loubes. Coalitional decomposition of quantities of interest. In *2023 Annual Meeting of MASCOT-NUM Research Group*, Le Croisic, France, 2023. URL: https://mascotnum2023.sciencesconf.org/

M. Il Idrissi, N. Bousquet, F. Gamboa, B. Iooss, and J. M. Loubes. Projection de mesures de probabilité sous contraintes de quantile par distance de Wasserstein et approximation monotone polynomiale. In *53èmes Journées de Statistique de la Société Française de Statistique (SFdS)*, Lyon, France, 2022. URL: https://hal.science/hal-03597059

M. Il Idrissi, N. Bousquet, F. Gamboa, B. Iooss, and J. M. Loubes. Robustness assessment of black-box models using quantile-constrained wasserstein projections. In *2022 Annual Meeting of MASCOT-NUM Research Group*, Clermont-Ferrand, France, 2022. URL: https://mascotnum2022.sciencesconf.org/. (Poster)

M. Il Idrissi, V. Chabridon, and B. Iooss. Shapley effects for reliability-oriented sensitivity analysis with correlated inputs. In *Proceedings of the 10th International Conference on Sensitivity Analysis of Model Output*, Tallahassee, Florida, United States of America, 2022. URL: https://samo2022.math.fsu.edu/

M. Il Idrissi, B. Iooss, and V. Chabridon. Mesures d'importance relative par décomposition de la performance de modèles de régression. In *52èmes Journées de Statistique de la Société Française de Statistique (SFdS)*, Nice, France, 2021. URL: https://hal.science/hal-03149764

General introduction

The topic of this CIFRE PhD originates from the close collaboration between EDF R&D and l'Université Toulouse III - Paul Sabatier. More precisely, this PhD has been hosted in the *PRISME* (Performance, Risque Industriel, Surveillance pour la Maintenance et l'Exploitation) department of EDF R&D, in the *GAIA* (Gestion d'Actifs, Incertitudes et Apprentissage) group. More generally, it was part of the TURING project, whose primary goal is to develop, experiment, and spread the bleeding edge of AI technologies to answer EDF's future electricity production and distribution challenges. A part of this collaboration involved the SINCLAIR (Saclay Industrial Collaborative Laboratory for Artificial Intelligence Research) laboratory, founded as a collaborative effort towards AI research between significant actors from the French industrial space. From an academic perspective, this PhD has also been hosted at the IMT (Institut de Mathématiques de Toulouse), whose scope covers all the domains of fundamental and applied mathematics.

This PhD aims to study the behavior of black-box models of critical systems and, more precisely, to strive towards a theoretically-grounded methodology for their interpretation. These models can be complex numerical simulators of physical phenomena or machine learning models with uncertain inputs. The developments in this thesis can be understood as an exploration of the mathematical aspects surrounding the interpretation of black-box models. This exploration aims to contribute to the establishment of a robust foundation for interpreting black-box models, ensuring their trustworthiness and effectiveness within the context of critical systems. The research outcomes are anticipated to bridge the gap between theoretical insights and practical applications, enhancing the reliability and interpretability of these models in critical domains and, hopefully, are a step towards establishing confidence in their use.

This manuscript comprises six chapters containing and expanding on the main scientific contributions made during this PhD.

Chapter 1 presents the overall context of this PhD and the scope of the presented contributions. It introduces a general mathematical framework for model interpretability upon which the following developments refer. Two central interpretability questions are introduced: input influence quantification and robustness assessment. Three main use-cases are presented and further studied in the remainder of the manuscript.

Chapter 2 explores measuring the influence of inputs. A link is drawn with the domain of combinatorics, leading to two approaches to building influence measures: an input-centric approach and a model-centric approach.

In **Chapter 3**, the input-centric approach is described using already-established interpretability methods relying on the framework of cooperative game theory. The use of allocations, i.e., the redistribution of resources, for quantifying importance is presented and discussed. Some methodological drawbacks to this approach are presented.

Chapter 4 focuses on the model-centric approach. It requires the ability to decompose black-box models with uncertain inputs. It is shown that such decompositions can be achieved whenever the inputs are dependent and lead to the definition of intuitive and theoretically grounded importance measures.

Chapter 5 presents the question of the robustness of black-box models and explores one of its aspects: the behavior changes whenever the inputs of a model are perturbed. It allows defining a generic methodology, which can, in-fine, be used to qualitatively assess these models' behavior under these perturbations.

Chapter 6 contains final concluding remarks, discusses the challenges ahead, and exposes promising perspectives to the presented work.

Appendices A to E contain technical preliminaries, the proofs of the results presented in each chapter, computational details, or more in-depth methodological regarding the use-cases to make this manuscript as self-contained as possible.

Appendix F presents three additional use-cases. A Covid-19 epidemiological model, the ultrasonic non-destructive control of a weld, and the study of a robot arm. These use-cases showcase the methods developped in the manuscript.

Finally, **Appendix G** contains an extended summary of the manuscript, written in French.

Additionally, a GitHub repository[1] containing all the codes to reproduce the presented results.

[1] https://github.com/milidris/phdThesis

Chapter 1

The need for interpretability

Contents

1.1 Context and motivations 2
1.1.1 Black-box modeling of complex critical systems 2
1.1.2 Sensivitity analysis meets post-hoc interpretability 2
1.1.3 Model interpretability for black-box models of critical systems 5
1.2 A mathematical framework for model interpretability 6
1.2.1 Black-box modeling: Random inputs, black-box model, and random output . . 6
1.2.2 Quantity of interest 7
1.2.3 Conundrums and interpretability methods 8
1.3 Two main interpretability methods 9
1.3.1 QoI decomposition for influence assessment 9
1.3.2 Input perturbations for the assessment of model robustness 10
1.4 Illustrative use-cases 10
1.4.1 Simplified hydrological model 11
1.4.2 Transmittance error of an optical filter 12
1.4.3 Acoustic fire extinguisher dataset 13
1.5 Intention and content of the manuscript 14

1.1 Context and motivations

1.1.1 Black-box modeling of complex critical systems

Instinctively, when confronted with physical phenomena, a natural reflex would be to perform repeated experiments. Traditionally, engineering studies would aim at extracting insights from these experiments by testing different configurations (e.g., changing the environment or initial conditions), recording the different results, and comparing them, treating the world as an experimental arena. This is what kick-started the field of *experimental physics*. However, as innovation blossomed and industrial needs grew larger in proportion and ambition, performing such experiments quickly became too costly, dangerous, too complex, or simply impossible to set up. More recently, *modern engineering* offered a solution: replace the experimental configurations with *physical models* of the phenomena, which would result in *numerical simulations* of the studied phenomena.

These *numerical models* proved their usefulness by shaping modern industrial practices. For example, determining the ad-hoc profitability of wind farms depending on their location, improving the design of nuclear power plants to ensure the utmost safety and prevent accidents, or their robustness to catastrophic events (e.g., natural disasters, targeted attacks). Électricité de France (EDF), and in particular its research and development branch (EDF R&D), plays an essential role in the development, certification, and spread of these numerical models for the electricity production industry[1]. Since these numerical models simulate *critical systems*, their reliability became paramount for the decision-making processes in industrial practices.

However, as these tools grew in ambition along with the industrial needs, they became *too complex to study analytically*, and performing simulations of the physical phenomenon, despite access to tremendous computing capabilities, *took increasingly longer and longer*. Additionally, some physical models encapsulate convoluted equations (e.g., Navier-Stokes equations), which can sometimes only be solved numerically. Due to their sheer complexity, these numerical models were considered as *black-boxes*.

Many of these simulations conducted over the years were consolidated in databases, in addition to the rapid improvement of sensing tools, standardizing the recording of on-site measurements. This affluence of data, coupled with the impressive performance of supervised learning methods for modeling complex phenomena, begged the question: *How can these data-driven modeling methods benefit industrial processes?*

In particular, these methods offer a solution to the ever-growing, time-consuming simulations of numerical models by offering fast-to-evaluate surrogates. They also promise to leverage sensing data to model complex phenomena that have not been modeled numerically yet or simply cannot be. However, since reliability is a primary focus in industrial engineering, confidence in these methods must be assessed for their adoption as part of critical system modeling. Recent advances in artificial intelligence led to the resurgence of over-parameterized but very effective machine learning models, also considered black-boxes.

The main difficulty comes from the fact that critical systems are usually subject to uncertainties. These uncertainties can stem from various reasons (e.g., lack of knowledge, measurement errors, or intrinsic to the studied phenomena). Understanding and controlling the effects of these uncertainties on the critical system is paramount for industrial decision-making and remains an active area of research. Dealing with uncertainties is a challenge in industrial engineering when dealing with black-box numerical models, but also in artificial intelligence when it comes to black-box supervised learning models.

The work presented in this thesis mainly revolves around the uncertainties surrounding black-box models (numerical or learned from data). This section offers a view on the "what, why, and how" uncertainties can be managed in industrial engineering and machine learning.

1.1.2 Sensivitity analysis meets post-hoc interpretability

In the contents of this thesis, a first parallel is drawn between two fields of applied mathematics. The first, *uncertainty quantification* (UQ), is deeply rooted in the study of uncertainties propagated in numerical models. In contrast, the second *machine learning* (ML) stems from the marriage between statistical

[1] https://www.edf.fr/en/the-edf-group/inventing-the-future-of-energy/rd-global-expertise/our-offers/our-software-and-calculation-codes

learning and computer sciences. However, while their goals may be fundamentally different, they share surprising similarities when drawing insights on the (numerical or learned) model of interest. In particular, many of the goals of *sensitivity analysis* (SA) are shared with *post-hoc interpretability*, a sub-field of explainable artificial intelligence (XAI) [182, 16], as described in the following.

Sensitivity analysis. Paraphrasing [182], in a nutshell, the field of SA can be summarized as:

> "The study of how the outputs of a system are related to, and are influenced by, its inputs. "

Historically, SA has been part of the UQ methodology [57], where the main goal is to extract insights from "black-box" computer models. These models are often specified in order to simulate physical phenomena, such as thermo-mechanical modeling for the structural analysis of manufacturing processes[2], or assessing the safety of industrial installations[3]. They are often comprised of a series of complex mathematical operations (e.g., solvers for differential equations, finite element models) designed by domain experts (e.g., physicists) to best approximate real-world physical phenomena' behavior. These models are crucial in industrial studies since they offer a cheaper and safer but complementary alternative to controlled repeated experiments.

These numerical models can be seen as "input-output" systems, where the inputs can represent initial conditions and related physical quantities (e.g., ambient temperature, pressure, humidity). In the UQ methodology, these inputs are considered as *uncertain*, either due to a reducible lack of knowledge (i.e., epistemic uncertainty) or due to controlled uncertainties (e.g., measurement errors). The uncertainties of the system are identified and quantified, and the inputs are subsequently endowed with a probabilistic structure (by domain experts' opinion or through real-world observations). In turn, the system's output also becomes random, better known as the *propagation of uncertainties* step in the UQ methodology. This is where SA comes into play. Given random inputs and a subsequent random model output, sensitivity analyses aim to draw insights from the modeled phenomena. In particular, four settings are of interest [48]:

- **Model exploration:** investigating the input-output relationship in the uncertain context to understand the behavior of the model better;
- **Factor fixing:** detect the "un-important" inputs (i.e., whose uncertainties have a limited impact on the output's uncertainty) to exclude them from the uncertainty study (by considering them as constants);
- **Factor prioritization:** identify the "most important" inputs, i.e., the ones whose uncertainty affects the output's (or a quantity of interest's) uncertainty the most;
- **Input distribution robustness:** study the variations of the output's distribution (or a quantity of interest) with respect to (w.r.t.) changes in the input's chosen probabilistic structure.

These settings can be approached either from a local (i.e., on a neighborhood around a particular input value) or a global (i.e., on the whole domain of the inputs) standpoint [158]. Many statistical methods have been developed in the SA literature to provide practical tools for these settings [120, 30, 48]. These tools provide *diagnostics* to the practitioner. These diagnostics can be understood as estimates of the quantities the SA method quantifies. Depending on the question at hand and the choice of the SA method, these diagnostics are an aid for scientific discoveries (e.g., improving the understanding of the studied phenomena) or for engineering extents (e.g., to assist decision-making processes). Figure 1.1 illustrates how and when sensitivity analyses can be performed to draw insights from numerical models.

Post-hoc interpretability. Taking inspiration from [16], in a nutshell, post-hoc interpretability can be summarized as:

> "The ability to explain and provide reasons for the behavior of a given ML model."

ML aims to offer tools for modeling various phenomena from observed data. Given a set of observations of input variables (i.e., features) and output variables (i.e., target) forming an observed *dataset*, the

[2] e.g., the use of the `code_aster` computer code for wire-arc additive manufacturing [101].
[3] e.g., the use of `CATHARE2` numerical code for loss-of-coolant incidents in nuclear power plants [3].

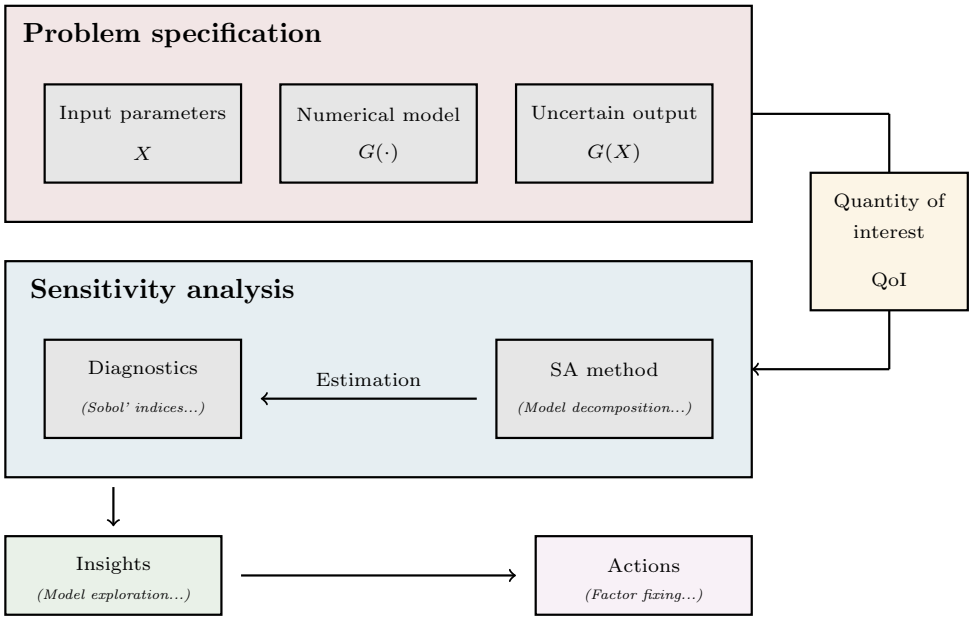

Figure 1.1: Sensitivity analysis to draw insights from numerical models.

overall aim is to produce a model able to fit the output best given the input. Traditionally, in the field of statistical learning, the input and the output variables are assumed to be random, and the dataset is comprised of realizations of these random variables [214]. However, the probabilistic structure of both the inputs and the outputs is often unknown and only observed. The true relation that links the inputs to the outputs is also unknown. Designing an ML model consists of assuming that this true relation can be approximated using a model belonging to a particular family (e.g., a linear model, an auto-regressive process, a neural network), often from a family of models characterized by parameters (e.g., linear coefficients, auto-regressive coefficients, weights and biases of neurons). The learning process can be described as leveraging the data to find the "best" values for these parameters, in the sense that they minimize an (empirical) error between the observed target values and values predicted by the model [97].

ML can be used in two related but fundamentally different settings:

- **Input-output relationship exploration:** determine if there exists a significant relationship between the input and the output and if there is, its nature (e.g., linear, nonlinear);
- **Predictive performance:** build the best-performing model to achieve a certain predictive task with high accuracy.

In statistics, ML models were seen as a tool for studying multivariate links, as a step up from traditional univariate and bivariate statistics [222]. Combined with the framework of hypothesis testing, the first setting was the main concern, which proved useful when applied to many areas of research (e.g., economics [7], biology [153], medicine [43], industrial processes [132]). In light of the powerful nature of such an approach, the second setting has recently seen an increasing amount of attention [124], especially with the introduction of deep learning approaches [90], which accomplished near-perfect prediction scores on highly non-trivial tasks (e.g., digit recognition [139], image classification [130]).

However, as the nature of the predictive tasks at hand became more and more challenging, the sheer complexity of the best-performing models grew accordingly, often endowed with an enormous number of parameters. These high-performing models were thus considered as "black-boxes". While the theory behind the learning process is well established [97], the mathematical reason behind why such over-parametrized models show such impressive performance is still unknown [163]: it is easy to show that a model works, but it is way more complicated to understand why. However, with the abundance of various data streams and the increasing efficiency of computing power, these models are attractive for modeling critical systems. Nevertheless, for such an adoption, the first setting is crucial for many domains: the decision-making process must be built upon theoretical guarantees to convince the relevant

safety and control authorities.

The field of XAI stemmed from this need to understand these black-box algorithms better [16]. In a nutshell, it encompasses every aspect of the "artificial intelligence explanation" process, from the development of suitable tools to the study of the interaction between the ML modeler and domain experts. *Post-hoc interpretability* is a part of the XAI field. The adjective "post-hoc" refers to the fact that the ML model of interest is already trained: the focus is put on trying to extract insights on the behavior of a specific model (i.e., with a fixed set of parameters) rather than developing novel families of "interpretable" models. A (non-exhaustive) list of settings that post-hoc interpretability aims at addressing is as follows [16]:

- **Trustworthiness:** the confidence of whether a model will act as intended when facing a given problem;
- **Transferability:** elucidation of the boundaries that might affect a model, allowing for a better understanding and implementation of unseen data;
- **Informativeness:** extracting information about the inner relations of a model;
- **Confidence:** ensure the robustness and stability of a model in which reliability is expected;
- **Fairness:** assess if a model is influenced by protected inputs, which may lead to unfair or unethical treatments.

These settings can be approached either from a local (i.e., on a particular prediction instance) or a global (i.e., on the whole domain of the inputs) standpoint [156]. Many methods have been proposed in the literature but are often justified empirically through popular benchmarks [16]. Figure 1.2 illustrates how and when post-hoc interpretability can be performed to draw insights from learned ML models.

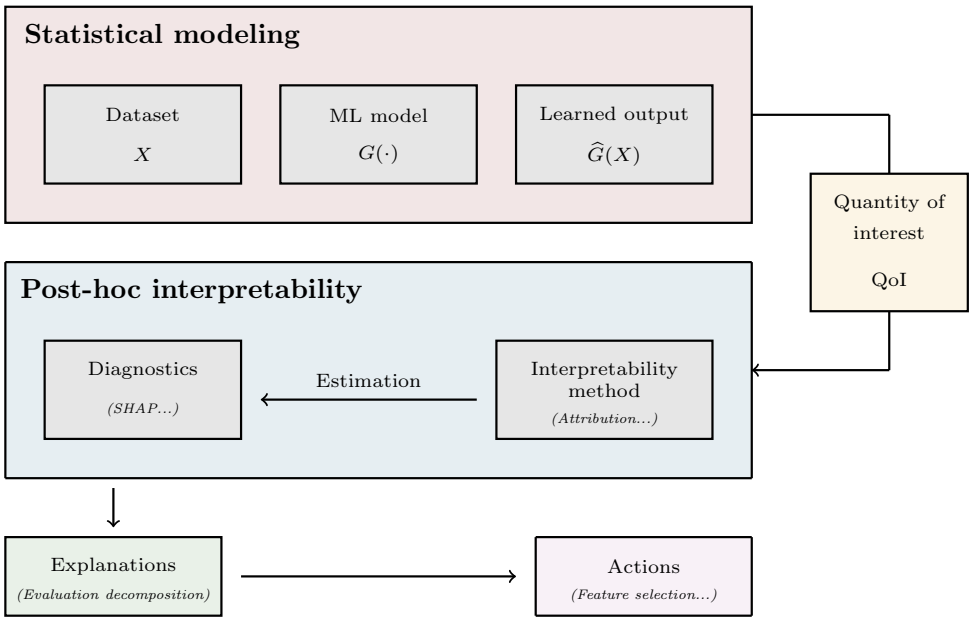

Figure 1.2: Post-hoc interpretability to draw insights from learned ML models.

1.1.3 Model interpretability for black-box models of critical systems

As the attentive reader may have noticed, post-hoc interpretability and SA share many aspects, and this overlap has been highlighted in the literature [182, 29, 142]. The work presented in this thesis is at the cornerstone between both fields in pursuing one particular goal: *the development of theoretically-grounded methods to interpret black-box models of critical systems to justify their adoption in practice*. The focus is put on *model-agnostic* methods, in the sense that they must not rely on a particular modeling structure since black-boxes can come in many different shapes and forms. Furthermore, they must be *theoretically-*

justified: more than empirical evidence is needed to adopt these methods for studying critical systems. Their characterization must be theoretically grounded and understood, their properties must be studied, their limits must be highlighted, and the meaning of the insights they bring forward must be clearly stated.

The starting point is to leverage the theoretical framework of UQ and SA due to its historical success in enabling the adoption of (black-box) numerical models for engineering studies. A unified mathematical framework is proposed to bridge the gap between SA and post-hoc: *model interpretability*. However, when it comes to fitting to the XAI needs, two challenges must be addressed: **dependence between the inputs** and **unknown probabilistic data-generating process** (i.e., the practitioner only has access to an observed dataset). These two constraints are at the heart of the developments made in this thesis. The proposed mathematical framework of model interpretability is introduced and discussed in the following section.

1.2 A mathematical framework for model interpretability

This section introduces the mathematical framework of model interpretability and the first set of notations that will be used in the remainder of the manuscript. To accommodate both SA and post-hoc interpretability, this probabilistic framework relies on a relatively general measure-theoretic standpoint. The interested reader is referred to [122] for some preliminaries and Appendix A for the relevant definitions of measure theory and probability theory. The following elements are introduced, defined, and discussed:

- **Random inputs:** they represent the uncertain inputs of numerical models or the relevant observed features related to an ML model. In this framework, random inputs take the form of vectors of *random elements*;
- **Black-box model:** they represent the black-boxes used to model (critical) systems. They can represent a numerical model of a physical phenomenon or an ML model trained from data. In this framework, black-box models take the form of functions mapping two suitable spaces;
- **Random output:** reminiscent of the *propagation of uncertainty* paradigm of UQ, random outputs are the evaluation of black-box models on the random inputs, thus becoming a random element valued in the codomain of the black-box model;
- **Quantity of interest:** it represents a meaningful quantity related to the random output, of which the effects of the random inputs need to be studied (e.g., a particular evaluation of a model, its variance). Quantities of interest (QoIs) are defined as mappings between the codomain of the black-box model to suitable types of spaces;
- **Interpretability methods:** they represent ways to solve a clearly stated *conundrum*, i.e., a key practical question one wishes to gain insight from.

In the remainder of the manuscript, the following notations are adopted. $\subset$ indicates a proper (strict) inclusion between two sets, while $\subseteq$ indicates that equality is possible, and for any set A, the set $\{B : B \subseteq A\}$ does not contain the empty set denoted $\emptyset$.

1.2.1 Black-box modeling: Random inputs, black-box model, and random output

This first section defines the elements that compose the first step of the proposed framework: *black-box modeling*. The focus is put on *what* the introduced notions can be formalized as, rather than *how* can one achieve black-box modeling (e.g., numerical codes, ML model).

Random inputs To accommodate both SA and XAI settings, many different "types" of inputs must be considered, which are not necessarily $\mathbb{R}$ valued (e.g., , non-tabular data such as text, images, and time-series). Hence, taking inspiration from [48], the random inputs are defined using the very general notions of *random elements* and *vectors of random elements* (see, Appendix A), which generalize the idea of random variables and random vectors.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be an abstract probability space (often referred to as the *sample space*), let $d \geq 1$ be a positive integer, and let $(E_1, \mathcal{E}_1), \ldots, (E_d, \mathcal{E}_d)$ be a collection of standard Borel measurable spaces. For every $A \subset D$, denote:

$$E_A := \bigtimes_{i \in A} E_i, \quad \mathcal{E}_A := \bigotimes_{i \in A} \mathcal{E}_i, \quad \text{and} \quad E := \bigtimes_{i \in D} E_i, \quad \mathcal{E} := \bigotimes_{i \in D} \mathcal{E}_i$$

where $\bigtimes$ denotes the Cartesian product between sets and $\bigotimes$ denotes the product of σ-algebras (see, [150], Section 2.4.2). Notice additionally that, for any $A \subset D$, $(E_A, \mathcal{E}_A)$ is also a standard Borel measurable space and $(E, \mathcal{E})$ as well (see, e.g., [126], Lemma 1.2).

The *random inputs* are represented by an E-valued, $(\mathcal{F})$-measurable mapping $X = (X_1, \ldots, X_d)^\top$ (i.e., a vector of random elements). For any $A \subset D$, the E_A-valued vector of random elements $X_A := (X_i)_{i \in A}$ defines a *subset of inputs*.

The *σ-algebra generated by the random inputs* (see, Definition A.2) is denoted σ_X, and for any $A \subset D$, the σ-algebra generated by the subset of inputs X_A is denoted σ_A. These generated σ-algebras can be understood as the relevant theoretical notion to formally identify the *sources of uncertainties* and are traditionally interpreted as the *information* brought forward by a random element.

The *joint distribution of the random inputs* is the probability measure induced by the measurable mapping X (see, Definition A.5), denoted P_X. For every $A \subset D$, the *marginal distribution of the subset of inputs* X_A is the probability measure induced by the measurable mapping X_A, denoted P_{X_A}.

Black-box model Again, in the spirit of generality, to accommodate both numerical and ML-based black-box models, particularly their variety of outputs (e.g., meshes, text, regression, classification), black-box models are defined in a rather abstract manner.

Let $(Y, \mathcal{Y})$ be a standard Borel measurable space. The *black-box model* is represented by a measurable mapping $G : E \to Y$.

Random output Rather naturally, and in the spirit of the *propagation of uncertainties* in UQ, the random output refers to the composition of the random inputs and the black-box model. Considering all the uncertainties it is subject to, it can be interpreted as the representation of the system model as a whole.

The random output is denoted by the measurable function $G(X) := G \circ X : \Omega \to Y$, i.e., a Y-valued random element. It is important to note that random outputs are necessarily σ_X-measurable functions. Additionally, denote by $\mathcal{G}_X$ *the space of random output* defined as:

$$\mathcal{G}_X = \{f : \Omega \to Y : f \text{ is } \sigma_X\text{-measurable}\}.$$

Additionally, for any $A \subseteq D$, denote $\mathcal{G}_A$ the subset of $\mathcal{G}_X$ of Y-valued, σ_A-measurable functions (where $\sigma_D = \sigma_X$). Additionally, denote $\mathcal{G}_\emptyset$ the space of functions measurable w.r.t. the $\mathbb{P}$-*trivial σ-algebra* (see, Definition A.6), denoted $\sigma_\emptyset$.

Figure 1.3 illustrates the relationships between the random inputs, the black-box model, and the random output. The proposed framework highly emphasizes the *functional relationships* between these three notions. In the context of black-box model interpretability, it is essential to note that the black-box modeling step is considered as *given*. The main focus of this thesis is not on *how one can model a phenomenon using black-boxes*, but instead on *drawing insights from the black-boxes once they have been modeled*. However, formally describing the elements that compose this first modeling step is paramount in the following developments. This formal take on black-box modeling allows for a sufficiently general framework, encompassing the diversity and complexity of real-world situations.

1.2.2 Quantity of interest

QoIs are paramount in the framework of model interpretability. They must be *meaningful to domain experts*, by bearing key information towards the *conundrum* the interpretability study aims at providing an answer to. In the spirit of generality and to accommodate the broad range of possible insights related to black-box models, these QoIs are also considered to be random, even though, in most cases, they are considered deterministic. The proposed definition of QoI expands the homonymous notion in SA [48] to consider a broader range of situations.

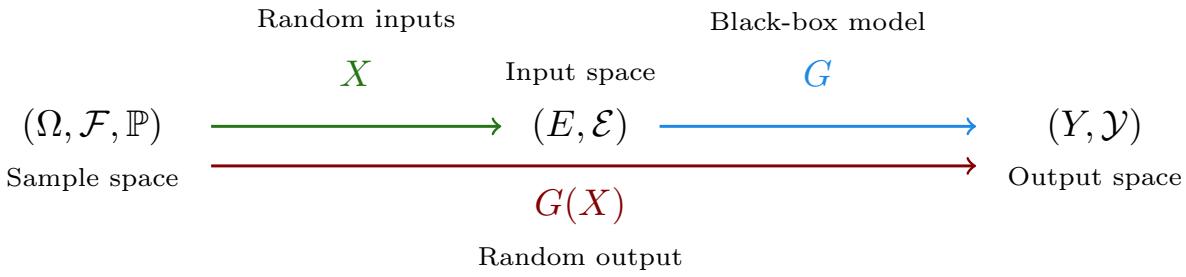

Figure 1.3: Black-box modeling: Relationship between the sample, input, and output spaces using mappings.

Let $(Q, \mathcal{Q})$ be a measurable space, called the *QoI space*. Let $\text{QoI} : \mathcal{G}_X \to Q$ be an operator, and the QoI refers to the random element result of its composition with the random output, i.e., $\text{QoI}\left(G\left(X\right)\right)$.

Examples of QoIs can be the random output itself, an *evaluation* (observation) of the random output, i.e., for an $\omega \in \Omega$ the quantity $G(X(\omega))$ [149], in case of $\mathbb{R}$-valued outputs, its *expectation*, i.e., $\mathbb{E}\left[G(X)\right]$, its *variance*, i.e., $\mathbb{V}\left(G(X)\right)$ [204], and in case of $\mathbb{R}^d$-valued outputs, its covariance matrix [85]. In the last two examples, the integral operator is taken w.r.t. the fixed probability measure $\mathbb{P}$ on Ω.

Special case (Special cases of the framework). In the remainder of this thesis, particular attention is paid to whether the presented theoretical results or empirical studies focus on special cases of this somewhat abstract framework. These restrictions (on the input, output, and QoI spaces) are encapsulated in these *special case* blocs and displayed before stating any theoretical or empirical result.

1.2.3 Conundrums and interpretability methods

In all generality, interpretability methods can be understood as *meaningful transformations of the QoI*. Although this definition is not very formal, it remains general enough to encompass the broad spectrum of methods proposed in the literature (see, e.g., [16]). A *transformation of the QoI* can be understood as performing a theoretically-grounded methodological study of the QoI. The term *meaningful* emphasizes the goal of the interpretability method, which is embodied by a *conundrum* [9]. Once a relevant interpretability method has been chosen to bring forward insights toward solving a conundrum, i.e., defining the meaningful theoretical quantities candidate to answering the practical question, computing estimates of these quantities lead to *diagnostics*, which can then be interpreted in order to explain studied domain-specific problem.

Conundrums Conundrums are embodied as *domain-oriented questions*. In the abstract modeling of *explanations games* [9], two individuals, an *explainer* (e.g., domain-experts, engineers) and an *explainee* (e.g., decision-maker, validation authority) interact in order to solve a *conundrum*, i.e., a question originating from the explainee which needs to be addressed by the explainer. To solve a particular conundrum, the explainer must provide an *explanation* to the explainee, which then decides if the explanation is *sufficient*. Studying the interactions between explainer and explainee is at the crossroads of many scientific research areas, such as logic, non-cooperative game theory, psychology, ergonomics, and microeconomics. They are leveraged in XAI to study the social aspects of *the acceptance of artificial intelligence* [16].

Examples of conundrums can be (but are not limited to):

- *"Why does the model provide this particular prediction on a particular datapoint?"*
- *"Which inputs are responsible for the uncertainty of the modeled system?"*
- *"What is the impact of the lack-of-knowledge about the inputs on the modeled system?"*

However, the work presented in this thesis focuses on one particular point of this complex interaction process: providing tools for the explainer to build relevant explanations for specific conundrums. Hence, these tools must provide insights whose meaning must be theoretically justified to support the subse-

quent explanations' relevancy. The tools are called *interpretability methods* in the remainder of this thesis and are discussed below.

Interpretability methods Once a conundrum has been stipulated, the first step is identifying relevant QoIs. They should be chosen as key indicators of the overall conundrum at play. Interpretability methods entail finding a methodology to solve the conundrum by studying the chosen QoIs. For example, if the main question revolves around the reasons behind a model prediction, a suitable QoI would be the prediction itself, with suitable methods revolving around causality (e.g., counter-factual methods [159, 9]), or rule-extraction approaches [18]. If the conundrum deals with identifying which inputs affect a QoI the most, one can refer to decomposition methods (e.g., coalitional decompositions [111], attribution methods [149]). If the explainee asks about the out-of-distribution behavior of the QoI, one can refer to input perturbation methods (e.g., probability measure projections [141, 13, 113], information-geometric-based perturbations [86]).

Many interpretability methods have been proposed in the XAI literature, offering insights into many different situations (see, e.g., [16, 143, 207] for an overview of the proposed methods). However, since the main goal of this thesis revolves around the modeling of systems, three desirability criteria are introduced:

- **Relevancy**: The power of an interpretability method to address a conundrum should be motivated;
- **Theoretical-groundedness**: An interpretability method should be built upon a strong theoretical framework with clearly stated assumptions, its properties studied, and shortcomings highlighted;
- **Practical coherence**: Aside from theory, the insights brought forward by the interpretability method must be studied empirically and validated on controlled use cases, and be in accordance with the domain-experts' opinions.

Producing relevant estimators of the theoretical quantities defined by interpretability methods also plays a pivotal role in bridging the gap between theory and practice. The estimates are referred to as *diagnostics* in the proposed framework. Despite their importance, apparent logic, and sometimes their demonstrated empirical usefulness [186], their adoption for decision-making for critical-system modeling remains subject to the theoretical study of the quantities they aim at approaching.

1.3 Two main interpretability methods

The work presented in this thesis introduces two main interpretability methods:

- **QoI decompositions**: The study of how QoIs can be decomposed w.r.t. the inputs and subsets of inputs of a black-box model. It is particularly suitable for conundrums dealing with influence quantification;
- **Input perturbations**: A methodology to perturb the distribution of the inputs in various settings. It proposes an answer to model robustness-related conundrums.

1.3.1 QoI decomposition for influence assessment

QoI decomposition methods, as their name suggests, entail being able to write $\mathrm{QoI}\left(G\left(X\right)\right)$ as a sum of elements of Q, provided it is endowed with a suitable "addition" operation (i.e., it is an Abelian group, see Appendix B). For instance *additive attributions methods* [149] consider the following sum:

$$\mathrm{QoI}\left(G\left(X\right)\right) = \phi_{\emptyset} + \sum_{i \in D} \phi_i,$$

where for every $i \in D$, $\phi_i \in Q$, and where each ϕ_i correspond to an *effect of the input* X_i. *Coalitional QoI decompositions* differ from attributions methods, in the sense that the sum is taken over the *power-set of* D (i.e., the set of subsets of D, including $\emptyset$), denoted $\mathcal{P}_D$. A coalitional decomposition of $\mathrm{QoI}\left(G\left(X\right)\right)$ would entail having:

$$\mathrm{QoI}\left(G\left(X\right)\right) = \sum_{A \in \mathcal{P}_D} \phi_A,$$

where for every $A \in \mathcal{P}_D$, $\phi_A \in Q$, and where each ϕ_A correspond to an *effect of the subset of inputs* X_A. The term "coalitional" comes from the fact that the effects of *coalitions* (i.e., subsets) of inputs are taken into account in the decomposition, in contrast to attribution methods which only focus on *individual* (i.e., univariate) effects. These two interpretability methods are intimately linked (see, Chapter 3).

The central paradigm behind QoI decompositions is that the resulting quantified effects may bear some information on an input's influence on the QoI. If Q is endowed with a natural total order, comparing the magnitudes of these effects can give insight into a potential "influence ranking" over the inputs.

In the literature, many techniques relying on these methods have been proposed. Examples of attributions methods proposed are, for evaluation decomposition of regression models (i.e., $Y = \mathbb{R}$, and $\mathrm{QoI}\left(G\left(X\right)\right)\left(G(X)\right) = G(X(\omega))$ for some $\omega \in \Omega$), LIME [186], or SHAP [149], and for variance decomposition (i.e., $Y = \mathbb{R}$ and $\mathrm{QoI}\left(G\left(X\right)\right) = \mathbb{V}\left(G(X)\right)$), Shapley effects [169] or proportional marginal effects [100]. The prime example of coalitional variance decomposition would be the well-known Sobol' indices [204], which are at the cornerstone of the field of variance-based SA.

1.3.2 Input perturbations for the assessment of model robustness

Input perturbation methods deal with modifying the inputs' distribution in a controlled manner. Once the modified distribution is achieved, input perturbation methods enable the study of the QoI *under the modified distribution*. In essence, for initial random inputs X and perturbed inputs $\widetilde{X}$, one can then study the differences between $\mathrm{QoI}\left(G\left(X\right)\right)$ and $\mathrm{QoI}(G(\widetilde{X}))$ caused by the particular perturbation. In-fine, it allows assessing the model's behavior (through its QoIs) on a different input probabilistic scheme than the initial one caused by a controlled perturbation, ultimately allowing *assessing the model's robustness to different input distributions*. This interpretability method can be used for prospective studies and exploratory analysis or to ensure the coherence of the model with domain-experts' knowledge to prevent domain misspecification.

Formally, let $\mathcal{C}$ be a *perturbation class*, i.e., a particular set of probability measures induced by E-valued random inputs, and $\mathcal{D}$ is a *discrepancy* between probability measures (i.e., not necessarily a distance). The *perturbation problem* can be written as the following constrained optimization problem:

$$\begin{aligned} P_{\widetilde{X}} \in \underset{P}{\operatorname{argmin}} \quad & \mathcal{D}\left(P_X, P\right) \\ \text{s.t.} \quad & P \in \mathcal{C}. \end{aligned}$$

Several choices of discrepancies and perturbations classes have been studied in the literature. Leveraging the pioneering work of [47] on entropic projections, the choice of the Kullback-Leibler (KL) divergence has been investigated by [141] in SA and by [13] in XAI, where $\mathcal{C}$ is defined through constraints on generalized moments. [86, 127] proposed to study parametric families of distribution, where the discrepancy is chosen utilizing the Fisher metric on the parameter space, leading to natural perturbations classes comprised of sequences of perturbed distributions along geodesics. In [113], the choice of the 2-Wasserstein distance is motivated, coupled with copula-preserving-quantile-constrained perturbations classes.

1.4 Illustrative use-cases

Throughout this thesis, the introduced methods and techniques are illustrated through use-cases. In this section, three main use-cases are presented, as well as relevant conundrums the presented methods aim to solve. Additional use-cases are presented in Appendix F, whose analyses are left as supplementary material.

Remark 1.1 (Data availability and reproducibility statements). All the presented datasets, data-generation processes, numerical models, model training, result computations and codes for the displayed figures are made available in the accompanying GitHub repository[a].

[a] https://github.com/milidris/phdThesis

1.4.1 Simplified hydrological model

Description of the use-case

This first use case is an example of a numerical model being easy to evaluate. It aims to represent a simplified model of the water level of a river. This model has been extensively used in the safety and reliability of industrial sites, where the occurrence of a flood can lead to dramatic human and ecological consequences. It consists of a substantial simplification of the one-dimensional Saint-Venant equation, with a uniform and constant flow rate, inspired from [120, 82]. The maximal annual water level from sea level is modeled as follows:

$$G(X) = Z_v + \left(\frac{Q}{BK_s \sqrt{\frac{Z_m - Z_v}{L}}} \right)^{3/5}, \tag{1.1}$$

where input variable and their explicit marginal probabilistic structure is detailed in Table 1.1.

Input	Unit	Distribution	Application Domain	Description
Q	$\mathrm{m}^3/\mathrm{sec}$	$\mathcal{G}(1013, 558)$ trunc.	$[500, 3000]$	River maximum annual water flow rate.
K_s		$\mathcal{N}(30, 7)$ trunc.	$[15, 55]$	Strickler riverbed roughness coefficient.
Z_v	m	$\mathcal{T}(49, 50, 51)$	$[49, 51]$	Downstream river level.
Z_m	m	$\mathcal{T}(54, 55, 56)$	$[54, 56]$	Upstream river level.
L	m	$\mathcal{T}(4990, 5000, 5010)$	$[4990, 5010]$	River length.
B	m	$\mathcal{T}(295, 300, 305)$	$[295, 305]$	River width.

Table 1.1: Inputs of the simplified river water level model and their explicit marginal distributions. $\mathcal{G}, \mathcal{N}, \mathcal{T}$ denote Gumbel, Normal and Triangular distributions, which may be truncated (trunc.).

For reliability studies, the modeled river can be considered to be located near an industrial site [140]. Hence, in addition to the random inputs, and as illustrated in Figure 1.4, a dyke surrounds the river, whose height is denoted by t. Hence, a reported maximal annual water level beyond this height characterizes the event of a flood of the industrial site.

Additionally, similarly to [38], a dependence structure between the inputs is modeled using a Gaussian copula, with the following correlation matrix:

$$R_P = \begin{pmatrix} 1 & 0.5 & 0 & 0 & 0 & 0 \\ 0.5 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0.3 & 0 & 0 \\ 0 & 0 & 0.3 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0.3 \\ 0 & 0 & 0 & 0 & 0.3 & 1 \end{pmatrix}, \quad \text{where} \quad \begin{pmatrix} Q \\ K_s \\ Z_v \\ Z_m \\ L \\ B \end{pmatrix} \sim P.$$

Conundrums

Concerning this use case, the following questions are answered in this thesis:

- *Which inputs are the most responsible for the uncertainty surrounding the river's maximal annual water level, and the event of a flood happening?* (see, Section 3.4.2)
- *Are there any inputs that do not contribute to this uncertainty?* (see, Section 3.4.2)
- *What would be the impact on the maximal annual water level of the river due to epistemic uncertainty on the Strickler riverbed roughness coefficient?* (see, Section 5.5.2)

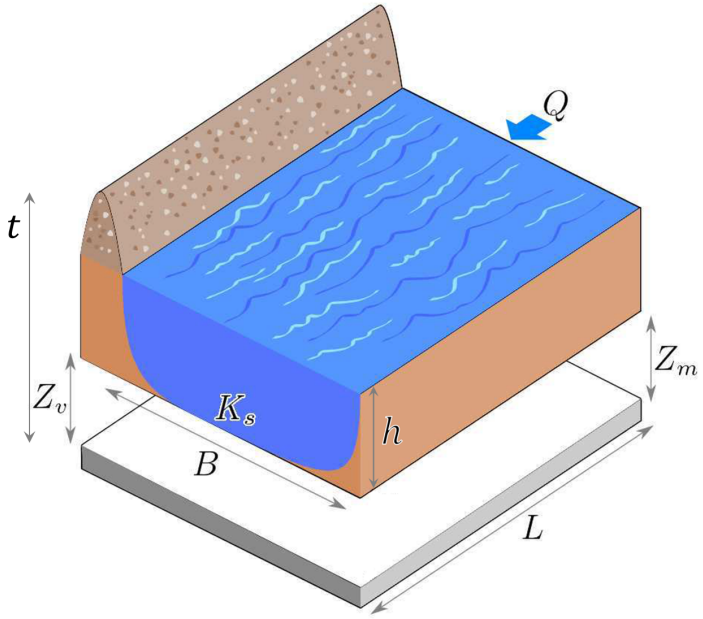

Figure 1.4: Illustration of the simplified hydrological numerical model.

- *If a surrogate replaced this model, would the meta-model capture these effects?* (see, Section 5.5.2)

1.4.2 Transmittance error of an optical filter

Description of the use-case

This second use case is an example of a complex numerical model that is too costly to evaluate, for which only a few simulations are available to the practitioner. In this use case, inspired by [217], the transmittance of an optical filter is studied. The studied system comprises 13 layers stacked on each other, each having the same thickness but varying refractive indices.

This filter aims at splitting a light wave into two or more parts, each taking different paths through the system before coming together. Due to the refraction of the wave on each successive layer of the system, the paths' length and amplitude can vary, resulting in varying system transmittance values. The ability to determine which layer is influential is crucial for optical filters and remains a complicated problem due to high levels of interaction between the layers. In the literature, previous global SA studies (see, e.g., [217, 216]) allowed providing some answers by assuming mutual independence between refractive indices.

Each of the 13 inputs $I_1, \ldots, I_{13}$ represents the refractive index error of a layer in the optical filter, which is assumed to vary uniformly between $[-0.05, 0.05]$. These errors are correlated, representing a deviation in the manufacturing process of the layers. The dependence structure is modeled using a Gaussian copula, where each pair of inputs exhibits a 0.9 correlation coefficient.

As depicted in [217], several light waves of varying frequencies are passed through the filter. The transmittance is then computed for each frequency, and their squared error w.r.t. the "perfect filter" (i.e., with no error) is computed. The model's output is the square root over the sum of these squared errors.

The practitioner only has access to a unique i.i.d. sample of size 1000 of the input-output simulations.

Conundrums

Concerning this use case, the following questions are answered in this thesis:

- *Which inputs are the most responsible for the uncertainty surrounding the optical's transmittance error?* (see, , Section 3.4.3)
- *Are there any inputs that do not contribute to this uncertainty?* (see, , Section 3.4.3)

- *If this model were to be replaced by a surrogate, would the importance ranking be suitable for feature selection?* (see, , Section 3.4.3)

1.4.3 Acoustic fire extinguisher dataset

Description of the use-case

The last illustrative use case represents a typical ML modeling procedure where only a dataset of experiments is available. The acoustic fire extinguisher dataset comprises 15390 experiments of fire extinguishing tests for three different liquid fire fuels. Amplified sub-woofers are placed in a collimator with an opening. When activated at different frequencies, the acoustic waves produce an air escape through the opening, which is used to extinguish fires. Three features are set using a design of experiment (DoE), and two are measured using appropriate equipment. One can refer to the in-depth descriptions in [128, 211] for more details on the experiment's settings. Table 1.2 gives additional details on the nature of the features.

Feature	Unit	Mode of measurement	Description
TankSize	cm	DoE	Discrete feature (5 levels) describing the tank size containing the fuel.
Fuel		DoE	Fuel type used (3 levels: Gasoline, Kerosene, Thinner).
Distance	cm	DoE	Distance of the flame to the collimator opening.
Frequency	Hz	DoE	Sound frequency range.
Decibel	dB	Measured	Sound pressure level.
Airflow	m/s	Measured	Airflow created by the sound waves.

Table 1.2: Description of the features of the acoustic fire extinguisher dataset.

For each experiment, a *binary output variable* is measured, representing the result of the experiment, i.e., whether the fire has been put out (1) or not (0). The two output classes are relatively balanced: 48.97% of the observations describe effectively extinguished fires. The empirical distribution, correlation structure, and relationship between the features and the output are represented in Figure 1.5. Some variables seem somewhat correlated in Spearman's sense [164], i.e., the linear correlation of the rank-transformed data. In addition, the correlation ratios [44] between the discrete inputs (TankSize and Fuel) and the continuous inputs are negligible.

The classification black-box model is a one-layer neural network (composed of 100 neurons), trained on 500 epochs, with a learning rate of 10^{-4}, similar to the study conducted in [212]. 5% of the data has been randomly selected for validation. The model resulted in a good prediction accuracy: 95.15% of the training data and 94.26% of the validation data are correctly classified. Figure 1.6 depicts the trained black-box model's ROC curve and confusion matrix. The model's predictive performance can be validated globally with an AUC of 0.992 and less than 3% of type 1 and 2 prediction errors.

Conundrums

Concerning this use case, the following questions are answered in this thesis:

- *Which inputs are the most responsible for the uncertainty surrounding the effective termination of a fire?* (see, , Section 3.4.4)
- *Are there any inputs with negligible effect on the uncertainty surrounding the effective termination of a fire* (see, , Section 3.4.4)

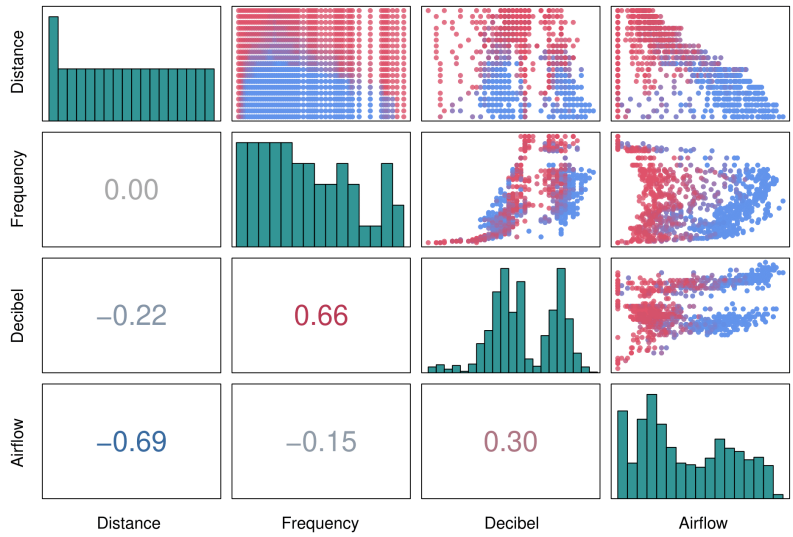

Figure 1.5: Histogram, cross-scatterplot, and Spearman's correlation coefficient of the input features of the acoustic fire extinguisher dataset. Red dots represent observations resulting in $Y = 0$, and blue dots for $Y = 1$.

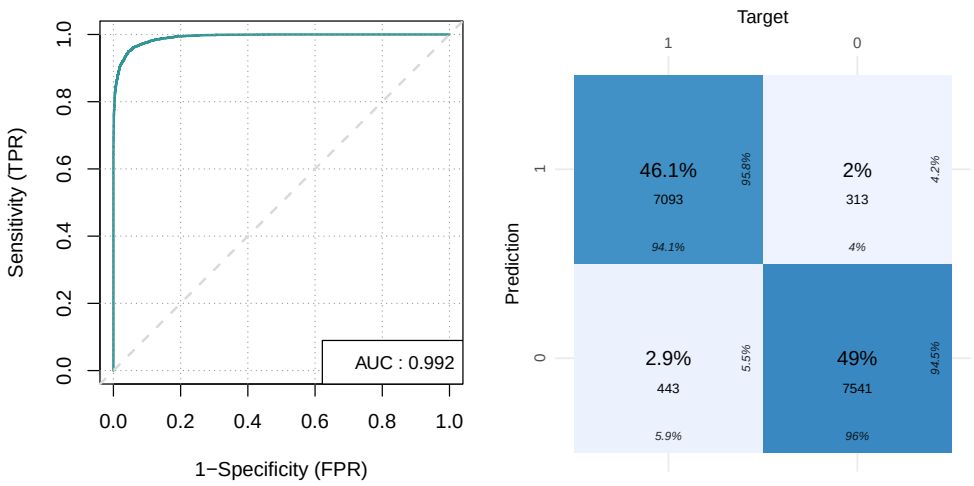

Figure 1.6: ROC curve (left) and confusion matrix (right) of the neural network model trained on the acoustic fire extinguisher dataset.

- *Would the model's prediction be robust to perturbations on the quantiles of the Airflow feature?* (see, , Section 5.5.1)
- *What would be the effects on the importance ranking of perturbations on the quantiles of the Airflow feature?* (see, , Section 5.5.1)

1.5 Intention and content of the manuscript

The contents of this thesis can be seen as a theoretical deep-dive into the two previously introduced interpretability methods: QoI decompositions and input perturbations. The main concern is formalizing these methods and going beyond empirically justified techniques to offer suitable tools for validating the use of black-boxes to model critical systems. Producing efficient estimates and diagnostics is essential, but it comes in second in the context of the presented work. Already established estimation schemes are presented for a plethora of practical situations. However, no novel estimators are proposed.

Additionally, as eluded in Section 1.2, generality is strongly emphasized. For instance, many considerations are put on the case of *dependent inputs* i.e., the variations of the inputs of a black-box model are influenced by each other, with mutual independence seen as a particular case. This general point of view is taken to offer suitable mathematical foundations to unify SA and post-hoc interpretability. Many methodological questions are addressed in the contents of this thesis, such as, for instance:

- How is it possible to generically decompose QoIs? (see, Chapter 2)
- What do existing interpretability techniques quantify? (see, Chapter 3)
- Is there a natural way to define importance when the inputs are not mutually independent?(see, Chapter 4)
- How can one define pure interaction effects in contrast to effects due to stochastic dependence between the inputs? (see, Chapter 4)
- What constitutes meaningful input perturbation, and how can the insights be interpreted? (see, Chapter 5)

The contents of this manuscript stand on several fields of mathematics: probability theory, statistics, abstract algebra, combinatorics, cooperative game theory, and functional analysis, to name a few. Some preliminaries are proposed in relevant appendices to make this thesis as self-sufficient as possible. Theoretical results from the literature are accompanied by references containing their proofs. The proofs of original results developed during the Ph.D. are postponed to the appendices.

Chapter 2 introduces the QoI decomposition interpretability method. Coalitional QoI decompositions are first introduced, and their link with the field of algebraic combinatorics through Rota's generalization of the Möbius inversion formula. It leads to two main ways to conceptually perform QoI decompositions: the input-centric and the model-centric approaches.

Chapter 3 introduces techniques coming from cooperative game theory. These attribution methods are shown to be aggregations of coalitional decompositions stemming from the input-centric approach. For the output variance decomposition task, the Shapley effects are introduced as an egalitarian redistribution of suitable dividends. Another proposed technique, called the proportional marginal effects, relies on a proportional redistribution of dividends. Although these methods are relevant in practice, they present methodological issues, which can be traced back to the input-centric approach.

Chapter 4 focuses on the model-centric approach. It amounts to performing functional decompositions, reminiscent of Hoeffding's functional analysis of variance (FANOVA). This result, which assumes mutual independence of the inputs, is shown to be a particular case of a more general result, allowing the inputs to be dependent. In particular, the variance decomposition leads to the definition of novel indices, allowing the separation of interaction and dependence effects.

Chapter 5 is concerned with the problem of input perturbation. The general framework of this particular interpretability method is introduced and discussed. The development of a particular instance of this general problem is then studied: the choice of the 2-Wasserstein distance as a suitable discrepancy between probability measures, along with quantile-based interpretable perturbations.

Chapter 2

Decompositions of quantities of interest

Contents

2.1 Input influence assessment for interpretability 19
2.2 Coalitional decompositions and influence measures 19
2.2.1 Influence order and influence measures 19
2.2.2 Coalitional decompositions 20
2.2.3 Möbius inversion on the Boolean lattice 21
2.2.4 Two approaches to measure influence 23
2.3 Importance quantification and the variance as a QoI 24
2.3.1 Mutually independent inputs and the model-centric approach 25
2.3.2 Dependent inputs and the input-centric approach 26
2.3.3 Illustration on controlled examples 27
2.4 Partial conclusion 28

Abstract (English). The very concept of measuring the influence of a group of inputs on a quantity of interest in a black-box model is based on algebraic considerations. More precisely, the search for an *influence measure* associating a value (real or more abstract) with a group of inputs is justified by the mere existence of a total order that would allow them to be ranked. *Coalitional decompositions* make it possible to associate with each subset of inputs a share of the quantity of interest. These particular influence measures can be linked to the notion of *generalized Möbius inversion*, a well-known result in combinatorics. This connection describes two ways to build coalitional decompositions: an approach focused on inputs and an approach based on an intrinsic decomposition of the model. These two approaches are illustrated in the case of decomposing the variance of a black-box model, leading to the study of Sobol' indices. Both approaches are then illustrated through analytical results on simple toy cases.

Abstract (Français). Le concept même de mesurer l'influence d'un groupe d'entrées sur une quantité d'intérêt d'un modèle boîte-noire repose sur des considérations algébriques. Plus précisément, la recherche de *mesure d'influence* associant une valeur (réelle ou plus abstraite) à un groupe d'entrées, est justifiée par la simple existence d'un ordre total qui permettrait de les hiérarchiser. Les *décompositions coalitionnelles* permettent d'associer à chaque sous-ensemble des entrées une part de la quantité d'intérêt. Ces mesures d'influence particulières peuvent être rapprochées de la notion d'*inversion de Möbius généralisée*, résultat phare du domaine de la combinatoire. Ce rapprochement permet de décrire deux manières de construire des décompositions coalitionnelles : une approche focalisée sur les entrées, et une approche reposant sur une décomposition intrinsèque du modèle. Ces deux approches sont illustrées dans le cas de la décomposition de la variance d'un modèle boîte-noire, amenant à l'étude des indices de Sobol'. Ces deux approches sont ensuite illustrées par le biais de résultats analytiques sur des cas jouets simples.

Keywords . Influence order • Influence measure • Möbius inversion formula • Coalitional decompositions • Sobol' indices

This chapter expands on the following contributions.

Journal articles:
M. Il Idrissi, N. Bousquet, F. Gamboa, B. Iooss, and J. M. Loubes. On the coalitional decomposition of parameters of interest. *Comptes Rendus. Mathématique*, 361:1653–1662, 2023. DOI: 10.5802/crmath.521

Conferences:
M. Il Idrissi, N. Bousquet, F. Gamboa, B. Iooss, and J. M. Loubes. Cooperative game theory and importance quantification. In *23rd European Young Statisticians Meeting of the Bernoulli Society*, Ljubljana, Slovenia, 2023. URL: https://sites.google.com/view/eysm2023

M. Il Idrissi, N. Bousquet, F. Gamboa, B. Iooss, and J. M. Loubes. Coalitional decomposition of quantities of interest. In *2023 Annual Meeting of MASCOT-NUM Research Group*, Le Croisic, France, 2023. URL: https://mascotnum2023.sciencesconf.org/

M. Il Idrissi, V. Chabridon, and B. Iooss. Shapley effects for reliability-oriented sensitivity analysis with correlated inputs. In *Proceedings of the 10th International Conference on Sensitivity Analysis of Model Output*, Tallahassee, Florida, United States of America, 2022. URL: https://samo2022.math.fsu.edu/

2.1 Input influence assessment for interpretability

A rather natural and intuitive question regarding a black-box model is whether its inputs have varying levels of "influence" on some given QoI. Take, for instance, the variance of the random output, which is usually interpreted as *the amount of uncertainty the modeled system is subject to*. Tracing these uncertainties back to subsets of inputs would allow comparing their *importance* w.r.t. to the model. For instance, in engineering studies, it could result in a preventive allocation of resources to the study of these particular sets of important inputs or allocating fewer resources to the fine measurements of less important inputs. Another example would be the fairness audit of black-box AI models [21]: if, for a particular evaluation of the model, a protected (i.e., sensitive) feature is shown to bear some influence, then the decision could not be considered as fair. Hence, the notion of influence is necessarily related to the QoI of a black-box model: an influential set of inputs, say, for the variance of the model, may not be influential on one of its quantiles. Measuring and assessing influence can be relevant for solving conundrums such as: *"What is the influence of subsets of inputs on the QoI?"* or *"Are there any inputs that do not influence the QoI?"*.

The question of measuring influence is central to many fields, such as sensitivity analysis [48] in order to interpret black-box models and to detect inputs with negligible effects on the system's overall uncertainty, but also in statistical learning [97], where importance quantification can be at the basis of feature selection or model comprehension [70, 92, 44]. More recently, in the XAI literature, many "explanation methods" have been proposed to measure the influence of inputs on the evaluation of learned models (e.g., SHAP [149], LIME [186]). These influence measures are usually expected to express some "influence ranking" between sets of inputs, making them comparable for decision-making purposes. However, aside from reasonable desirability criteria or rationales justifying their conception, these methods lack a general framework for their theoretical study.

In this chapter, the algebraic roots of this problem are highlighted. The two notions of *influence order* and *influence measure* of subsets of inputs are disentangled and formally introduced. The definition of the latter comes naturally as order embeddings. It motivates the definition of *coalitional decompositions*. These decompositions require desirability criteria to ensure the influence measures are relevant and suitable and quantify understood theoretical quantities. Coalitional decompositions are intimately related to *Rota's generalization of the Möbius inversion formula*, a theoretical result from the field of combinatorics. This result offers two approaches to define influence measures: input-centric and model-centric. These approaches are illustrated for variance decomposition, using the well-known Sobol' indices [204] as examples. These indices are analytically computed for two simple use cases, where mutual independence is not necessarily assumed and then interpreted.

2.2 Coalitional decompositions and influence measures

In this section, the general framework of influence quantification is explored. The notion of *influence measures*, i.e., functions aiming at quantifying the influence of subsets of inputs on a QoI, comes naturally as soon as the existence of an *influence order* is assumed. This abstract ranking can express that some inputs can have varying degrees of influence on a QoI and thus be compared using a binary relation. Measuring the influence on a QoI amounts to finding an order embedding onto the QoI space and can be done through coalitional decompositions, i.e., additive decompositions of QoIs. Coalitional decompositions are intimately linked with the notion of *Möbius transforms*, which leads to two different methodological approaches to characterizing influence measures.

2.2.1 Influence order and influence measures

First, the focus is put on the rationale behind the construction of influence measures and their overall goal. It relies heavily on notions of abstract algebra and order theory. The interested reader can find some preliminaries in [52, 197] and in Appendix B.1.

The idea behind "influence" is deeply tied to the existence of an *ordering* between the coalitions of inputs. This ordering would express a ranking w.r.t. to the influence of subsets of inputs on a particular QoI. In other words, it assumes the existence of some abstract (and unobserved) binary relation between subsets of inputs, forming a *total order*.

Formally, let $\preceq_{\mathrm{QoI}}$ be a binary relation acting on $\mathcal{P}_D$, and suppose that $(\mathcal{P}_D, \preceq_{\mathrm{QoI}})$ forms a *totally ordered*

set (see, Definition B.5). For two subsets of inputs X_A and X_B, $A, B \in \mathcal{P}_D$, $A \preceq_{\text{QoI}} B$ means that "X_B has the same, or more influence on QoI $(G(X))$ than X_A". However, the binary relation $\preceq_{\text{QoI}}$ can never be observed nor inferred directly. In the following, only its existence is assumed.

Despite these drawbacks, it remains possible to gather insights on this binary relation using *influence measures*, i.e., comparable elements on Q, the QoI space (assuming Q can be endowed with a total order). The formal definition of influence measures is motivated by the following result.

Proposition 2.1. *Let $(T, \preceq)$ be a finite totally ordered set, and let $(M, \leq)$ be an infinite totally ordered set. There always exists a function $\phi : T \to M$, such that, $\forall A, B \in T$,*

$$A \preceq B \quad \iff \quad \phi(A) \leq \phi(B).$$

Proof: Finite totally ordered sets can always be embedded in a chain [52] of an infinite totally ordered set.

Special case . From this point on, Q is assumed to be an infinite Polish space (i.e., the topological space is not finite), and $(Q, \mathcal{Q})$ is assumed to be standard Borel.

Actually, Proposition 2.1 can be understood as the fact that there will always exist an *order-embedding* (see, Definition B.7) between $\mathcal{P}_D$ and the QoI space Q, preserving the total influence order, under the sole assumption that the inputs can be ranked w.r.t. their influence. These order-embeddings are called *influence measures*. Finding relevant influence measures for subsets of inputs can be seen as searching for suitable functions $\phi : \mathcal{P}_D \to Q$, hoping it expresses the influence order. Since the influence order can neither be observed nor be inferred directly, proposed influence measures in the literature are often justified by desirability criteria, i.e., intuitive properties of the influence measure that are deemed reasonable.

2.2.2 Coalitional decompositions

The sole assumption of the existence of an influence order over the subsets of inputs justifies the search for suitable influence measures, i.e., a function $\phi : \mathcal{P}_D \to Q$. However, any arbitrary influence measure may not be suitable to express the influence order. Without any additional assumption on $\preceq_{\text{QoI}}$, the relevancy of influence measures must be motivated.

Coalitional decompositions [111] are a particular class of influence measures ϕ that are inherently related to a QoI. In essence, they define influence measures that are *additive decomposition of the QoI* over $\mathcal{P}_D$, where each evaluation of influence measure related to a subset of inputs aims at quantifying its influence. However, additively decomposing a QoI requires additional assumptions on Q (i.e., addition over the elements of Q must be properly defined).

Special case . From this point on, Q is assumed to be an Abelian group (see, Definition B.2) with the addition operation "$+$".

Remark 2.1. The assumption that Q forms an Abelian group when endowed with an addition operation is not too restrictive. Since $\mathbb{R}$ (with the usual addition) is, in particular, an Abelian group, $\mathbb{R}$-valued QoIs remain valid. Furthermore, spaces of real or complex matrices (with the usual elementwise addition) are also Abelian groups, along with vector spaces (with the usual vector addition). Thus, a vast range of QoIs can be taken into account. In fact, it generalizes many of the developments from the literature to more abstract QoIs [111].

Formally, coalitional decompositions are defined as follows.

Definition 2.1 (Coalitional decompositions). Let X be random inputs, G be a black-box model, $G(X)$ a random output, and QoI $(G(X))$ be a Q-valued QoI. Let $\phi : \mathcal{P}_D \to Q$ be an influence measure. If the additive decomposition

$$\text{QoI}\,(G(X)) = \sum_{A \in \mathcal{P}_D} \phi(A)$$

hold, then ϕ is said to be a *coalitional QoI decomposition*.

While being fairly general, for an influence measure to be a coalitional decomposition, as defined in Definition 2.1, is not enough. Take, for instance, the influence measure

$$\phi(A) = \begin{cases} \mathrm{QoI}\left(G\left(X\right)\right) & \text{if } A = D; \\ 0 & \text{otherwise.} \end{cases}$$

In this case, ϕ is indeed a coalitional decomposition of $\mathrm{QoI}\left(G\left(X\right)\right)$, but as one can notice, its ability to express the influence order of the subsets of inputs can be questioned. Other than its evaluation on D, the evaluation of ϕ on another set $A \subset D$ is not directly linked to the subset of inputs X_A and, thus, fails at measuring its influence. Hence, additional restrictions on the influence measure are required. A desirability criterion called *graduality* is introduced to rule out trivial influence measures.

Gradual coalitional decompositions. Intuitively, one would want the evaluation $\phi(A)$, for $A \in \mathcal{P}_D$ of an influence measure, to be a *quantity related to the subset of inputs* X_A. In order to formally define graduality, the notion of *representants* of a subset of inputs is introduced. It is closely related to the *high-dimensional model representation* of Rabitz [179].

Definition 2.2 (Representant of a subset of inputs). Let X be random inputs, G be a black-box model, $G(X)$ a random output. Suppose that $G(X)$ can be written as

$$G(X) = \sum_{A \in \mathcal{P}_D} G_A(X_A) \text{ a.s,}$$

where $G_A(X_A) \in \mathcal{G}_A$.
In this case, $G_A(X_A)$ is said to be the *(G-)representant* of X_A. Additionally, if either

- $\forall B \subset A$, σ_B is strictly included in the σ-algebra generated by $G_A(X_A)$;
- The σ-algebra generated by $G_A(X_A)$ is included in $\sigma_\emptyset$;

then $G_A(X_A)$ is said to be *the proper representant of* X_A.

Remark 2.2. Thanks to the Doob-Dynkin lemma (see, Lemma A.2), representants are Y-valued random elements that are functions of X_A. *Proper representants* is the subset of $\mathcal{G}_A$ of Y-valued functions of X_A that *cannot solely be expressed as functions of the proper subsets of* X_A (i.e., that are in $\mathcal{G}_A \setminus \left(\bigcup_{B \subset A} \mathcal{G}_B\right)$). Hence, in essence, a representant of X_A can be understood as a Y-valued random element which is either *exactly* a function of the inputs in X_A, or constant a.s.

Influence measures that can be expressed as the QoI of representants are called *gradual* [111]. Hence, defining suitable gradual influence measures entails finding suitable representants $G_A(X_A) \in \mathcal{G}_A$ for each subset of input X_A, such that the sum of the QoIs of each of these representants is equal to $\mathrm{QoI}\left(G\left(X\right)\right)$. Formally, the graduality of an influence measure is defined as follows.

Definition 2.3 (Gradual coalitional decomposition). Let X be random inputs, G be a black-box model, $G(X)$ a random output, and $\mathrm{QoI}\left(G\left(X\right)\right)$ be a Q-valued QoI, and $\phi : \mathcal{P}_D \to Q$ be a coalitional decomposition. If ϕ can be written, $\forall A \in \mathcal{P}_D$, as

$$\phi(A) = \mathrm{QoI}\left(G_A(X_A)\right),$$

where each $G_A(X_A) \in \mathcal{G}_A$ is a representant of X_A, then ϕ is said to be *gradual*. If, in addition, $G_A(X_A)$ is a proper representant of X_A, then ϕ is said to be *properly gradual*.

Gradual influence measures are suitable candidates for expressing the influence order since, by design, they are the inherent expression of the QoI on functions of subsets of inputs.

2.2.3 Möbius inversion on the Boolean lattice

As detailed in the previous sections, defining influence measures entails finding suitable functions $\phi : \mathcal{P}_D \to Q$. To that extent, one can first note that $\mathcal{P}_D$ admits a very particular *algebraic structure*.

When endowed with the usual inclusion (i.e., $\subseteq$) binary relation, $(\mathcal{P}_D, \subseteq)$ forms a very particular *partially ordered set* (poset). In order theory, posets of power-sets characterize *Boolean lattices* [52], which can be illustrated using *Hasse diagrams*, as in Figure 2.1.

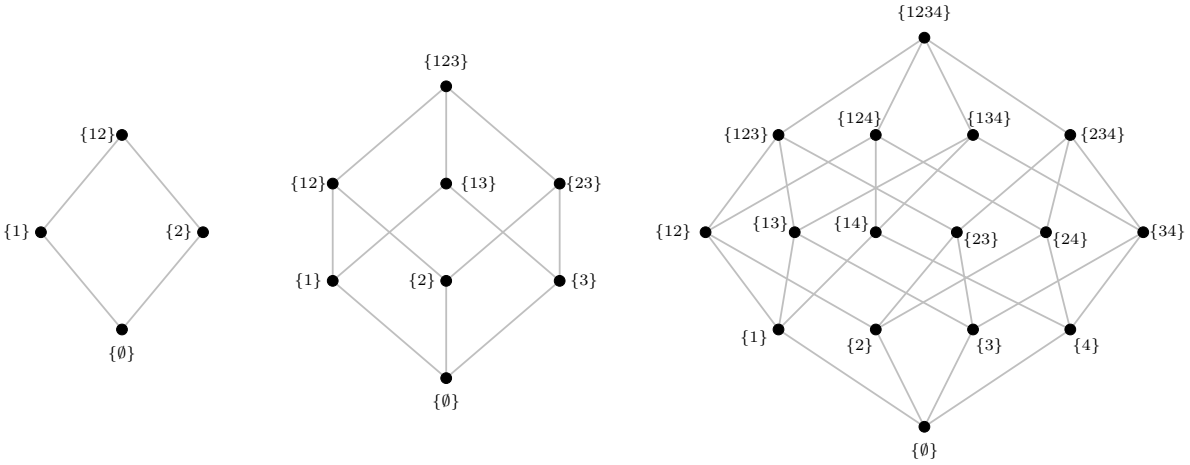

Figure 2.1: Hasse diagram of the Boolean lattice formed by the power-set of D for d being equal to 2, 3, and 4 (from left to right). It should be read from the bottom to the top. If two vertices are linked, the bottom element "is included" (in the sense of $\subseteq$) in the above element.

The *classical Möbius inversion formula* has been first discovered in the field of number theory [155]. It provides a particular relation between pairs of arithmetic functions (i.e., defined on the natural numbers and valued in $\mathbb{R}$) by leveraging the order structure of the natural numbers w.r.t. the binary relation of divisibility. This result has been generalized to functions defined on (locally) finite posets by Gian-Carlo Rota [187]. The interested reader is referred to Appendix B.1.3 for further details about Rota's generalization.

In particular, when dealing with functions defined on the power-set, leveraging that it forms a Boolean lattice, Rota's extension of the Möbius inversion formula entails the following.

Corollary 2.1 (Generalized Möbius inversion on the power-set). *Let d be a finite positive integer, $\mathcal{P}_D$ be the power-set of D, and let $\mathbb{A}$ be an Abelian group. For any two set functions*

$$v : \mathcal{P}_D \to \mathbb{A}, \quad \text{and} \quad \phi : \mathcal{P}_D \to \mathbb{A},$$

the two following statements are equivalent:

(i) $\forall A \in \mathcal{P}_D, \quad v(A) = \sum_{B \in \mathcal{P}_A} \phi(B)$;

(ii) $\forall A \in \mathcal{P}_D, \quad \phi(A) = \sum_{B \in \mathcal{P}_A} (-1)^{|A|-|B|} v(B)$.

Proof: see, [133] p.108 or [136] Lemma A.2.

Particular cases of Corollary 2.1 are widely used in many fields, sometimes without acknowledging its deep algebraic roots. For instance, the *inclusion-exclusion principle* for probability measures is an expression of this result (see, [74] Proposition 2.2). In Dempster-Shaffer (evidence) theory [199], this formula is used to link belief and mass functions, which share links with decision theory and the study of capacities [39]. More importantly, for the developments proposed in this thesis, this result is directly linked to the field of *cooperative game theory*, and in particular, to a broad set of *allocations* known as the *Harsanyi set* [26].

Methodologically, the equivalence in Corollary 2.1 offers two ways to approach influence measures. The first one can be understood as leveraging (ii) in order to obtain (i): this is called the *input-centric approach*. It requires defining a particular influence measure v, called a *value measure*, and studying the influence measure ϕ defined as its *Möbius transform* (see, Definition B.10). The second approach requires

an *intrinsic decomposition of* $G(X)$ in order to leverage (i) to obtain (ii) in Corollary 2.1: this is the *model-centric approach*. Provided ϕ is a coalitional decomposition, one can define a value measure v and express ϕ as its Möbius transform. Both of these approaches are further detailed in the following.

2.2.4 Two approaches to measure influence

There are two ways to define Möbius decompositions based on the equivalence relation of Corollary 2.1.

The input-centric approach Simply said, the input-centric approach focuses first on choosing an influence measure $v : \mathcal{P}_D \to Q$, referred to as the *value measure* in the following. Once a value measure is chosen, one can define an influence measure $\phi : \mathcal{P}_D \to Q$ as the Möbius transform of v. The value measure must respect a straightforward condition for ϕ to be a Möbius decomposition, as the following result highlights.

Proposition 2.2. *Let $v : \mathcal{P}_D \to Q$ be an value measure, and define*

$$\forall A \in \mathcal{P}_D, \quad \phi(A) = \sum_{B \in \mathcal{P}_A} (-1)^{|A|-|B|} v(B).$$

ϕ is a coalitional decomposition if and only if $v(D) = \text{QoI}\,(G\,(X))$.

Proof of Proposition 2.2 on p.120.

The input-centric approach to defining Möbius decomposition follows the rationale:

1. Let $\text{QoI}\,(G\,(X))$ be any QoI defined on a Abelian group Q;
2. Chose a value measure $v : \mathcal{P}_D \to Q$ such that $v(D) = \text{QoI}\,(G\,(X))$;
3. Define $\phi : \mathcal{P}_D \to Q$, such that $\forall A \in \mathcal{P}_D,\ \phi(A) = \sum_{B \in \mathcal{P}_A} (-1)^{|A|-|B|} v(B)$;
4. ϕ is an input-centric Möbius decomposition of $\text{QoI}\,(G\,(X))$.

Remark 2.3 (Input-centric mechanism). Proposition 2.2 and the construction of input-centric Möbius decompositions rely on the particular combinatorial mechanism of Corollary 2.1. One can notice that, in general, $\forall A \in \mathcal{P}_D$,

$$\sum_{B \in \mathcal{P}_A} \sum_{C \in \mathcal{P}_B} (-1)^{|B|-|C|} v(C) = v(A).$$

Hence, an influence measure ϕ, defined as

$$\phi(A) = \sum_{C \in \mathcal{P}_B} (-1)^{|B|-|C|} v(C),$$

will necessarily define a decomposition of $v(A)$ for every $A \in \mathcal{P}_D$, and in particular of $v(D)$.

Input-centric Möbius decompositions are at the heart of interpretability methods inspired from *cooperative game theory*. This intrinsic link is further explained, explored, and discussed in Chapter 3. The interested reader is referred to [111] for additional examples of input-centric decompositions for various QoIs.

The model-centric approach The model-centric approach can be understood as an *intrinsic QoI decomposition*, without a prior definition of a value measure v. It can rely, for instance, on the existence of a coalitional decomposition of $G(X)$. The main difference with the input-centric approach is that it does not rely on a prior choice of value measure but instead focuses on the intrinsic properties of the inputs X and the model G.

As its name suggests, the model-centric approach starts with a coalitional decomposition of $G(X)$, i.e., assuming that the decomposition

$$G(X) = \sum_{A \in \mathcal{P}_D} G_A(X_A) \ \text{ a.s,}$$

hold, and where each $G_A(X_A)$ is a representant of X_A. From this model decomposition, define the influence measure:

$$\begin{aligned}\phi : \mathcal{P}_D &\to Q \\ A &\mapsto \text{QoI}\left(G_A(X_A)\right),\end{aligned}$$

and the subsequent value measure:

$$\begin{aligned}v : \mathcal{P}_D &\to Q \\ A &\mapsto \sum_{B\in\mathcal{P}_A} \phi(B).\end{aligned}$$

Model-centric coalitional decompositions are intimately linked with the search for *properly gradual decompositions*. This approach is at the center of Chapter 4, where Hoeffding's decomposition of the Lebesgue space $\mathbb{L}^2\left(\sigma_X\right)$ is generalized for dependent inputs, leading to the definition of properly gradual decompositions of square-integrable real-valued models.

2.3 Importance quantification and the variance as a QoI

This section deals with *importance quantification* as a particular case of the above framework. The distinction is made between influence measures, which up until now have been broadly defined as the decomposition of QoIs in some abstract space Q, to the notion of *importance measures*, which refer to the special case of real-valued models, and their variance as a QoI. As eluded previously, in SA, the variance of the random output, i.e., $\mathbb{V}\left(G(X)\right)$ can be interpreted as "the overall amount of uncertainties of the modeled phenomenon" [57]. Hence, if a subset of inputs "is responsible" for a significant part of these uncertainties, it is deemed *important*. However, this interpretation of the variance is not exclusive to UQ. In statistics, a parallel is often made between the variance and notions such as *information* or *dispersion*, in the sense that a random variable with zero variance becomes deterministic (i.e., constant) and hence does not motivate probabilistic studies.

When studying the variance of a random variable, one first needs to make sure that it exists. To that extent, Lebesgue spaces $\mathbb{L}^2$ are introduced.

Definition 2.4 (Lebesgue space $\mathbb{L}^2$). Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let $\mathcal{B} \subseteq \mathcal{F}$ be a sub-σ-algebra. Let

$$\mathcal{L}^2(\mathcal{B}) := \left\{ Z \ : \ \Omega \to \mathbb{R} : \int_\Omega Z(\omega)^2 d\mathbb{P}(\omega) < \infty \ \text{ and } \ \sigma_Z \subseteq \mathcal{B} \right\},$$

be the space of square-integrable, $\mathcal{B}$-measurable random variables. Define the subspace of $\mathcal{L}^2(\mathcal{B})$

$$\mathbf{N} := \left\{ Z \in \mathcal{L}^2(\mathcal{B}) \ : \ \int_\Omega Z(\omega)^2 d\mathbb{P}(\omega) = 0 \right\}$$

The Lebesgue space $\mathbb{L}^2\left(\mathcal{B}\right)$ is defined as the canonical quotient space $\mathcal{L}^2(\mathcal{B})/\mathbf{N}$, i.e., relations (e.g., equalities, inequalities) between any two elements of $\mathbb{L}^2\left(\mathcal{B}\right)$ hold *($\mathbb{P}$-)almost surely (a.s.)*.

In addition, denote $\mathbb{E}_A\left[G(X)\right] := \mathbb{E}\left[G(X) \mid \sigma_A\right]$ the *conditional expectation of* $G(X)$ *given* X_A (see, e.g., [126]), and let $\mathbb{E}_D\left[G(X)\right] = \mathbb{E}\left[G(X) \mid \sigma_X\right] = G(X)$.

In this section, the general framework presented in Section 1.2 is restricted according to the following assumptions.

Special case . In this section, the following is assumed:

- The output space $Y = \mathbb{R}$, i.e., the model is $\mathbb{R}$-valued;
- The space of random outputs $\mathcal{G}_X$ is restricted to $\mathbb{L}^2\left(\sigma_X\right)$;
- The QoI is $\mathbb{V}\left(G(X)\right)$, and hence the QoI space $Q = \mathbb{R}$;

2.3.1 Mutually independent inputs and the model-centric approach

Random inputs $X = (X_1, \dots X_d)^\top$ are said to be *mutually independent* whenever their induced probability measure can be written, for every $B = (B_1, \dots, B_d) \in \mathcal{E}$, as

$$P_X(B) = \prod_{i=1}^{d} P_{X_i}(B_i),$$

i.e., the product of the probability measures induced by each input. Under this assumption, a particular decomposition of a random output $G(X) \in \mathbb{L}^2(\sigma_X)$ hold, known in the literature as *Hoeffding's decomposition* [102].

Theorem 2.1 (Hoeffding's decomposition). *Let X be mutually independent random inputs and $G(X) \in \mathbb{L}^2(\sigma_X)$ be a black-box model. There exists a unique decomposition of the form*

$$G(X) = \sum_{A \in \mathcal{P}_D} G_A(X_A), \quad \text{a.s.}$$

such that:

- *$G_\emptyset$ is constant a.s.;*
- *$\forall A \subseteq D$, $\forall i \in A$, $\int_{E_i} G_A(X_{A\setminus\{i\}}, x_i) dP_{X_i}(x_i) = 0$ (Annihilating property).*

Furthermore, one has that

$$\forall A \in \mathcal{P}_D, \quad \mathbb{E}_A[G(X)] = \sum_{B \in \mathcal{P}_A} G_B(X_B), \tag{2.1}$$

and for every $A \neq B \in \mathcal{P}_D$, $G_A(X_A)$ and $G_B(X_B)$ are orthogonal, i.e.,

$$\int_E G_A(x_A) G_B(x_B) dP_X(x) = 0.$$

Proof: See [48], Theorem 3.3.

This result allows defining an importance measure through a coalitional decomposition of $\mathbb{V}(G(X))$, also known as the *functional analysis of variance* (FANOVA).

Corollary 2.2 (FANOVA). *Let X be mutually independent random inputs and $G(X) \in \mathbb{L}^2(\sigma_X)$ be a black-box model. Then,*

$$\mathbb{V}(G(X)) = \sum_{A \in \mathcal{P}_D} \mathbb{V}(G_A(X_A)),$$

and furthermore,

$$\forall A \in \mathcal{P}_D, \ \mathbb{V}(\mathbb{E}_A[G(X)]) = \sum_{B \in \mathcal{P}_A} \mathbb{V}(G_B(X_B)). \tag{2.2}$$

One can notice that, thanks to Eq. (2.1), Eq. (2.2) is reminiscent of Corollary 2.1 (i), which lead to the following characterization:

$$\forall A \in \mathcal{P}_D, \quad \mathbb{V}(G_A(X_A)) = \sum_{B \in \mathcal{P}_A} (-1)^{|A|-|B|} \mathbb{V}(\mathbb{E}_B[G(X)]).$$

thanks to Corollary 2.1 (ii). The importance measure is defined as

$$\begin{aligned} S : \mathcal{P}_D &\to \mathbb{R} \\ A &\mapsto S_A = \mathbb{V}(G_A(X_A)) = \sum_{B \in \mathcal{P}_A} (-1)^{|A|-|B|} \mathbb{V}(\mathbb{E}_B[G(X)]). \end{aligned} \tag{2.3}$$

are known as the *Sobol' indices*[1] in the variance-based global SA literature [204]. As a by-product, the

[1] Sobol' indices are usually normalized between 0 and 1, i.e., dividing the importance measure by $\mathbb{V}(G(X))$. However, for conciseness, the focus is put on their un-normalized version in this thesis.

value measure

$$\begin{aligned} S^{\text{clos}} : \mathcal{P}_D &\to \mathbb{R} \\ A &\mapsto S_A^{\text{clos}} = \mathbb{V}\left(\mathbb{E}_A\left[G(X)\right]\right). \end{aligned} \tag{2.4}$$

can be defined thanks to Eq. (2.2). It is known as the *closed Sobol' indices* in the global SA literature [146, 48]. The Sobol' and closed Sobol' indices are illustrated in Figure 2.2. The Venn diagram on the left represents $\mathbb{V}\left(G(X)\right)$, which can be decomposed using the importance measure S. The value measure S^{clos} then, as in Eq. (2.2), is the sum of the Sobol' indices S related to each subset of inputs.

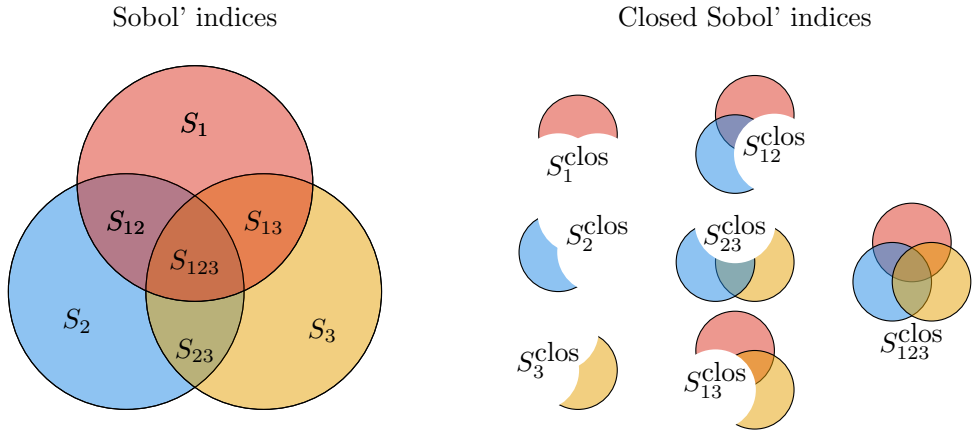

Figure 2.2: Illustration of the Sobol' and closed Sobol' indices for three inputs.

In this example, one leverages a *coalitional decomposition of the random output* $G(X)$, leading to Eq. (2.1), which particularizes to Eq. (2.2) for the variance as a QoI, which, thanks to Corollary 2.1, lead to the characterization of the importance measure in Eq. (2.3). This illustrates the model-centric approach to defining influence measures since it originates from a decomposition of $G(X)$.

Whenever the inputs are assumed to be mutually independent, it can be shown that the Sobol' indices (i.e., Eq. (2.3)) *form a properly gradual influence measure*, since, for every $A \in \mathcal{P}_D$, the summand $G_A(X_A)$ in Hoeffding's decomposition is *a proper representant of* X_A (see, Chapter 5).

2.3.2 Dependent inputs and the input-centric approach

Now, suppose that *the inputs are not necessarily mutually independent*. Hence, the decomposition provided in Theorem 2.1 does not hold. However, notice that since $G(X) \in \mathbb{L}^2\left(\sigma_X\right)$, the quantities $\mathbb{V}\left(\mathbb{E}_A\left[G(X)\right]\right)$ exist for every $A \in \mathcal{P}_D$, since conditional expectations are contractive operators. Choose, for instance, the closed Sobol' indices as a *value measure*, i.e.,

$$\begin{aligned} S^{\text{clos}} : \mathcal{P}_D &\to \mathbb{R} \\ A &\mapsto S_A^{\text{clos}} = \mathbb{V}\left(\mathbb{E}_A\left[G(X)\right]\right). \end{aligned}$$

It then leads to the following input-centric importance measure:

$$\forall A \in \mathcal{P}_D, \quad S_A = \sum_{B \in \mathcal{P}_A} (-1)^{|A|-|B|} \mathbb{V}\left(\mathbb{E}_B\left[G(X)\right]\right) = \sum_{B \in \mathcal{P}_A} (-1)^{|A|-|B|} S_B^{\text{clos}}. \tag{2.5}$$

Notice that since $S_D^{\text{clos}} = \mathbb{V}\left(G(X)\right)$, thanks to Proposition 2.2, S is a coalitional decomposition of $\mathbb{V}\left(G(X)\right)$. It is interesting to note that *even if the inputs are not mutually independent*, the importance measure S *remains a coalitional decomposition of the variance of the random output*, without the need for a coalitional decomposition of $G(X)$ itself.

As its name suggests, the input-centric approach begins with choosing a *value measure*, which can be interpreted as an *a-prior* way to quantify the importance of a subset of inputs. Here, the choice of the closed Sobol' indices can be justified by the fact that it represents the variance *of the best approximation of* $G(X)$ *by a function of* $\mathbb{L}^2\left(\sigma_A\right)$ (the conditional expectation being an orthogonal projection) [48, 104].

Remark 2.4 (Sobol' indices and dependence). One can notice that Eq. (2.3) and Eq. (2.5) are similar. One is defined in a model-centric manner, which requires mutual independence, and the other is defined without the need for mutual independence. Hence, the input-centric variance decomposition in Eq. (2.5) *is equivalent to the one in* Eq. (2.3) **if and only if** the inputs are mutually independent (see, Chapter 4). The main remark is that the dependence structure between the inputs **does play a role** in the characterization and the properties of influence measures.

2.3.3 Illustration on controlled examples

Gaussian inputs, linear model, and a correlated exogenous input. Consider the model:

$$G(X) = X_1 + X_2, \quad X = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 & \rho \\ 0 & 1 & 0 \\ \rho & 0 & 1 \end{pmatrix} \right). \tag{2.6}$$

Here, $G(X)$ is a function of X_1 and X_2, but the inputs include an *exogenous input* X_3, possibly correlated to X_1. The notion of exogenous input, defined formally in Chapter 3, can be broadly understood as "an input that is not in the model". In this particular use-case, since the inputs are Gaussian, it is possible to compute the analytically valued of the Sobol' indices, and thus the importance measures defined above. They are presented in Table 2.1.

Subset of inputs	$S^{\text{clos}}/\mathbb{V}(G(X))$	$S/\mathbb{V}(G(X))$
$\emptyset$	0	0
$\{1\}$	$1/2$	$1/2$
$\{2\}$	$1/2$	$1/2$
$\{3\}$	$\rho^2/2$	$\rho^2/2$
$\{1,2\}$	1	0
$\{1,3\}$	$1/2$	$-\rho^2/2$
$\{2,3\}$	$(1+\rho^2)/2$	0
$\{1,2,3\}$	1	0

Table 2.1: Analytical values for normalized S^{clos} and S for the illustration in Eq. (2.6).

First, recall that the inputs are mutually independent, and hence, the influence measure is model-centric whenever $\rho = 0$. In this case, the Sobol' indices indicate that half of the importance is granted to X_1 and the other half to X_2. This interpretation is reasonable since $G(X)$ is defined as the sum of these two inputs having the same variance. However, one can notice that whenever $\rho \neq 0$, and the influence measure becomes input-centric, a positive share of importance is granted to X_3, compensated by a negative share granted to the coalition X_{13}. This can be understood as being due to the correlation between X_1 and X_3. However, several questions arise:

- *How can a negative share of importance be interpreted?*
- *Why is the importance given to X_3 substituted to the importance given by X_{13}? Why is it not the other way around?*
- *How would one detect that X_3 is exogenous, even though it has a non-zero importance?*
- *Is it possible to distinguish the effects due to the dependence structure between the inputs from the effects intrinsically due to the model?*

Gaussian inputs, linear model with an interaction term. Consider the model:

$$G(X) = X_1 + X_2 X_3, \quad X = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 & \rho \\ 0 & 1 & 0 \\ \rho & 0 & 1 \end{pmatrix} \right). \tag{2.7}$$

Here, $G(X)$ is a linear function of the three inputs with an interaction term between X_2 and X_3, and X_3 may be correlated with X_1. In this case, the value measure S^{clos} and influence measure S can be computed analytically and are listed in Table 2.2. Whenever the inputs are mutually independent (i.e., $\rho = 0$),

Subset of inputs	$S^{\text{clos}}/\mathbb{V}\left(G(X)\right)$	$S/\mathbb{V}\left(G(X)\right)$
$\emptyset$	0	0
$\{1\}$	$1/2$	$1/2$
$\{2\}$	0	0
$\{3\}$	$\rho^2/2$	$\rho^2/2$
$\{1,2\}$	$(1+\rho^2)/2$	$\rho^2/2$
$\{1,3\}$	$1/2$	$-\rho^2/2$
$\{2,3\}$	$(1+\rho^2)/2$	$1/2$
$\{1,2,3\}$	1	$-\rho^2/2$

Table 2.2: Analytical values for normalized S^{clos} and S for the illustration in Eq. (2.7).

one can notice that half the importance is attributed to X_1, and the other half to X_{23}, which is comprehensible because $G(X)$ has an interaction term between X_2 and X_3. However, if ρ is different from zero, one can notice that X_3 and X_{12} do receive a positive share of importance, which is compensated by negative shares distributed to X_{13} and X_{123}. The interpretation becomes complicated, and having access to the influence measures (and not the model) does not draw an accurate picture of the model $G(X)$, e.g., one might wonder if X_3 has some individual importance or if there is an interaction (i.e., not due to the dependence) between X_1 and X_2.

Whenever the inputs are mutually independent, the Sobol indices, as defined in Eq. (2.3), seem to be accurate according to the intricacies of the model. However, in correlated settings, when defined from an input-centric scheme, as in Eq. (2.5), their interpretation is not as clear. However, input-centric influence measures can still be helpful, especially when aggregated in a certain way, which is the topic of the following chapter.

2.4 Partial conclusion

This chapter highlights the combinatorics and algebraic roots of the question behind influence measurement. The sole assumption that subsets of inputs may have different degrees of influence on QoIs of black-box models and thus can be ranked justifies the definition of influence measures as order-embeddings. One particular class of influence measures, the *coalitional decompositions*, are particularly interesting. As their name suggests, these influence measures can be understood as an additive decomposition of a QoI over the power-set of D and hence be endowed with an intuitive interpretation. However, not every coalitional decomposition can be a suitable candidate to express the total order over the subsets of inputs. To that extent, *(proper) graduality* is introduced: the evaluations of the influence measure related to a subset of inputs must be the QoI of a representant of this subset.

Rota's generalization of the Möbius inversion formula for power-sets allows two approaches to define such influence measure. The first, called *input-centric*, relies on a combinatorial mechanism and is not dependent on the inputs' dependence structure. The second, called *model-centric*, relies on the ability to conditionally decompose $G(X)$, seen as a random element. This connexion introduces the notion of *value measure*, a set function whose Möbius transform is an influence measure.

These two approaches are illustrated for the problem of importance quantification, i.e., variance decomposition. Regarding mutually independent inputs, the model-centric approach is similar to the well-known definition of the Sobol' indices. When choosing the closed Sobol' indices as a suitable value measure, the input-centric approach allows defining importance measures even if the inputs are *not mutually independent*. However, in this case, their interpretation raises some questions, which are highlighted through analytical computations on some simple use cases.

In Chapter 3, the input-centric approach, using the closed Sobol' indices, is used in order to define *attribution methods* inspired from *cooperative game theory*, i.e., leverage a coalitional decomposition in order to define importance measure *for each input* (and not the subsets of inputs). These *allocations* can be interpreted as aggregations of input-centric coalitional decompositions.

In Chapter 4, the model-centric approach is further explored. Hoeffding's decomposition, presented in Theorem 2.1, is effectively generalized to *not mutually independent inputs*. It paves the way towards the definition of *properly gradual influence measures*, with interesting and intuitive theoretical properties.

Chapter 3

Input-centric approach and cooperative games

Contents

3.1 Introduction **31**
3.2 Cooperative games and allocations **31**
3.2.1 Cooperative games and allocations 32
3.2.2 Egalitarian allocation: the Shapley values 35
3.2.3 Proportional allocation: the proportional values 37
3.3 Importance attribution with dependent inputs **37**
3.3.1 Sobol' cooperative games and exogeneity 38
3.3.2 The Shapley effects 39
3.3.3 The proportional marginal effects 40
3.3.4 Illustrative examples 42
3.4 Illustration on use-cases **44**
3.4.1 Estimation schemes 44
3.4.2 River water level 46
3.4.3 Optical filter transmittance 48
3.4.4 Acoustic fire extinguisher 49
3.5 The fundamental problem of the input-centric approach **50**

Abstract (English). One way to construct coalitional decompositions of a quantity of interest through the input-centric approach is by drawing parallels between influence measures and resource allocations provided by cooperative game theory. This research domain can be summarized as follows: Given a set of players and a function measuring the value of each coalition, one seeks to redistribute the total value produced to each player. A general way to construct efficient allocations, i.e., the ones redistributing the entire total value, relies on the concept of *Harsanyi dividends*. These dividends can be interpreted as influence measures and the allocations as aggregations of these measures. The *Shapley values* are an example of such allocation, as well as *proportional values*. For the problem of importance quantification, i.e., the decomposition of the variance of a black-box model, Shapley values for a specific choice of value function, known as *Shapley effects*, do not detect exogenous inputs (those not in the model). To address this issue, proportional values have been adapted to importance quantification, giving rise to *proportional marginal effects*, providing an interpretable importance quantification while detecting exogenous inputs. The behavior of these indices is studied through analytical and real case studies. The use of the input-centric approach is discussed.

Abstract (Français). Une manière de construire des décompositions coalitionnelles de quantité d'intérêt par l'approche focalisée sur les entrées, est en faisant un parallèle entre les mesures d'influence et les allocations de ressources offertes par la théorie des jeux coopératifs. Ce domaine de recherche peut être résumé ainsi : étant donné un ensemble de joueurs et une fonction permettant de mesurer la valeur de chaque coalition, comment redistribuer la valeur totale produite à chaque joueur. Une manière assez générale de construire des allocations efficaces, i.e., permettant de redistribuer l'entièreté de la valeur totale, repose sur la notion de *dividendes d'Harsanyi*. Ces dividendes peuvent être interprétés comme des mesures d'influence, et les allocations comme des agrégations de cette mesure. Les *valeurs de Shapley* en sont un exemple, ainsi que les *valeurs proportionnelles*. Pour le problème de la quantification d'importance, i.e., la décomposition de la variance d'un modèle boîte-noire, les valeurs de Shapley pour un choix de fonction de valeur particulier, connues sous le nom d'*effets de Shapley*, ne permettent pas de détecter des entrées exogènes (qui ne sont pas dans le modèle). Pour remédier à ce problème, les valeurs proportionnelles ont été adaptées à la quantification d'importance, donnant naissance aux *effets proportionnels marginaux*, offrant une quantification d'importance interprétable tout en détectant les entrées exogènes. Le comportement de ces indices est étudié sur des cas d'études analytiques et réels. L'utilisation de l'approche focalisée sur les entrées est finalement discutée.

Keywords . Cooperative game theory • Sensitivity analysis • Harsanyi dividends • Shapley values • Proportional values • Exogeneity detection • Importance quantification

This chapter expands on the following contributions.

Journal articles:

M. Herin, M. Il Idrissi, V. Chabridon, and B. Iooss. Proportional marginal effects for global sensitivity analysis. *SIAM/ASA Journal on Uncertainty Quantification*, 2024. URL: https://hal.science/hal-03825935. In press

M. Il Idrissi, V. Chabridon, and B. Iooss. Developments and applications of Shapley effects to reliability-oriented sensitivity analysis with correlated inputs. *Environmental Modelling and Software*, 143:105115, 2021. ISSN: 1364-8152. DOI: 10.1016/j.envsoft.2021.105115

Pre-prints:

A. Foucault, M. Il Idrissi, B. Iooss, and S. Ancelet. Shapley effects and proportional marginal effects for global sensitivity analysis: application to computed tomography scan organ dose estimation. Preprint, 2023. URL: https://hal.science/hal-04114533

L. Clouvel, B. Iooss, V. Chabridon, M. Il Idrissi, and F. Robin. A review on variance-based importance measures in the linear regression context. Preprint, 2023. URL: https://hal.science/hal-04102053

Conferences:

M. Il Idrissi, N. Bousquet, F. Gamboa, B. Iooss, and J. M. Loubes. Coalitional decomposition of quantities of interest. In *2023 Annual Meeting of MASCOT-NUM Research Group*, Le Croisic, France, 2023. URL: https://mascotnum2023.sciencesconf.org/

M. Il Idrissi, V. Chabridon, and B. Iooss. Shapley effects for reliability-oriented sensitivity analysis with correlated inputs. In *Proceedings of the 10th International Conference on Sensitivity Analysis of Model Output*, Tallahassee, Florida, United States of America, 2022. URL: https://samo2022.math.fsu.edu/

M. Il Idrissi, B. Iooss, and V. Chabridon. Mesures d'importance relative par décomposition de la performance de modèles de régression. In *52èmes Journées de Statistique de la Société Française de Statistique (SFdS)*, Nice, France, 2021. URL: https://hal.science/hal-03149764

3.1 Introduction

When defining suitable influence measures, the first approach is to focus on the interdependencies of the inputs and aim at quantifying their effects on the QoI. There may be various sources for these interdependencies, e.g., functional interaction due to the model or stochastic dependence between the inputs that can affect the output differently. Each possible subset of inputs, when considered jointly, due to such interdependencies, can affect the propagated uncertainties of the random output, and hence, in-fine, affect the QoI. Hence, the analogy between inputs affecting a model and interacting agents evolving in a system is rather natural.

Game theory is dedicated to studying interactions between agents in a system [168]. The agents are called *players*, and the system in which they interact is called a game. Game theory can be divided into two categories: *cooperative* and *non-cooperative*. Non-cooperative games aim at modeling, in a *dynamic manner*, the best available *moves* each player can make in the game, usually in order to maximize their marginal utility. *Cooperative game theory* aims at studying the *outcomes of games* when the players are *arranged in different combinations*, i.e., considered jointly, in every possible way.

Hence, tackling the problem of measuring the influence of subsets of inputs is inherently similar to the framework of cooperative game theory. For instance, the Lindeman-Merenda-Gold importance indices for linear regression models [144]: they were developed independently of cooperative game theory, but, when studied under the paradigm of cooperative games, can be characterized as the Shapley values [201], a particular allocation of resources [210].

More recently, in global SA, [169] proposed novel variance-attributing indices, which promise to quantify the importance *even if the inputs are dependent*. These indices are introduced by analogy between the inputs of a black-box model and the players of a cooperative game based on the Shapley allocation of value. In XAI, the well-known SHAP method [149] relies on the same rationale to provide "explanations" on black-box model *evaluations*, often called "local explanations" in the literature.

In this chapter, the main focus is placed on the usage of cooperative games in order to define influence indices. First, the framework of cooperative game theory is introduced, as well as two classes of allocations: the Weber and the Harsanyi sets. The latter, which includes the former, is characterized as an *aggregation of dividends*. These dividends can be seen as input-centric influence measures. Then, two allocations are introduced: the Shapley values and the proportional values. Finally, the question of input importance quantification (i.e., the QoI is the variance of the random output) is addressed and illustrated analytically, as well as on three use-cases.

3.2 Cooperative games and allocations

The paradigm of cooperative game can be understood rather intuitively. The key idea behind this field is to study and characterize how "wealth" can be redistributed among players. These players are bound to interact with each other, producing some *value* (e.g., monetary). The value produced by each set of players (coalitions) interacting is measured using a *value function*. Given a set of players and a value function that associates some value to each coalition of players, the main question tackled in cooperative games is the question of *allocation*: *How can one redistribute the value of the grand coalition* (i.e., the set formed by *every player*) *among each individual player?*

This section formally introduces the framework of cooperative games and allocations. Two particular sets of allocations are presented and discussed: the *Weber set* and the *Harsanyi set*. Two allocations are studied in this manuscript: the *Shapley values* and the *proportional values*. Finally, a connection is with the input-centric approach to defining influence measures presented in Chapter 2.

Remark 3.1. The notion of cooperation (in the usual sense) between players is not intrinsic to cooperative games, nor is competition exclusive to *non-cooperative* games. The former focuses on a more global view of interactions between players, focusing on the resources produced by "combining players" (i.e., considering coalitions of players), while the latter aims at studying these interactions in much finer detail. Cooperative games can model competition and conflicts among players, and non-cooperative games can describe cooperation behaviors. Alternate naming scheme, although not standard, better translates the goals of each of these two approaches [31]: *procedural* game theory for non-cooperative games, and *combinatorial* game theory for the study of cooperative games.

3.2.1 Cooperative games and allocations

A (transferable-utility) cooperative game is a tuple (D, v) where $D = \{1, \dots, d\}$ is a set of d players and $v : \mathcal{P}_D \to \mathbb{R}$ is a *value function*, i.e., an application that maps a real value to every possible coalition (i.e., subset) of players, with the convention that $v(\emptyset) = 0$.

Whenever v is assumed to be *monotonically increasing*, meaning that, for any $A, B \in \mathcal{P}_D$, $B \subseteq A$ implies that $v(B) \leq v(A)$, or said differently, a "bigger coalition" bring the same or more value than a "smaller one", the cooperative game is said to be *monotonic*. Additionally, if v is positive (resp. nonnegative), the game is said to be positive (resp. nonnegative).

Duality: worth instead of value Cooperative games can be approached from a dual perspective. Under the initial paradigm presented above, the value function v can be interpreted as the amount of the value produced by a coalition of players. The dual approach focuses on the "worth" or "bargaining power" of a coalition, i.e., the shortfall in value due to a coalition [70, 69]. Formally, the dual of a cooperative game (D, v) is the cooperative game usually denoted by (D, w) where w is defined, for any $A \in \mathcal{P}_D$ as:

$$w(A) = v(D) - v(D \setminus A). \tag{3.1}$$

The quantities $w(A)$ can be interpreted as a measure of how crucial a coalition is in producing $v(D)$, i.e., the value the grand coalition loses by removing the coalition. In the following, one refers to $w(A)$ as the *marginal contribution of the coalition* A. The duality between (D, v) and (D, w) can be understood in the sense that the dual of (D, w) is the initial game (D, v). Since the dual game (D, w) is also a cooperative game, and $w(D) = v(D)$, it also offers a way to study how the value of the grand coalition can be allocated among players.

Allocations One of the main goals of cooperative game theory is to build *allocations* (also called solution concepts, or payoffs) [168]. An allocation of a cooperative game (D, v) can be understood as defining a *redistribution scheme* of $v(D)$ (i.e., the value produced by the grand coalition) amongst each player in D. Formally, an allocation can be understood as a mapping $\psi : D \to \mathbb{R}$, which, to every player $i \in D$, associates a real-value ψ_i.

Without additional constraints, there are infinite possible (and trivial) allocations for a particular cooperative game. Hence, to ensure their relevance, desirability criteria on the redistribution process are usually sought after. In the literature, these criteria are usually called "axioms". A large portion of the field of cooperative game theory amounts to defining allocations uniquely characterized by a set of suitable axioms. For the purposes of this thesis, two axioms are of interest:

- **Efficiency:** the allocation sums to $v(D)$, i.e., $\sum_{i \in D} \psi_i = v(D)$;
- **Non-negativity:** the allocation must be non-negative, i.e., $\forall i \in D,\ \psi_i \geq 0$.

There are many ways to define allocations. Different approaches lead to different sets of allocations. In particular, two are of interest in this thesis. The first one is called the *Weber set* and relies on random order schemes, i.e., considering *dynamic interactions* between players in every order possible. The second one, called the *Harsanyi set*, defines allocations as aggregations of dividends derived from the game's value function.

Random orders and the Weber set The random order point of view to define allocations can be understood under the following rationale. Suppose that players are bound to interact *dynamically*, following a certain pre-defined order. For instance, for three players $D = \{1, 2, 3\}$, consider the order $(2, 3, 1)$. Suppose that the players interact dynamically according to this order, i.e., first $\{2\}$ alone, then $\{2, 3\}$ and finally $\{1, 2, 3\}$. At each step of this dynamic, compute the *marginal contribution* relative to each step, i.e.,

1. Player $\{2\}$ is alone, the marginal contribution (to $\emptyset$) is $v(\{2\})$;
2. Now, player $\{3\}$ is introduced, and its marginal contribution to $\{2\}$ is then $v(\{2, 3\}) - v(2)$;
3. Finally, player $\{1\}$ is introduced, and its marginal contribution to $\{2, 3\}$ is then $v(D) - v(\{2, 3\})$.

Hence, each order induces a sequence of marginal contributions relative to introducing a single player in the dynamic governed by the order in which the players are introduced. The main idea behind random orders is to consider that each possible order is endowed with a certain chosen *probability*. To define allocations from random orders, say for a player $i \in D$, one can consider the *expectation* of the *marginal contribution due to the introduction of* i, over every possible ordering of players (i.e., weighted by their probability). Allocations that can be characterized in this fashion are often called *probabilistic values* (or random order allocations) [69, 100]. The set of probabilistic values of a cooperative game (D, v) is called the *Weber set* [223].

Formally, let $\mathcal{S}_D$ be the symmetric group on D (the set of all permutations of elements of D). To be consistent with the notation of [69], let $\pi = (\pi_1, \ldots, \pi_d) \in \mathcal{S}_D$ be a particular permutation, and for any $i \in D$, denote $\pi(i) := \pi_i^{-1}$ its inverse image (i.e., the position of i in π, such that $\pi_{\pi(i)} = i$). Then, one can define the following set of players for any $i \in \{0, \ldots, d\}$:

$$C_i(\pi) = \{\pi_j : j \leq i\}. \tag{3.2}$$

where, by convention, $C_0(\pi) = \emptyset$.

Remark 3.2. The set $C_i(\pi)$ can be understood as the set containing the i-th first players in the ordering π. As an illustration, let $D = \{1, 2, 3\}$, and let $\pi = (2, 1, 3) \in \mathcal{S}_D$. Then,

$$\pi(1) = 2, \quad \pi(2) = 1, \text{ and } \quad \pi(3) = 3.$$

Moreover,

$$C_{\pi(1)}(\pi) = C_2(\pi) = \{1, 2\}, \quad C_{\pi(2)}(\pi) = C_1(\pi) = \{2\}, \quad C_{\pi(3)}(\pi) = C_3(\pi) = \{1, 2, 3\}$$

As their names suggest, random order models endow $\mathcal{S}_D$ with a probabilistic structure. For a game (D, v), its Weber set contains every probabilistic value allocation ψ that can be written, for any $i \in D$, as:

$$\begin{aligned} \psi_i &= \sum_{\pi \in \mathcal{S}_D} p(\pi) \left[v\left(C_{\pi(i)}(\pi)\right) - v\left(C_{\pi(i)-1}(\pi)\right) \right] \\ &= \mathbb{E}_{\pi \sim p}\left[v\left(C_{\pi(i)}(\pi)\right) - v\left(C_{\pi(i)-1}(\pi)\right) \right] \end{aligned} \tag{3.3}$$

where p is a given probability mass function over the orderings of D. For a player i, its random order allocation can thus be interpreted as the expectation over the permutations π of D w.r.t. p, of the marginal contributions of i to the coalitions formed by $C_{\pi(i)-1}(\pi)$.

It is important to note that allocations in the Weber set are always *efficient* and, if the cooperative game is monotonic, they are *nonnegative* [223]. Hence, defining a probabilistic value amounts to *choosing a probability mass function* v *over the orderings* $\mathcal{S}_D$.

A parallel between the rationale behind random orders and the well-known "forward" and "backward" variable selection procedures can be drawn. Formally, one can notice that, for a player i and any permutation $\pi \in \mathcal{S}_D$, one has:

$$w\left(C_{\pi(i)}(\pi)\right) - w\left(C_{\pi(i)-1}(\pi)\right) = v\left(D \setminus C_{\pi(i)-1}(\pi)\right) - v\left(D \setminus C_{\pi(i)}(\pi)\right). \tag{3.4}$$

A random order model allocation of the dual of a cooperative game can be understood as the expected (w.r.t. a probability mass function p over $\mathcal{S}_D$) marginal contribution of a player i to the players that *follows* in the orderings' dynamic. In contrast, for the initial cooperative game, it is the expected marginal contribution of i to the players that *precedes* in the orderings' dynamic. As illustrated in Figure 3.1, considering the initial game (D, v) can be seen as considering a "forwards procedure" (players are added sequentially), whereas considering the dual game (D, w) can be seen as considering a "backward procedure" (players are removed from the grand coalition sequentially).

Harsanyi dividends and the Harsanyi set The overall philosophy behind the Harsanyi differs from the Weber set. It revolves around the *Harsanyi dividends* of a cooperative game (D, v). Introduced in [95], the dividend of a coalition of players $A \in \mathcal{P}_D$ of a cooperative game (D, v) is defined as:

$$\mathcal{D}_v(A) = \sum_{B \in \mathcal{P}_A} (-1)^{|A|-|B|} v(B).$$

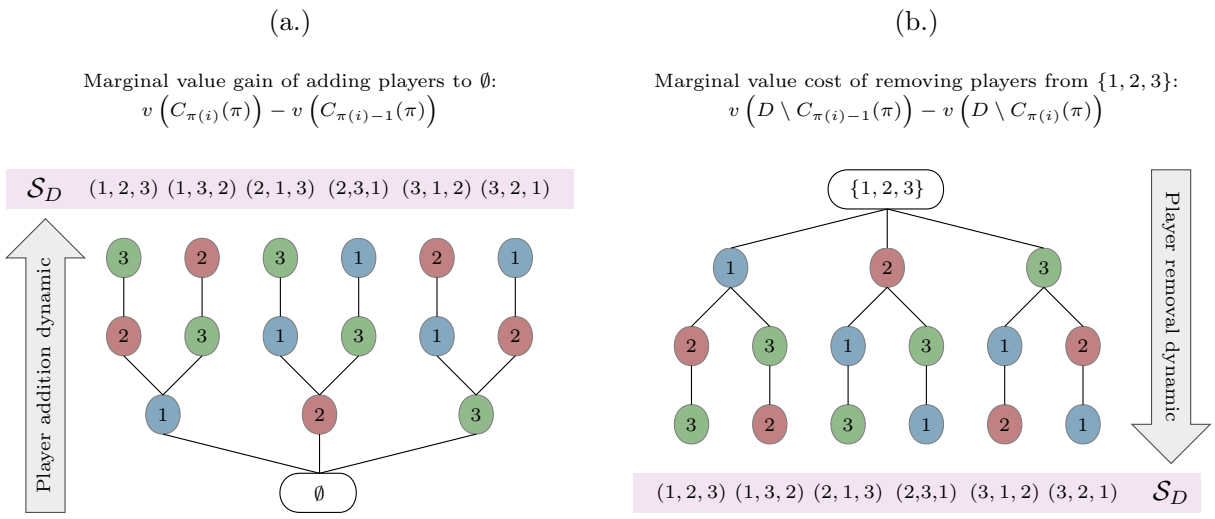

Figure 3.1: Analogy between random order model allocations and the forward-backward procedures for $D = \{1, 2, 3\}$: (a.) represents the allocation of a cooperative game as a forward procedure; (b.) illustrates the allocation of its dual as a backward procedure. The allocation of player 1 (resp. player 2 and 3) is the expected marginal gain (for a cooperative game (D, v)) or cost (for its dual (D, w)) computed for the blue (resp. red and green) ordering positions, weighted according to a probabilistic distribution over $\mathcal{S}_D$.

Coming from Chapter 2, one can notice that the Harsanyi dividends are none other than the *Möbius transform* of the value function, and thus, from Corollary 2.1, notice that *they sum up to* $v(D)$, i.e.,

$$\sum_{A \in \mathcal{P}_D} \mathcal{D}_v(A) = v(D).$$

Hence, the Harsanyi dividends can be understood as an *input-centric coalitional decomposition of* $v(D)$, defined as the Möbius transform of some chosen value function v. These dividends are usually interpreted as *the added value (or surplus) created by a coalition of players* in the literature, which is illustrated in Figure 3.2.

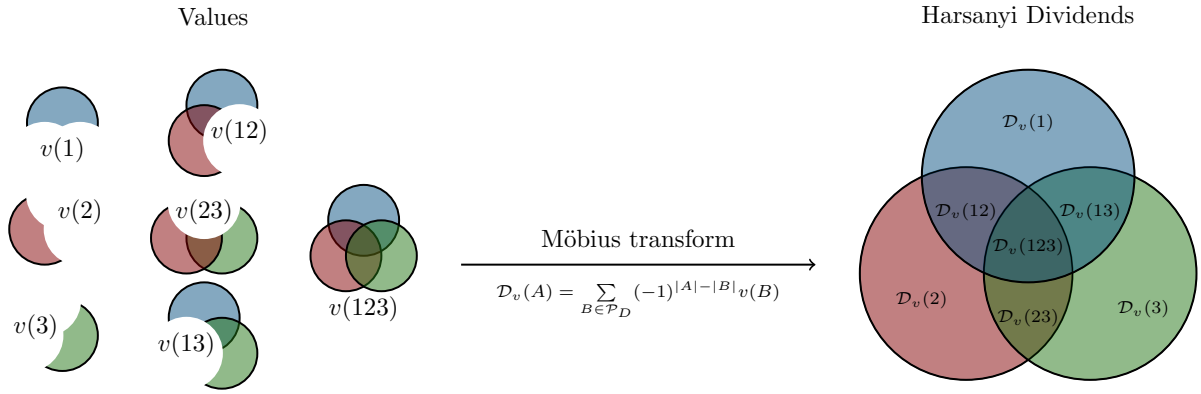

Figure 3.2: Illustration of the Harsanyi dividends for 3 players.

The *Harsanyi set* of a cooperative game (D, v), is the set of allocations that can be written as an *aggregation of the Harsanyi dividends*. Formally, an allocation ψ is in the Harsanyi set if it can be written, for every $i \in D$, as:

$$\psi_i = \sum_{A \in \mathcal{P}_D \,:\, i \in A} \lambda_i(A) D_v(A), \quad \text{where} \quad \begin{cases} \forall i \in D, \forall A \in \mathcal{P}_D,\ \lambda_i(A) \geq 0, \\ \forall A \in \mathcal{P}_D,\ \sum_{i \in D} \lambda_i(A) = 1. \end{cases} \tag{3.5}$$

As with the allocations in the Weber set, the allocations in the Harsanyi set are always efficient, and if v is monotonic, they are also nonnegative. In fact, *the Harsanyi set generalizes the Weber set*, in the sense that any allocation in the Weber set can be expressed as in Eq. (3.5), thanks to the following result.

Theorem 3.1. *For any cooperative game (D, v), the Weber set of allocations on (D, v) is included in (in the sense that they can be expressed as an allocation in) the Harsanyi set of allocations on (D, v).*

Proof: [56] or [215], Theorem 4.1

3.2.2 Egalitarian allocation: the Shapley values

The Shapley values [201] of a cooperative game (D, v) are a fairly well-known allocation in the theory of cooperative games. Its popularity extends outside the game theory realm due to its rather intuitive formulation and the reasonable set of axioms that characterize it. The Shapley values can be interpreted through different (but equivalent) approaches. The original formulation in [201] defines the Shapley values as the allocation $\mathrm{Shap} : D \to \mathbb{R}$ of a cooperative game (D, v), for every $i \in D$, as

$$\mathrm{Shap}_i = \frac{1}{d} \sum_{A \subset \mathcal{P}_{D_{-i}}} \binom{d-1}{|A|}^{-1} \left[v\left(A \cup \{i\}\right) - v\left(A\right) \right],$$

where $D_{-i} = \{1, \dots, i-1, i+1, \dots, d\}$. The interested reader is referred to [201, 156] for an interpretation of this formula. In the context of this thesis, the focus is put on equivalent Shapley values characterizations as a member of the Weber and Harsanyi set and their subsequent interpretation through random orders and dividend-sharing paradigms.

Random order interpretation The Shapley values can be expressed as a random-order model allocation [223], as in Eq. (3.3), and are characterized by the particular choice of the *(discrete) uniform* probability mass function over the orderings $\mathcal{S}_D$:

$$\mathrm{Shap}_i = \frac{1}{d!} \sum_{\pi \in \mathcal{S}_D} \left[v\left(C_{\pi(i)}(\pi)\right) - v\left(C_{\pi(i)-1}(\pi)\right) \right], \tag{3.6}$$

i.e., $p(\pi) = 1/d!$, for every $\pi \in \mathcal{S}_D$. Hence, they can be interpreted as the choice that *maximizes the entropy* in the class of probability distributions supported on $\mathcal{S}_D$ [66]. From a Bayesian standpoint, this entails that, without any prior knowledge of the dynamic of the players, choosing the Shapley values constitutes the "best least-informative guess". In [200], Shapley himself qualified these values as "*[...] an a priori assessment of the situation, based on either ignorance or disregard of the social organization of the players*".

In light of this characterization, the Shapley values can be understood as the natural allocation if no information about the interdependencies of the players can be either inferred (e.g., through v) or gathered externally (e.g., from experts' opinion). Additionally, since every permutation is granted the same probability, the formulation in Eq. (3.6) traduces a first glimpse at the "egalitarian treatment" of the Shapley values: every ordering dynamic has the same weight. However, this egalitarian redistribution becomes clearer when characterized as an element of the Harsanyi set.

Dividend sharing interpretation In [95], John C. Harsanyi showed that the Shapley values could be written as an aggregation of dividends. It writes, for every player $i \in D$

$$\mathrm{Shap}_i = \sum_{A \in \mathcal{P}_D \,:\, i \in A} \frac{\mathcal{D}_v(A)}{|A|}. \tag{3.7}$$

In other words, the dividend produced by a coalition $A \in \mathcal{P}_D$ is split into $|A|$ equal shares, which are redistributed among the players in A, without acknowledging their individual contributions, or the contributions due to their interdependencies. This egalitarian redistribution process is illustrated in Figure 3.3.

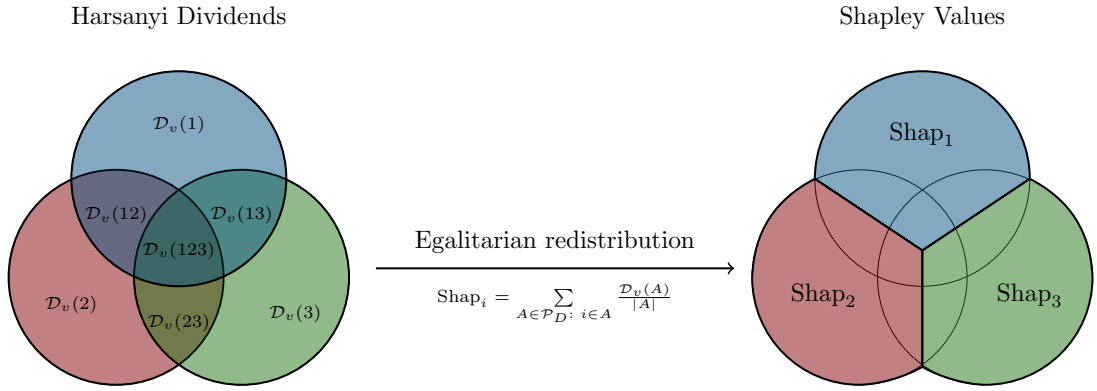

Figure 3.3: Illustration of the egalitarian redistribution of dividends of the Shapley values.

Axiomatic interpretation They values can also be characterized axiomatically (see [69]), as the only allocation ψ respecting the following two axioms:

- **Efficiency**: $\sum_{i=1}^{d} \psi_i = v(D)$;
- **Balanced contributions:** for all $A \in \mathcal{P}(D)$, and for all $i, j \in A$, $i \neq j$:

$$\psi_i(A, v) - \psi_i(A_{-j}, v) = \psi_j(A, v) - \psi_j(A_{-i}, v).$$

The second axiom entails that for any two different players i and j, the difference in each allocation by removing the other player to any sub-game (A, v) such that $i,\ j \in A$ must remain equal, for any $A \in \mathcal{P}_D$. In other words, the difference in allocation of the two players induced by the removal of the other player must be equal, implicitly entailing a balanced redistribution process where individual and coalitional contributions are favored equally.

More commonly, four sets of axioms are highlighted when it comes to the characterization of the Shapley values, as the unique allocation ψ of a game (D, v) respecting:

- **Efficiency**: $\sum_{j=1}^{d} \psi_j = v(D)$;
- **Symmetry**: If $v(A \cup \{i\}) = v(A \cup \{j\})$ for all $A \in \mathcal{P}_d$, then $\psi_i = \psi_j$, meaning that if two players show the same marginal contribution to *every* coalition, their payoff should be the same;
- **Dummy**: If $v(A \cup \{i\}) = v(A)$ for all $A \in \mathcal{P}_d$, then $\psi_i = 0$, meaning that if a player has a zero marginal contribution to *every* coalition, its payoff should be zero;
- **Additivity**: If two games (D, v) and (D, v') have Shapley values ψ and ψ' respectively, then the game $(D, v + v')$ has Shapley values $\psi + \psi'$.

Remark 3.3 (Shapley values and fairness). In the recent literature, especially in applied fields, the use of Shapley values is motivated by the notion of "fairness" [149], instead of choosing the more precise adjective of "egalitarian". One can think of many situations where an equal redistribution would be unfair.

For instance, suppose that two software engineers produce lines of code. The first engineer produces 10.000 lines by itself, while the second only produces 5.000. When working together, the second engineer decides not to work, and thus, together, they only produce 10.000 lines of code since only the first engineer has been hard at work. Hence, the dividend of the coalition of both players is a penalization of -5.000 lines of code, and, according to the Shapley values, *both engineers should be penalized equally*, even though *the first engineer has done all the work*.

In this situation, the equal redistribution of the dividends is not a *fair decision* because the Shapley values are blind to *the interdependencies between the players*. Hence, choosing the Shapley values as an allocation does not ensure a fair redistribution process. In game theory, the notions of *fair division* [160] do not necessarily imply equal treatment.

When considering dual games, the Shapley values display a particular behavior.

Proposition 3.1. *Let (D, v) be a cooperative game, and (D, w) be its dual. Then, the Shapley values of (D, v) are equal to the Shapley values of (D, w).*

Proof: See, [83] Lemma 2.7.

It is important to note that this behavior is inherent to the Shapley values and can not hold, in general, for other allocations.

3.2.3 Proportional allocation: the proportional values

The *proportional values*, as their name suggests, entail a *proportional redistribution of value* instead of the egalitarian redistribution offered by the Shapley values. Here, the notion of proportionality goes beyond the proportionality w.r.t. to the value of individual players (which characterize the *proportional Shapley values*, see, e.g., [17]). Here, the redistribution is to be understood as being proportional to a player's contribution to *every coalition*.

Formally, for a *positive game* (D, v), its *proportional values* refers to the allocation $\mathrm{PV} : D \to \mathbb{R}$ in the Weber set, associated with the particular choice of probability mass function over $\mathcal{S}_D$ [70],

$$p(\pi) = \frac{L(\pi)}{\sum_{\sigma \in \mathcal{S}_D} L(\sigma)}, \quad \text{where} \quad L(\pi) = \exp\left(- \sum_{j \in D} \log\left(v\left(C_j\left(\pi \right) \right) \right) \right). \tag{3.8}$$

This choice of probability mass function is motivated by the following axioms, which uniquely characterize the proportional values [69], as the unique allocation ψ such that:

- **Efficiency**: $\sum_{i=1}^{d} \psi_i = v(D)$;
- **Equal proportional gains**: for all $A \in \mathcal{P}(D)$, and for all $i, j \in A$, $i \neq j$:

$$\frac{\psi_i\big((A, v)\big)}{\psi_j\big((A, v)\big)} = \frac{\psi_i\big((A_{-j}, v)\big)}{\psi_j\big((A_{-i}, v)\big)}.$$

The equal proportional gains axiom sheds light on the redistribution dynamic of this particular allocation scheme. For any two different players i and j, the ratio of their allocations in any subgame (A, v) (for every $A \in \mathcal{P}_D$ such that $i,\ j \in A$) must be invariant to removing each player's contribution to the other's allocation. In other words, the magnitude of the ratios must be preserved, independently of the possible interaction between i and j, within any coalition they can be a part of. It implies that the allocation favors the players proportionally to their (marginal) contributions to every possible coalition in the redistribution process.

Ratio potential and recursive formulation The proportional values can also be characterized recursively [71, 167], for every $i \in D$, as:

$$\mathrm{PV}_i = \frac{R(D, v)}{R(D_{-i}, v)} \tag{3.9}$$

where, for all $A \in \mathcal{P}(D)$, $R(A, v) = v(A) \left(\sum_{j \in A} R(A_{-j}, v)^{-1} \right)^{-1}$, with $R(\emptyset, v) = 1$. The function R is commonly known as a *ratio potential* (see, [71], Section 3). This equivalent formulation is important in order to *extend the proportional values to nonnegative games*. Additionally, unlike the Shapley values, one can notice, from Eq. (3.9), that in general, the proportional values of a cooperative game are different from the ones of its dual.

3.3 Importance attribution with dependent inputs

This section focuses on *variance decompositions* and building *importance measures*. To that extent, the framework presented in Section 1.2 is restricted according to the following assumptions.

Special case . In this section, the following is assumed:

- The output space $Y = \mathbb{R}$, i.e., the model is $\mathbb{R}$-valued;
- The space of random outputs $\mathcal{G}_X$ is restricted to $\mathbb{L}^2(\sigma_X)$;
- The QoI is $\mathbb{V}(G(X))$, and hence the QoI space $Q = \mathbb{R}$;

3.3.1 Sobol' cooperative games and exogeneity

Sobol' cooperative games. The analogy between the players D of a cooperative game (D, v) and the inputs $X_1, \dots, X_d$ of a black-box model has been made in [169]. In this pioneering paper, the author chose the *closed Sobol' indices*, defined as the input-centric value measure for dependent inputs, as in Section 2.3.2, for the value function of choice. Recall that these indices exist (i.e., are not infinite) as long as $G(X) \in \mathbb{L}^2(\sigma_X)$. This choice of value function defines a particular cooperative game, called *the Sobol' cooperative game* [100].

Definition 3.1 (Sobol' cooperative game). Let $X = (X_1, \dots, X_d)$ be random inputs, and let $G(X) \in \mathbb{L}^2(\sigma_X)$ be an $\mathbb{R}$-valued random output. Let the value measure S^{clos} be defined as:

$$\begin{aligned} S^{\text{clos}} : \mathcal{P}_D &\to [0, \infty) \\ A &\mapsto \mathbb{V}(\mathbb{E}_A[G(X)]) \end{aligned}$$

The Sobol' cooperative game with inputs X and output $G(X)$ is the cooperative game $\left(D, S^{\text{clos}}\right)$, where S^{clos} are the closed Sobol' indices.

Sobol' cooperative games are *nonnegative* and *monotonic*. The choice of S^{clos} as a value function is usually motivated as representing the variance of the best approximation of $G(X)$ of functions in the subspaces $\mathbb{L}^2(\sigma_A)$ of $\mathbb{L}^2(\sigma_X)$ [104, 48].

The dual of a Sobol' cooperative game $\left(D, S^{\text{clos}}\right)$ is the nonnegative, monotonic cooperative game $\left(D, S^T\right)$, where $S^T : \mathcal{P}_D \to [0, \infty)$ denotes the total Sobol' indices [48], defined, for any $A \in \mathcal{P}(D)$, as

$$S_A^T := \mathbb{V}(G(X)) - \mathbb{V}(\mathbb{E}_{-A}[G(X)]) = \mathbb{E}_{-A}\left[(G(X) - \mathbb{E}_{-A}[G(X)])^2\right] \tag{3.10}$$

where $-A := D \setminus A$, and the equality in Eq. (3.10) comes from the *law of total variance*.

From this choice of value measure, as in Eq. (2.5), define the *input-centric* coalitional decomposition of the variance:

$$\begin{aligned} S : \mathcal{P}_D &\to \mathbb{R} \\ A &\mapsto \sum_{B \in \mathcal{P}_A} (-1)^{|A|-|B|} S_B^{\text{clos}}, \end{aligned}$$

which, under the cooperative game theory paradigm, are *none other than the Harsanyi dividends of the game* $\left(D, S^{clos}\right)$. As highlighted in Section 2.3.2, this influence measure remains a coalitional decomposition even if the inputs are not mutually independent due to the input-centric mechanism.

A mathematical definition of exogeneity. As introduced in Chapter 1, one of the main goals of global SA is to be able to detect exogenous inputs. There is a difference between inputs having a negligible effect on the model's uncertainty and the notion of exogeneity. In Section 2.3.3, the notion of exogeneity has been introduced as "an input or a set of inputs that are not in the model". This intuitive definition can be formalized as follows:

Definition 3.2 (Exogeneity). Let $X = (X_1, \dots, X_d)$ be random inputs, $G(X)$ be a random output, and let $i \in D$. The X_i is said to be *an exogenous input* if, there exists some $f(X_{-i}) \in \mathbb{L}^2(\sigma_{-i})$ such that $G(X) = f(X_{-i})$ a.s..

Moreover for $A \in \mathcal{P}_D$, if there exists some $f(X_{-A}) \in \mathbb{L}^2(\sigma_{-A})$ such that $G(X) = f(X_{-A})$ a.s., then the subset of inputs X_A is said to be *an exogenous vector*.

It is important to note that, according to the proposed definition, a set of exogenous inputs does not necessarily form an exogenous vector. For instance, consider three inputs (X_1, X_2, X_3) such that $X_1 = X_2$ a.s. Then, for the model

$$G(X_1, X_2, X_3) = X_3 + X_1 = X_3 + X_2,$$

X_1 and X_2 both appear to be exogenous, but the random vector (X_1, X_2) is not. However, these situations are often related to functional equality between the inputs, which can be easily remediated with an appropriate assumption (see, Proposition 4.2 in Chapter 4), but require extensive developments. For the context of this chapter, in order to remediate these situations, the following assumption is introduced:

Assumption 1. *Let $E \in \mathcal{P}(D)$. If for every $i \in E$, X_i is exogenous, then X_E forms an exogenous vector.*

3.3.2 The Shapley effects

The *Shapley effects* are none other than the Shapley values of the Sobol' cooperative game $\left(D, S^{\text{clos}}\right)$ [169]. They were first introduced as an attribution method to quantify individual input importance for *not necessarily mutually independent inputs*. They can be written, using its characterization as an aggregation of the coalitional decomposition S, as the allocation $\text{Sh} : D \to \mathbb{R}$ defined, for every $i \in D$, as:

$$\text{Sh}_i = \sum_{A \in \mathcal{P}_D \,:\, i \in A} \frac{S_A}{|A|}. \tag{3.11}$$

Remark 3.4. In [169], the equality in Eq. (3.11) is shown to hold if the inputs are mutually independent. This comes from the fact that, in this paper, the definition of the Sobol' indices S is directly bound to Hoeffding's FANOVA in Corollary 2.2, which requires mutual independence. In the present case, the definition of S is broader, as the Möbius transform of S^{clos}, which is well-defined even if the inputs are not mutually independent and happens to be equivalent whenever the inputs are.

These indices have been extensively studied in the literature [170, 121, 23]. In [206], the equivalence of the Shapley effects of the Sobol' cooperative game $\left(D, S^{\text{clos}}\right)$ and the ones of its dual $\left(D, S^T\right)$ has been highlighted, which is none other than the expression of Proposition 3.1 on this particular game, which enabled interesting (and more efficient) estimation schemes.

These indices are particularly interesting since Sobol' cooperative games (and their dual) are *monotonic*. Since the Shapley values are part of the Harsanyi set, it implies that they are *nonnegative*, meaning that, if divided by $\mathbb{V}\left(G(X)\right)$, the Shapley effects *can be interpreted as percentages of the output's variance*. Effectively, it offered a solution to the lack of functional decomposition (such as in Theorem 2.1) for non-mutually independent inputs. Their popularity is strengthened by the fact that their estimation only requires the ability to estimate S^{clos} for every possible subset of inputs, which, leveraging the plethora of methods from the SA literature (see, e.g., [206, 33, 178, 89, 19]), effectively paved the way towards plug-in estimation schemes.

Shapley's joke. The Shapley effects display a peculiar behavior when it comes to exogenous inputs. When the random inputs are correlated, the Shapley effects can allocate variance shares to exogenous inputs. This behavior has been spotted in [121] and has been informally called "Shapley's joke" in the SA community. One can illustrate this behavior thanks to the following model:

$$G(X) = X_1 + X_2, \quad X = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 & \rho \\ 0 & 1 & 0 \\ \rho & 0 & 1 \end{pmatrix} \right). \tag{3.12}$$

The Shapley effects can be computed analytically, and are equal to:

$$\frac{\text{Sh}_1}{\mathbb{V}\left(G(X)\right)} = 1/2 - \rho^2/4, \quad \frac{\text{Sh}_2}{\mathbb{V}\left(G(X)\right)} = 1/2, \quad \frac{\text{Sh}_3}{\mathbb{V}\left(G(X)\right)} = \rho^2/4.$$

One can notice that, as soon as $\rho \neq 0$, X_3 receives a non-zero share of the output's variance even though it is exogenous. In highly correlated settings, X_3 can even be interpreted as being almost as important as X_1. This interpretation can be meaningful because the correlation between X_3 and X_1 may be relevant

to the underlying studied phenomenon. However, the practitioner usually cannot access the model G. Hence, relying only on the Shapley effects, the practitioner could not determine the exogenous nature of X_3.

In order to define a suitable variance attribution that overcomes this peculiar behavior, one can refer to *different allocations of the Sobol' cooperative game*.

3.3.3 The proportional marginal effects

The proportional marginal effects (PME) are inspired by the proportional marginal variance decomposition (PMVD) indices, introduced in the context of linear regression models [70, 91]. These indices were developed as an exogenous-detecting replacement of the Lindeman-Merenda-Gold (LMG) indices. The LMG indices are none other than the Shapley values of a cooperative game, with the value function being equal to the determination coefficient R^2 computed using the linear regression using only subsets of inputs [92, 44]. These indices suffered from the identical drawback as the Shapley effects: they could not detect exogenous inputs. However, [70] proposed to use the proportional values instead of the Shapley values, leading to suitable importance indices with the added property of exogeneity detection. Hence, studying the proportional values (as defined in Section 3.2.3) of Sobol' cooperative games in order to define attribution with exogeneity detection abilities is only natural.

However, since the proportional values are only well-defined for positive games, they cannot be directly employed for Sobol' cooperative games and their dual. The first step is to *extend this allocation to nonnegative monotonic games*.

Extending the proportional values for nonnegative and monotonic cooperative games. By leveraging the method of [68], it is possible to define a continuous extension of the PVs for games with coalitions of zero value. The following result builds upon this extension.

Theorem 3.2 (Proportional value extension [100]). *Let (D, v) be a nonnegative and monotonic game with value function $v : \mathcal{P}_D \to [0, \infty)$.*

Denote $\mathcal{K}$ the set of largest (w.r.t. their cardinality) coalitions with zero value, i.e.,

$$\mathcal{K} = \left\{ A \in \mathcal{P}_D \ : \ |A| = \max_{B \in \mathcal{P}_D} \{|B|\} \ \textit{s.t.}\ v(B) = 0 \right\}.$$

Additionally, for any $i \in D$, denote the sets of largest zero coalitions that do not contain i by $\mathcal{K}_{-i}$, i.e.,

$$\mathcal{K}_{-i} = \left\{ A \in \mathcal{K} : i \notin A \right\}.$$

Define, for every $A \in \mathcal{K}$, the (necessarily) positive value function:

$$\begin{aligned} v_A : \mathcal{P}(D \setminus A) &\to (0, \infty) \\ B &\mapsto v_A(B) := v(B \cup A). \end{aligned}$$

Let $PV^* : D \to \mathbb{R}$ be the allocation defined as:

$$PV_i^* = \begin{cases} \dfrac{\sum_{A \in \mathcal{K}_{-i}} R(D_{-i} \setminus A, v_A)^{-1}}{\sum_{A \in \mathcal{K}} R(D \setminus A, v_A)^{-1}} & \text{if } i \notin \bigcap\limits_{A \in \mathcal{K}} A \\ 0 & \text{otherwise.} \end{cases} \tag{3.13}$$

Then, PV^* is a continuous extension of the proportional values to the set of nonnegative monotonic games, i.e., for a positive monotonic game (D, v),

$$PV_i^* = PV_i, \quad \forall i \in D.$$

Proof of Theorem 3.2 on p.123.

The function R in Eq. (3.13) refers to the ratio potential defined in Eq. (3.9). Interestingly, the definition of this extension precisely identifies the players whose allocation is equal to zero: a player is granted a

zero attribution if and only if it is part of every largest coalition with zero value.

This extension naturally leads to the definition of the allocation $\mathrm{PME} : D \to \mathbb{R}$.

Definition 3.3 (Proportional marginal effects [100]). Let X be random inputs, and let $G(X) \in \mathbb{L}^2(\sigma_X)$ be a random output, and let (D, S^{clos}) be their Sobol' cooperative game. The allocation $\mathrm{PME} : D \to \mathbb{R}$ is defined as the (extended) proportional values (i.e., Eq. (3.13)) of the dual game $\left(D, S^T\right)$.

Choosing the dual of a Sobol' cooperative game is especially suitable for exogenous input detection. This fact becomes clear thanks to the following result, inspired by the work in [96].

Lemma 3.1 (Total Sobol' indices and functional representation [100]). *Let $X = (X_1, \ldots, X_d)$ be random inputs and $G(X) \in \mathbb{L}^2(\sigma_X)$ be a random output. One has, $\forall A \subseteq D$,*

$$S_A^T = 0 \quad \Longleftrightarrow \quad G(X) = \mathbb{E}_{-A}\left[G(X)\right] \text{ a.s.}$$

Proof of Lemma 3.1 on p.125.

Naturally, as part of the Weber set (and thus the Harsanyi set), the PME are efficient, and since the dual of Sobol' cooperative games are monotonic, they result in a nonnegative allocation. Thus, as the Shapley effects, they can be interpreted as percentages of the output's variance. However, when it comes to exogenous inputs, they behave differently than the Shapley effects.

The PMEs and exogeneity detection. The PME, in the same fashion as the PMVD and unlike the Shapley effects, allows the detection of exogenous inputs, as shown in the following result.

Proposition 3.2 (PME exogeneity detection [100]). *Let $X = (X_1, \ldots, X_d)$ be random inputs and $G(X) \in \mathbb{L}^2(\sigma_X)$ be random output, and suppose that Assumption 1 holds. Then, for any input $i \in D$, the following equivalence holds:*

$$X_i \text{ is exogenous} \iff \mathrm{PME}_i = 0.$$

Proof of Proposition 3.2 on p.125.

In conjunction with the computation of the Shapley effects, this property can offer a more complete picture of the studied random output $G(X)$. For instance, coming back to the model in Eq. (3.12), the PMEs can be computed analytically and are equal to:

$$\frac{\mathrm{PME}_1}{\mathbb{V}\left(G(X)\right)} = 1/2, \quad \frac{\mathrm{PME}_2}{\mathbb{V}\left(G(X)\right)} = 1/2, \quad \frac{\mathrm{PME}_3}{\mathbb{V}\left(G(X)\right)} = 0.$$

The PMEs do indeed detect X_3 as being an exogenous input by granting it a zero allocation. Moreover, the PMEs are not influenced by the linear correlation ρ between X_1 and X_3. In combination with the Shapley effects, additional insights on $G(X)$ can be extracted from the initial study: while X_3 can affect $G(X)$ through its correlation with other inputs (supposedly known by the practitioner), it is exogenous to $G(X)$. Additionally, by allocating half the output's variance to X_1 and X_2, the PMEs also indicate an equal importance between both inputs. Hence, by combining the Shapley effects and the PME interpretation, one can interpret the results as follows: X_3 is an exogenous variable (PMEs). However, it affects $G(X)$ through its correlation with X_1 (Shapley effects) and X_1 and X_2 seem to have an equal influence on the output's variance, whenever X_3 is detected as exogenous (PMEs).

Hence, instead of taking the PMEs as a "better attribution method" because it does detect exogenous inputs, both allocation-inspired attribution methods can be used in conjunction in order to draw a more precise insight into the importance dynamics inherent to the modeled phenomena (the output $G(X)$ but also the probabilistic structure governing the inputs X).

Moreover, since the underlying redistribution process is *proportional*, unlike the *egalitarian* redistribution process of the Shapley values, they can end up in a fundamentally different attribution of the output's variance. These differences are studied using analytical examples in the following section.

3.3.4 Illustrative examples

Unbalanced linear model with three Gaussian inputs Beyond detecting exogenous inputs, the Shapley effects and the PMEs fundamentally differ in their redistribution process. While the Shapley effects allocate importance in an egalitarian fashion, the PME follows a proportional principle. This toy-case aims to highlight this difference by introducing a coefficient in a linear model with three correlated Gaussian inputs. This use-case is referred to as *unbalanced* since the three linear coefficients differ. This toy-case writes:

$$G(X) = X_1 + \beta X_2 + X_3, \quad X = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & \rho \\ 0 & \rho & 1 \end{pmatrix} \right),$$

and in this case, $\mathbb{V}(G(X)) = 2 + \beta^2 + 2\rho\beta$.

The analytical (unnormalized) shares of output variance, according to the Shapley effects and the PMEs, are given by

$$\mathrm{Sh}_1 = 1, \quad \mathrm{Sh}_2 = \beta^2 + \beta\rho + \frac{1}{2}\rho^2(1-\beta^2), \quad \mathrm{Sh}_3 = 1 + \beta\rho - \frac{1}{2}\rho^2(1-\beta^2),$$

and

$$\mathrm{PME}_1 = 1, \quad \mathrm{PME}_2 = \frac{\beta^2(1+\beta^2+2\rho\beta)}{(1+\beta^2)}, \quad \mathrm{PME}_3 = \frac{(1+\beta^2+2\rho\beta)}{(1+\beta^2)}.$$

One can notice that, by considering the *balanced* case (i.e., $\beta = 1$), the Shapley effects and PMEs are equal. However, as soon as the model is unbalanced, one can notice in Figure 3.4 that both allocations behave in a completely different fashion when it comes to the importance attribution of X_2 and X_3.

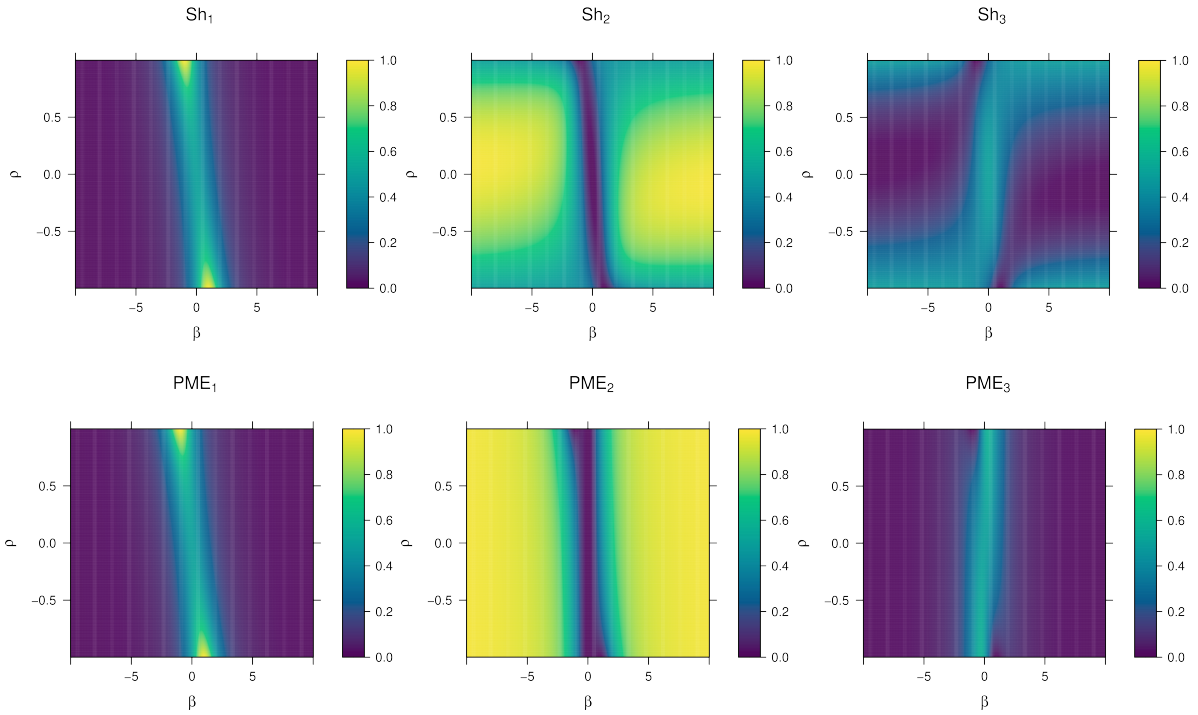

Figure 3.4: Normalized Shapley effects and PMEs for the unbalanced linear model with three Gaussian inputs.

In extreme cases of positive linear correlation between X_2 and X_3, the Shapley effects allocate half the importance to each input despite a fairly high β value in favor of X_2. The PMEs, on the other hand, tend to favor X_2 by granting it almost the whole output's variance despite its high correlation with X_3. This behavior highlights the "egalitarian vs. proportional" behavior of both effects: the Shapley effects consider X_2 and X_3 equally important due to their high correlation, while the PMEs highly favor X_3.

While these results inform on the asymptotic behavior of both indices, their difference can also be highlighted for punctual values of ρ and β. Figure 3.5 illustrates the behavior of both indices w.r.t. ρ, for two different values of β (namely, 2 and 10). Whenever $\beta = 2$, one can notice that PME_2 increases w.r.t. ρ,

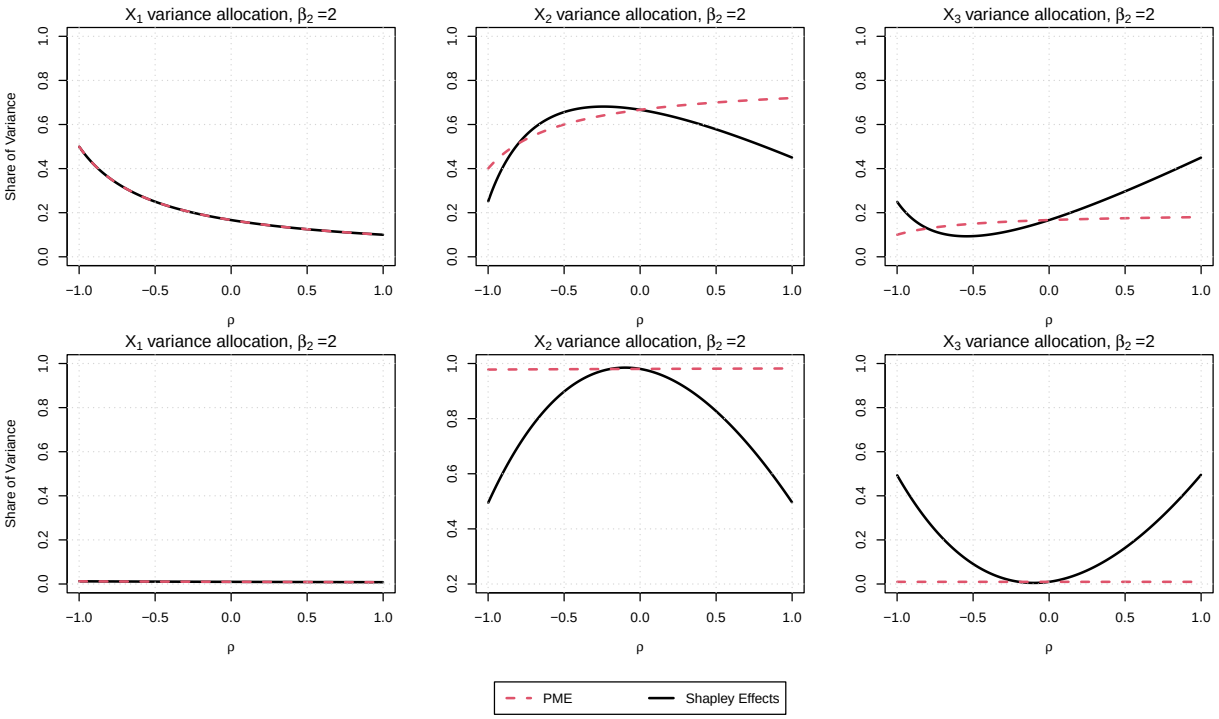

Figure 3.5: Normalized PMEs and Shapley effects w.r.t. ρ. The top row depicts the allocations for $\beta = 2$ while the bottom row is for $\beta = 10$.

while Sh_2 decreases after $\rho \simeq -0.24$, and both indices are concave w.r.t. ρ. On the other hand, Sh_3 is convex w.r.t. ρ and becomes increasing at $\rho \simeq -0.54$, while PME_3 remains concave increasing. At extreme values of ρ (i.e., close to -1 or 1), one can notice that Sh_2 and Sh_3 are considered equally important. Furthermore, one can notice that $\mathrm{PME}_2 > \mathrm{PME}_3$, whatever the magnitude of their correlation. Increasing β to 10 exacerbates this behavior of the Shapley effects. However, the PMEs behave differently: X_1 and X_3 are given a negligible part of the variance, while X_2 is granted a seemingly constant share, w.r.t. ρ, hovering around 98%.

In conclusion, in this unbalanced case, the proportional redistribution property of the PME allows for a more apparent importance hierarchy, even in situations of extreme correlation. On the other hand, the Shapley effect tends to even importance out between the correlated inputs, leading to a potentially indecisive importance hierarchy.

Linear model with two Gaussian inputs and an unbalanced interaction term. This toy case aims to study and compare the behavior in a trade-off between individual and interaction effects. This particular unbalanced linear model is given as follows:

$$G(X) = X_1 + (1-\alpha)X_2 + \alpha X_1 X_2, \quad X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right),$$

and in this case, $\mathbb{V}\left(G(X)\right) = 2 + (1-\alpha)^2 + 2(1-\alpha)\rho + \rho^2$. The parameter α aims at controlling the "trade-off" between the individual effect of X_2 and its interaction term with X_1. When $\alpha = 0$, there is no interaction term between X_1 and X_2, and when $\alpha = 1$, X_2 does not have any individual effect. In addition, both inputs are linearly correlated through their covariance $\rho \in (-1, 1)$.

The analytical (unnormalized) shares of output variance, according to the Shapley effects and the PMEs, are given by

$$\mathrm{Sh}_1 = \frac{3 + \rho^2(1-\alpha)^2 + 2\rho(1-\alpha)}{2}, \quad \mathrm{Sh}_2 = \frac{1 + 2\rho^2 + (2-\rho^2)(1-\alpha)^2 + 2\rho(1-\alpha)}{2},$$

and

$$\mathrm{PME}_1 = \frac{2}{3+(1-\alpha)^2} \times \mathbb{V}\left(G(X)\right), \quad \mathrm{PME}_2 = \frac{(1-\alpha)^2+1}{3+(1-\alpha)^2} \times \mathbb{V}\left(G(X)\right).$$

To illustrate the redistribution differences between the Shapley effects and the PMEs w.r.t. both correlation and interaction, (α, ρ)-plane plots are provided in Figure 3.6. First, one can notice that when $\alpha = 0$, the Shapley effects and the PMEs are equal, granting each input half of the output's variance. However, when α deviates from zero, both indices display different behaviors. Secondly, and interestingly, the analytical formulas of the PMEs do not depend on the correlation coefficient ρ.

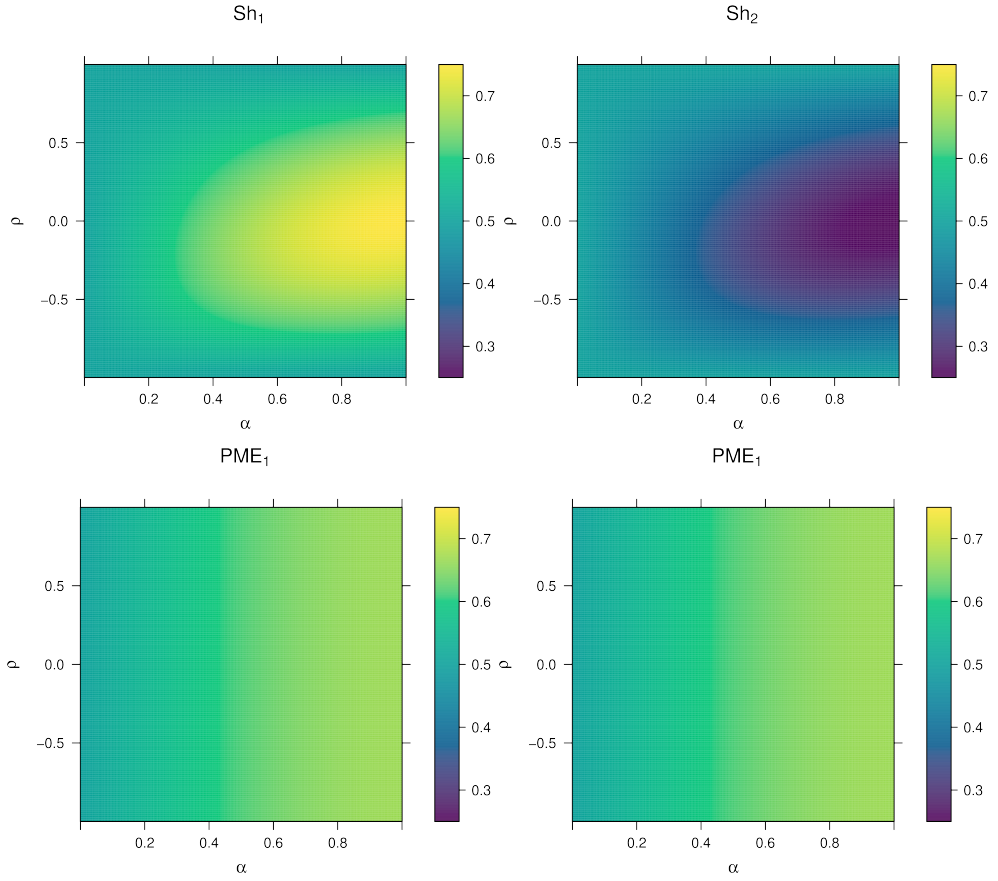

Figure 3.6: PMEs and Shapley effects in the (α, ρ)-plane for the linear model with unbalanced interaction term.

Focusing on the behavior of both effects w.r.t. the interaction, one can first focus on the α-axis of the plots in Figure 3.6. Whenever α is close to 0, one can notice that both indices tend to allocate an equal share of the output variance to both inputs. As α increases, the PME grants an increasing share of the output variance to X_1, independently of ρ. However, on the other hand, the Shapley effects display a sharing mechanism dependent on ρ. When ρ is between -0.5 and 0.5, and α is close to 1, Sh_1 increases, with a maximum allocation of 0.75 taken at $(\rho = 0, \alpha = 1)$, while Sh_2 decreases, with a minimum allocation of 0.25 at the same point on the plane. Additionally, one can notice that when both inputs are highly correlated, the Shapley effects redistribute the output's variance equally, whatever the value of α is.

3.4 Illustration on use-cases

3.4.1 Estimation schemes

Following the two-step methodology presented in [32], the Shapley effects and the PMEs can be estimated in two distinct steps:

- **Step 1:** Estimate the *conditional elements*, i.e., $S_A^T, \forall A \in \mathcal{P}_D$;

- **Step 2:** Perform an *aggregation procedure* via a direct plug-in of the estimated conditional elements.

Only the aggregation procedure differs between the estimation of the PMEs and the Shapley effects. It entails that the estimation cost in terms of model evaluations is the same for the PMEs as for the Shapley effects. Furthermore, both indices can be evaluated "at once" using the same conditional elements estimates. Two situations the practitioner may encounter are taken into account.

Estimating the conditional elements. First, if the practitioner can randomly sample from (i) every possible conditional distribution of the conditional random variables $X_A|X_{\overline{A}}$ and (ii) every marginal distribution, i.e., to simulate i.i.d. observations of X_A, for all $A \in \mathcal{P}(\mathcal{D})$, then the conditional elements can be estimated via a Monte Carlo scheme. This estimation scheme has been studied and proven to yield consistent estimates in [206, 33, 48, 115], and is recalled in Appendix C.1.1. However, it is essential to note that the ability (i) to sample from the conditional distributions can be difficult in practice (especially if the inputs are dependent).

Second, if the practitioner can only access an i.i.d. input-output sample (coming from the joint distribution of the inputs), they can perform a given-data estimation scheme. Such a scheme has been proposed in the literature and relies on approximating the conditional samples using nearest-neighbors [33]. One can refer to [33, 48, 115] for additional theoretical and computational details on this estimation method, which is also briefly recalled in Appendix C.1.2.

In both situations, the practitioner must estimate $2^d - 1$ conditional elements, which is exponential w.r.t. the number of inputs. As stated in [206, 32, 115], some Monte Carlo-inspired methods can require a number of evaluations proportional to $d!(d-1)$, which may be prohibitive for costly numerical models. The given-data procedure avoids the need to simulate and evaluate data, but the sheer number of elements to estimate can render the estimation very long. However, both indices can be estimated with the same set of conditional elements, with the only differentiating factor being the aggregation procedures, which are less computationally expensive in comparison.

Aggregation procedures Coming from Eq. (3.7), the aggregation procedure for the Shapley effects is relatively straightforward (and linear). Recent developments showed that the computational cost of the Harsanyi dividends, i.e., Möbius transforms, can be significantly reduced (regarding machine operations, not model evaluations) [152], once the conditional elements are available.

The aggregation procedure for the PMEs can be broken down as follows. Given estimates of every conditional element, i.e., $\widehat{S_A^T}$ for every $A \subseteq D$, the PME can be computed using its recursive definition in Eq. (3.13).

Ratio potential computation First, recall that for any value function v, $R(\emptyset, v) = 1$ and for any $i \in D$, $R(i, v) = v(\{i\})$. The computation of $R(A, v)$ can be broken down as follows:

1. Let $A \in \mathcal{P}(D)$, $A \neq \emptyset$, $|A| \geq 2$.
2. Compute $v(B)$, for every $B \in \mathcal{P}(A)$.
3. For $m = 1, \ldots, |A| - 1$:
   - ↪ For $B \subseteq A$ such that $|B| = m$:
     - ↪ Compute $R(B, v) = v(B) \left(\sum_{j \in B} R(B_{-j}, v)^{-1} \right)^{-1}$.
4. Compute $R(A, v) = v(A) \left(\sum_{j \in A} R(A_{-j}, v)^{-1} \right)^{-1}$.

Following this algorithm and given conditional element estimates, one can then compute $R\left(A, \widehat{S^T}\right)$ for any $A \in \mathcal{P}(D)$.

Aggregation procedure for PME computation With the ability to compute the ratio potential $R(A, \widehat{S^T})$ for any $A \in \mathcal{P}(D)$ and any set function v, one can proceed to compute the PME. First, define

the function, $\forall A \in \mathcal{P}(D)$:

$$\begin{aligned} \widehat{\zeta_A} : \mathcal{P}(D \setminus A) &\to \mathbb{R}^+ \\ B &\mapsto \widehat{\zeta_A}(B) := \widehat{S^T_{A \cup B}} \end{aligned}$$

The aggregation procedure of the PME can then be broken down as follows:

1. Compute $\widehat{S^T_A}$, for every $A \in \mathcal{P}(D)$.
2. Compute $\mathcal{K} = \underset{A \in \mathcal{P}(D) \text{ s.t. } \widehat{S^T_A}=0}{\operatorname{argmax}} |A|$.
3. For every $A \in \mathcal{K}$, compute $R\left(D \setminus A, \widehat{\zeta_A}\right)$.
4. Let $R_{\mathcal{K}} = \sum_{A \in \mathcal{K}} R\left(D \setminus A, \widehat{\zeta_A}\right)^{-1}$.
5. For $i = 1, \dots, d$:
   (a) If $i \in \underset{A \in \mathcal{K}}{\cap} A$, set $\text{PME}_i = 0$.
   (b) If $i \notin \underset{A \in \mathcal{K}}{\cap} A$:
      i. Compute $\mathcal{K}_{-i} = \left\{A \in \mathcal{K} : i \notin A\right\}$.
      ii. For every $A \in \mathcal{K}_{-i}$, compute $R\left(D_{-i} \setminus A, \widehat{\zeta_A}\right)$.
      iii. Let $\text{PME}_i = \sum_{A \in \mathcal{K}_{-i}} R\left(D_{-i} \setminus A, \widehat{\zeta_A}\right)^{-1} / R_{\mathcal{K}}$.

In the following sections, particular care is put into highlighting which estimation scheme is used and every hyper-parameter related to it. An accompanying GitHub repository[1] containing all the codes used to produce the presented figures is made available as well, for reproducibility purposes.

3.4.2 River water level

Figure 3.7 presents the Shapley effects and PMEs for the river water level model, presented in Section 1.4.1. These allocations have been computed using a scheme (see, Appendix C.1.1), with simulation sample sizes equal to $N_v = 50.000$ for estimating $\mathbb{V}(G(X))$, as well as $N_o = 2000$ and $N_i = 300$ to estimate the total Sobol' indices for every $A \in \mathcal{P}_D$. This experience has been repeated 150 times to compute the empirical mean and the 5% and 95% quantiles. The two allocations have been computed on the same conditional element estimates (i.e., only the aggregation step differs).

One can notice that the importance ranking does not differ between the Shapley effects and the PMEs, with Q and Z_v being granted the majority of the output's variance. In third place, the Strickler coefficient K_s receives around 9.3% of the output variance according to the Shapley values and around 14% according to the PMEs. Finally, the other three inputs, Z_m, L, and B, can be considered negligible. In particular, L receives less than 0.01% of the output's variance according to the PME and, hence, is not considered exogenous (since it is still greater than zero). However, its effect is still considered relatively minimal. This importance quantification is similar to more in-depth studies of this model in the literature [141, 38].

In conclusion, the three most important factors when it comes to studying the variability of the annual maximal water level of a river are the flow rate Q, the downstream river level (Z_v), and the Strickler riverbed roughness coefficient, totaling a combined allocation of around 99% of the output's variance, using either the Shapley effects and the PME. The three remaining coefficients can be considered negligible.

[1] https://github.com/milidris/phdThesis

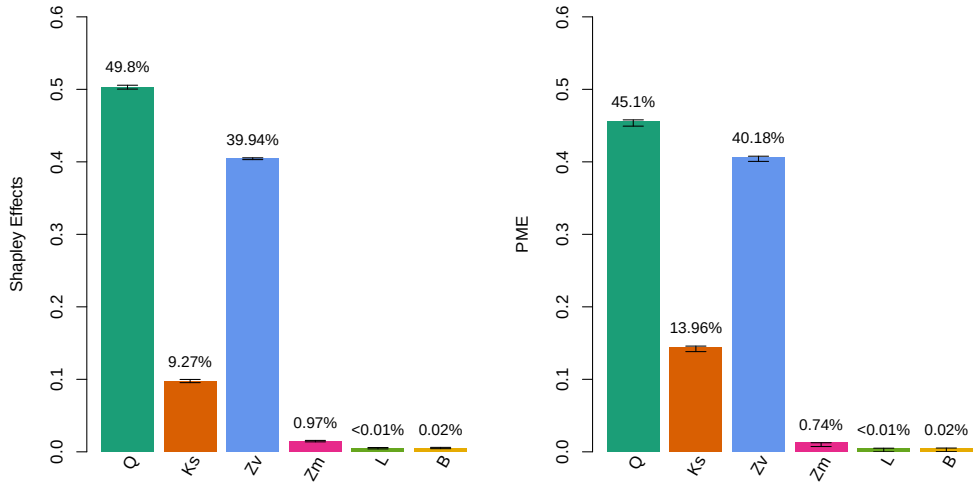

Figure 3.7: Normalized Shapley effects and PMEs for the model of river water level.

Industrial site flooding Suppose that the studied river is located near an industrial site, and a dam of height $t = 54.25\text{m}$ has been placed to prevent floods. The occurrence of a flood can thus be modeled as the random output:

$$T(X) = \mathbb{1}_{\{G(X)>t\}}(X).$$

$T(X)$ follows a Bernoulli distribution. In reliability-oriented sensitivity analysis [37, 175], and more precisely, in *target SA* (see, e.g., [180, 115]), the occurrence probability of $T(X)$ (i.e., its means) is called *the failure probability* and the variance of $T(X)$ is called the *failure variance*. The failure probability has been estimated at around 1%, on 10^7 i.i.d. simulations. The Shapley effects and PME can be computed on Bernoulli outputs (i.e., $T(X)$ remains in $\mathbb{L}^2(\sigma_X)$ as demonstrated in [115]), providing a decomposition of the failure variance, which can be interpreted as importance measures.

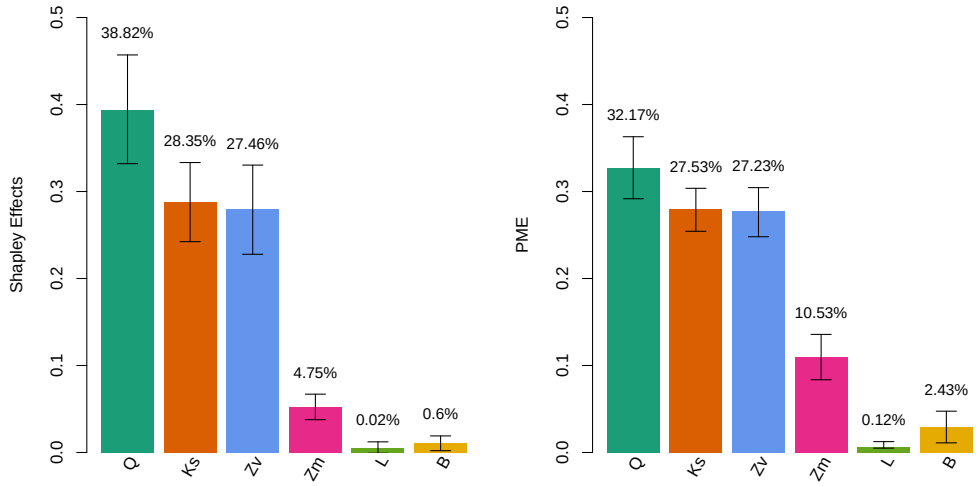

Figure 3.8: Normalized Shapley effects and PMEs for the failure variance of the occurrence of a flood.

The most apparent difference in the importance attribution between the inputs, when compared with Figure 3.7, concerns the Strickler coefficient K_S. Its importance share, initially around 9-14%, is now around 30%. Q and Z_v have a drastically lower share of variance. Z_m and B also are granted an increased share of variance compared to the initial study. These interpretations are expected, especially when taking a more in-depth look at the model in Eq. (1.1) since low values of K_S, Z_m, and B will tend to shoot the height of the river water level up. This interpretation is also similar to the relevant studies from the literature [115].

When studying the river water level variability, only three inputs appeared to be relevant. However, in the case of the flooding occurrence of an industrial site, the Strickler coefficient K_s seems to account for much more of the failure variability of this phenomenon. Additionally, the upstream river level Z_m seems to bear significant importance, 5% of the failure variance according to the Shapley effects, and

around 11% according to the PME. Hence, studying the occurrence of a flood instead of the river water level itself leads to a different interpretation of the model.

3.4.3 Optical filter transmittance

A unique i.i.d. sample of size 1000 of these 13 inputs has been simulated, on which the model's output has been computed. A given-data estimation method is used since this model is pretty expensive to evaluate (here, using the Monte Carlo scheme is not feasible). Hence, the Shapley effects and PMEs are computed using the nearest-neighbor procedure (see, Appendix C.1.2), with an arbitrarily chosen number of neighbors equal to 6, using the `sensitivity` R package.

Figure 3.9 displays the Shapley effects and PMEs estimates. The intervals are the 5% and 95% empirical quantiles computed on 100 estimation repetitions. For each repetition, both indices have been estimated on a random selection of 80% of the initial dataset.

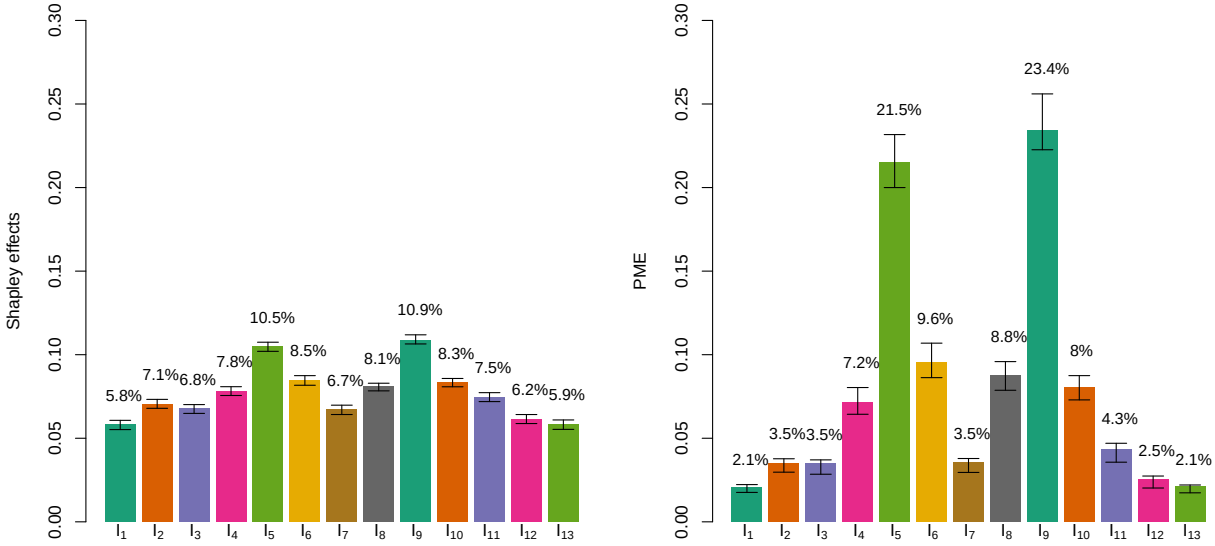

Figure 3.9: Normalized Shapley effects and PMEs estimates using the nearest-neighbor procedure for the interference filter model. The vertical error bars represent the 90% intervals of the estimates.

The Shapley effects of the different inputs vary between 5% and 11%, while the PMEs vary between 2% and 24%. Even if the Shapley effects of I_5 and I_9 are slightly larger than the others, no particular input emerges as predominantly influential, and none emerges as fairly non-influential. However, the PME is more discriminant in the influence repartition. I_5 and I_9 stand out as very influential, I_4, I_6, I_8 and I_{10} seem to bear some importance, while I_1, I_2, I_3, I_7, I_{11}, I_{12} and I_{13} can be considered as non-influential.

This more pronounced discriminating power can be explained by the difference in the redistribution process of the PMEs and the Shapley effects, especially in this case where the inputs are highly correlated. It highlights the more discriminatory ability of the PMEs for influence ranking in situations of highly correlated inputs, where the Shapley effects tend to equalize the influence between the inputs in this situation.

Surrogate modeling and feature selection The PME values of non-influential inputs are not worth zero but are relatively close to zero (the PMEs of I_1 and I_{13} are smaller than 2%, and the PMEs of I_2, I_3, I_7, I_{11}, and I_{12} are smaller than 3%). However, as the nearest-neighbor procedure used to estimate the PME is known to have a bias, we cannot infer the non-exogeneity of these inputs. Models with subsets of inputs have been trained to verify the presence of spurious inputs.

The predictive capabilities of three different Gaussian process (GP) surrogate models [192] are compared. For each model, dimension reduction is performed by selecting subsets of inputs according to the previously discussed importance rankings:

- **GP_1** - The inputs are selected with a 5% importance threshold applied on the Shapley effects: the 13 inputs are kept in the GP. Then, this GP corresponds to the one without dimension reduction;

- **GP_2** - The inputs are selected with a 5%-threshold applied on the PMEs: only 6 inputs (I_4, I_5, I_6, I_8, I_9 and I_{10}) are kept to train the GP;
- **GP_3** - The inputs are selected with a 2.2%-threshold applied on the PMEs: 2 inputs (I_1 and I_{13}) are removed from the initial 11 to train the GP.

The three surrogate models are trained on the initial 1000 observations and are parameterized by a constant trend and a $5/2$-Matérn covariance kernel. The parameters have been estimated using a maximum likelihood scheme utilizing the `DiceKriging` R package [188].

To measure the predictive power of the models, their "predictivity coefficients" (i.e., the Q^2-metric, see, e.g., [67]) are computed and displayed in Table 3.1. Removing the two inputs with the lowest PMEs has a negligible impact on the model predictivity (shortfall in Q^2 of less than 0.4%), and removing the seven inputs with the lowest PMEs has a minor impact on the model predictivity (shortfall in Q^2 of less than 1%).

Model	Number of inputs	Selection Threshold	Q^2
$\mathbf{GP_1}$	13	Shapley Effects - 5%	99.48%
$\mathbf{GP_2}$	6	PMEs - 5 %	98.79%
$\mathbf{GP_3}$	11	PMEs - 2.2%	99.14%

Table 3.1: Predictivity coefficient of the three GP surrogate models.

This use case illustrates the PMEs' usefulness in variable selection with highly correlated inputs for dimension reduction and surrogate modeling purposes. Overall, the PMEs favor the already influential inputs at the expense of the correlated inputs, while the Shapley effects equalize importance amongst them. Combined with the ability to detect exogenous inputs, it makes the PME particularly suitable for screening purposes.

3.4.4 Acoustic fire extinguisher

The allocations have been computed on the predictions provided by the trained neural network (presented in Section 1.4.3) on the whole dataset. The nearest-neighbor estimation scheme is used, with an arbitrarily chosen number of neighbors equal to 6. The results are presented in Figure 3.10, and both allocations have been computed on the same conditional elements. One can notice that the importance

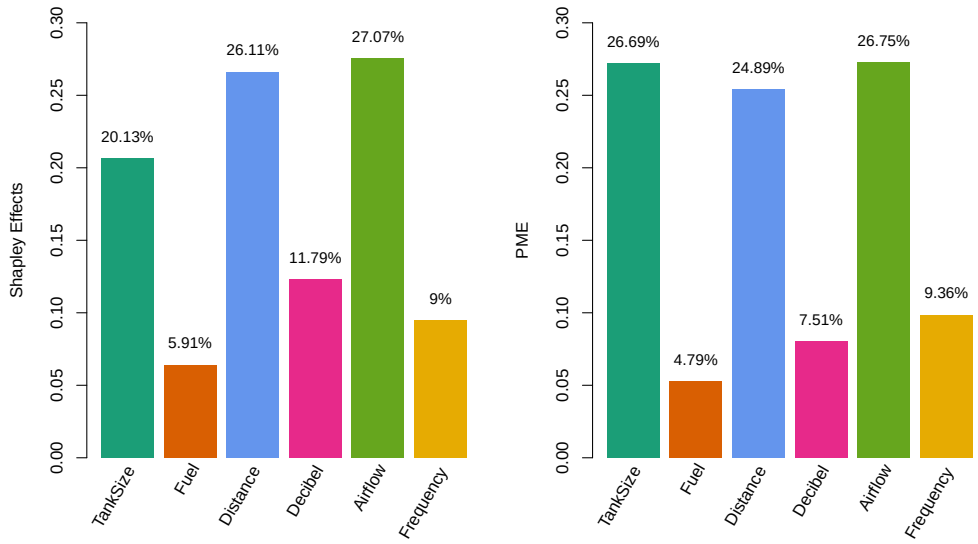

Figure 3.10: Normalized Shapley effects and PMEs estimates using the nearest-neighbor procedure the acoustic fire extinguisher model.

ranking provided by the two allocations is fundamentally different. The most important change is due to the tank size and the decibel value. The former is granted more of the output's variance due to the proportional redistribution than the egalitarian one, while the latter sees a decrease in its allocated importance. Additionally, no inputs seem to be detected as exogenous or negligible: all of them seem to

bear some importance for the neural network's propagated uncertainty. Finally, one can notice that in both cases, the airflow remains the most important input.

Thus, according to this neural network classifier, the three most important parameters for predicting an effective termination of a fire are the airflow coming out of the acoustic fire extinguisher, the distance to the active flame, and the fuel tank size. These three inputs account for around 73% of the neural network's variance on the whole dataset according to the Shapley effects and 78% according to the PME. The fuel type only accounts for around 5% according to both allocations, and thus, its effect can be considered limited but not entirely negligible. While the frequency's importance seems consistent between the two allocations, a significant difference can be spotted in the decibels: the Shapley effects grant around 12% of importance. In comparison, the PME only grants them around 8%.

3.5 The fundamental problem of the input-centric approach

In a nutshell, the analogy between players of a cooperative game and inputs of a random output is fruitful when defining attribution methods of several QoIs using allocations from the literature. In particular, the allocations in the Harsanyi set (which includes the Shapley values and the PMEs) can be directly linked with the input-centric approach to defining influence measures:

1. Choose a value function v;
2. Define its dividends $\mathcal{D}_v$ as the Möbius transform of v. The dividends are then coalitional decompositions of $v(D)$;
3. Aggregate the dividends to have an attribution of $v(D)$.

This last step in the definition of allocations in the Harsanyi set is what is added to the rationale described in Section 2.2.4. Hence, this last step is paramount to characterizing allocations. From this point of view, several remarks can be brought forward.

Choice of value function As highlighted in Remark 2.3, the chosen value measure (which plays the same role as the value function) only needs to exist to define an input-centric coalitional decomposition. However, as brought forward in Section 2.3, in the case of variance decomposition, the final interpretation of such influence measures remains misunderstood, *unless the inputs are mutually independent*.

Hence, the aggregation of the Sobol' indices S, built on the choice of S^{clos} as a value function/measure, can be seen as *an aggregation of (well-defined, but) poorly understood quantities*. This observation affects the ability to provide a clear interpretation of the resultant attributions whenever the inputs are not mutually independent, which is the primary purpose of these indices. Said differently, the analogy of importance quantification and cooperative games *did not answer any of the questions brought forward in Section 2.3.3*, due to the choice of value function being ill-chosen w.r.t. the dependence structure of the inputs.

Hence, while the analogy between players and inputs may seem like a good idea to define variance attribution for dependent inputs, the overall question of the choice of value function remains central. In the recent literature, many attempts have been made to study cooperative games with different value functions (see, e.g., [198]), supposedly aimed at quantifying different aspects of the influence of the inputs on a QoI. However, the choice of value function remains arbitrary and is subject to the same critiques formulated above.

Furthermore, the overall philosophy behind cooperative games is to place the players in a central position in the development. When it comes to Sobol' cooperative games, this can be understood as follows: the choice of the value function allows to define *an* input-centric decomposition of $\mathbb{V}(G(X))$ *independently of whether* $G(X)$ *can be decomposed*. Hence, the inherent decomposition of $G(X)$ is wholly disregarded and is the root of the lack of interpretation of input-centric importance measures. The main argument brought forward in favor of the choice of S^{clos} is the fact that it quantifies the variance of the best approximation of $G(X)$ as a function of the subspaces $(\mathbb{L}^2(\sigma_A))_{A\subseteq D}$, which can be understood as a value *for the input*, but completely disregarding the random output $G(X)$.

Axioms The axioms are often considered "mathematical arguments to justify the theoretical nature of the importance measure". While their role is central to cooperative game theory to seek uniqueness,

for defining influence measures on the Harsanyi set (including the Shapley values), these axioms can be understood as properties on the *aggregation process of dividends*. However, the previous remark still holds: if the choice of value function leads to misunderstood influence measures, then allocations uniquely characterized through axioms can be considered as *theoretically-grounded aggregations of misunderstood quantities*.

Hence, since these axioms only affect the aggregation step in the definition of allocations, their weight in justifying the whole method remains somewhat limited. In the recent literature, many papers criticize the axiomatic justification of the Shapley values by showing their limited relevance for importance quantification (see, e.g., [218]).

Multiplicities of importance measures Since the allocations can be understood as aggregations of dividends, many different allocations can easily be defined, by choosing an aggregation step, resulting in different importance attributions, where each one of them can be studied in order to find interesting behavior (e.g., the exogeneity detection of the PMEs). However, this multiplicity of allocations begs the question: *which ones are suitable for quantifying importance?*

This chapter highlights that the Shapley effects and the PMEs can be widely different, even if computed on the same data. The resulting practical interpretation can thus change drastically based on the choice of method. But, *which attribution is the "right" one?* A path towards studying multiple allocations, and for each one, considering multiple value functions seems like an arduous and infinite task.

Considering that the Shapley effects and the PME are two different methods is subject to confusion. Allocations, as aggregations of dividends, should be treated as such: they summarize the information (in different ways) contained in the dividends. Moreover, since, for influence quantification purposes, the input-centric approach to defining dividends has been proven to lead to misunderstood influence measures, the central question circles back to the choice of the value function.

Conclusion Defining a suitable value function is paramount in the in-fine interpretation of attribution methods based on allocations. For importance quantification, as highlighted in Remark 2.4, the choice of S^{clos} as a value measure does coincide with a model-centric approach to defining importance measures *whenever the inputs are mutually independent*. In this particular case, the Sobol' indices (which are the Harsanyi dividends of the Sobol' cooperative game) can be interpreted as shares of variances due to functional interactions induced by the model between the inputs. The resulting allocations can be easily understood as particular aggregations of these interaction effects. Hence, the choice of value function *is ultimately bound to the probabilistic structure of the inputs*, and especially to *their dependence structure*.

Understanding the problem from this point of view motivates exploring the *model-centric approach* to defining influence measures, which first requires decomposing $G(X)$ itself. Thanks to Corollary 2.1, the resulting value measures would be prime candidates as a suitable value function. The following chapter is dedicated to this problem.

Chapter 4

Model-centric approach and output decomposition

Contents

4.1 Introduction **55**
4.2 Preliminaries **56**
4.2.1 The Lebesgue space $\mathbb{L}^2(\sigma_X)$ and its subspaces 56
4.2.2 Angles between closed subspaces of a Hilbert space 57
4.2.3 Direct sums, complemented subspaces and projections 60
4.3 Coalitional output decomposition with dependent inputs **62**
4.3.1 Two reasonable assumptions 63
4.3.2 Output decomposition and geometric interpretation 64
4.3.3 Mutual independence and Hoeffding's decomposition 68
4.4 Model-centric influence measures **70**
4.4.1 Orthocanonical evaluation decomposition 70
4.4.2 Variance decomposition 71
4.5 Analytical example: two Bernoulli inputs **74**
4.5.1 Orthocanonical decomposition as solving equations 74
4.5.2 Angle, comonotonicity and definite positiveness of Δ 75
4.6 Conclusion **75**

Abstract (English).

The first step in defining model-centric coalitional decomposition is the ability to uniquely decompose square-integrable functions of non-mutually independent random inputs into a sum of functions of every possible subset of variables. The well-known Hoeffding decomposition allows achieving such a task whenever the inputs are mutually independent. However, no such result has been achieved whenever the inputs are not mutually independent, except under very restrictive assumptions. A novel view on this problem is proposed, linking three domains of mathematics: probability theory, functional analysis, and combinatorics. The problem of random output decomposition can be seen as trying to express a direct-sum decomposition of Lebesgue spaces of functions of the inputs. Under two reasonable assumptions on the inputs (non-perfect functional dependence and non-degenerate stochastic dependence), it is always possible to uniquely decompose such functions. This "orthocanonical decomposition" is intuitive and unveils the linear nature of non-linear functions of non-linearly dependent inputs, effectively generalizing Hoeffding's pioneering result. They can be expressed using oblique projections and enable the definition of intuitive and interpretable model-centric coalitional decompositions of quantities of interest. This result offers a path towards a more precise uncertainty quantification, which can benefit sensitivity analyses and interpretability studies whenever the inputs are dependent. This decomposition is illustrated analytically, and the challenges to adopting these results in practice are discussed.

Abstract (Français).

La première étape dans la définition de la décomposition coalitionnelle centrée sur le modèle est la capacité à décomposer une sortie aléatoire, dont les entrées ne sont pas forcément mutuellement indépendantes. La décomposition bien connue d'Hoeffding permet d'accomplir cette tâche lorsque les variables d'entrée sont mutuellement indépendantes. Cependant, aucun résultat de ce type n'a été obtenu lorsque les entrées dépendent les unes des autres, sauf sous certaines hypothèses très restrictives. Une nouvelle perspective sur ce problème est présentée, établissant un lien entre trois domaines des mathématiques : la théorie des probabilités, l'analyse fonctionnelle et la combinatoire. La décomposition de sortie aléatoire peut être vu comme une décomposition en somme directe des espaces de Lebesgue de fonctions des entrées. Sous deux hypothèses raisonnables sur les entrées (dépendance fonctionnelle non parfaite et dépendance stochastique non dégénérée), il est toujours possible de décomposer de manière unique de telles fonctions. Cette "décomposition orthocanonique" est intuitive et révèle une nature linéaire des fonctions non linéaires, dont les entrées peuvent être non-linéairement dépendantes, généralisant ainsi de manière efficace le résultat pionnier de Hoeffding. Elles peuvent être exprimées au moyen de projections obliques et permettent de définir des décompositions coalitionnelles de quantités d'intérêt centrées sur le modèle intuitives et interprétables. Ce résultat ouvre la voie à une quantification de l'incertitude plus précise, qui pourrait bénéficier aux analyses de sensibilité et aux études d'interpretabilite avec entrées dépendantes. Cette décomposition est illustrée de manière analytique, et les défis liés à l'adoption de ces résultats en pratique sont discutés.

Keywords . Hoeffding decomposition • Functional analysis • Angles between Hilbert spaces • Orthogonal and oblique projection • Direct-sums

This chapter expands on the following contribution.

Pre-prints:
M. Il Idrissi, N. Bousquet, F. Gamboa, B. Iooss, and J. M. Loubes. Hoeffding decomposition of black-box models with dependent inputs. Preprint, 2023. URL: https://hal.science/hal-04233915

Conference:
M. Il Idrissi, N. Bousquet, F. Gamboa, B. Iooss, and J. M. Loubes. Hoeffding decomposition, revisited. In *SIAM Conference on Uncertainty Quantification 2024*, Trieste, Italy, 2024. URL: https://www.siam.org/conferences/cm/conference/uq24

4.1 Introduction

As shown in Chapter 2, there are two main ways to define influence measures. The input-centric approach, highlighted in Chapter 3, offers tools to define influence measures for dependent inputs mechanically. However, as highlighted in Section 3.5, this approach revolves around the arbitrary choice of a value measure to quantify a subset of inputs' value. The resulting interpretation of the influence measure is ultimately related to the dependence structure of the inputs.

In this chapter, the focus is put on the *model-centric approach*. The main idea is to find a suitable *coalitional decomposition* of $G(X)$ and then define indices based on it. In other words, one wishes to obtain a so-called *high-dimensional model representations* (HDMR) [179]. Formally, for random inputs $X = (X_1, \dots, X_d)^\top$, and a random output $G(X)$, it amounts to finding the decomposition

$$G(X) = \sum_{A \in \mathcal{P}_D} G_A(X_A), \tag{4.1}$$

where each $G_A(X_A) \in \mathcal{G}_A$ is a representant (see, Definition 2.2) of input $X_A = (X_i)_{i \in A}$. Whenever the X_i are assumed to be mutually independent, such a decomposition is known as *Hoeffding's decomposition* (see, Theorem 2.1) [102]. In the literature, the proposed influence measures (or methods based on influence measures) usually assume mutual independence of the inputs [204, 149], either for the simplicity of the resulting estimation schemes or for the lack of a proper framework. However, the inputs are often endowed with a dependence structure intrinsic to the (observed or modeled) studied phenomena. Always assuming mutual independence can be seen as expedient and can lead to improper insights [96]. One of the main challenges to a better understanding of black-box models is to consider this dependence structure [182] and, above all, to formally justify the proposed methods without heavily relying on empirical observations or specific benchmarks.

Whenever the inputs are not assumed to be mutually independent, many approaches have been proposed in the literature. Notably, [96] proposed an approximation theoretic framework to address the problem and provide useful tools for importance quantification. However, they lack a proper and intuitive understanding of the estimated quantities. In [38], the authors approached the problem differently and brought forward an intuitive view on the subject, but under somewhat restrictive assumptions on the probabilistic structure of the inputs. In [105] and [134] proposed a projection-based approach under constraints derived from desirability criteria. However, most of these proposals do not offer a completely satisfactory and unequivocal answer to the interpretation of the resulting influence measure. Other approaches rely on a transformation of the dependent inputs to achieve mutual independence using, for instance, Nataf or Rosenblatt transforms [138, 137, 151], offering meaningful indications on the relationship between X and $G(X)$. While these approaches can be applied to a broad range of probabilistic structures, they can be seen as lacking in generality (e.g., existence of probability density functions, being in an elliptical family of distribution, restricted to $\mathbb{R}^d$ valued inputs).

This chapter is dedicated to exploring and studying the problem of random output decomposition (i.e., as in Eq. (4.1)) whenever the inputs are not assumed to be mutually independent. To that extent, a different point of view is adopted at the crossroads of probability theory, functional analysis, and abstract algebra. This point of view allows seeing HDMRs as a *direct-sum decomposition* of the Lebesgue space $\mathbb{L}^2(\sigma_X)$. It is shown that such HDMRs hold under two reasonable assumptions on the inputs:

1. Non-perfect functional dependence (i.e., the inputs cannot be functions of each other);

2. Non-degenerate stochastic dependence (i.e., there cannot be a perfect stochastic dependence between the inputs).

Under these two assumptions, the subspaces of the Lebesgue space $\mathbb{L}^2(\sigma_X)$ (see, Section 1.2), where σ_X is the σ-algebra generated by the inputs X,(see, Definition 2.4) involved in the direct-sum decomposition can be characterized, and lead to a geometric understanding of importance quantification. In addition, Hoeffding's decomposition can be seen as a very particular case of this more general result. The definition of inherently interpretable influence measures is discussed for evaluation decomposition, as well as importance quantification. Finally, the proposed indices are studied analytically on a particular toy-case of a black-box model of two Bernoulli random inputs.

4.2 Preliminaries

In this section, some preliminaries are introduced to justify that the HDMR of random outputs in $\mathbb{L}^2(\sigma_X)$ can be seen as a direct-sum decomposition problem. The framework presented in Section 1.2 is restricted according to the following assumptions.

Special case . In this chapter, the following is assumed concerning the random output:

- The output space $Y = \mathbb{R}$, i.e., the model is $\mathbb{R}$-valued;
- The space of random outputs $\mathcal{G}_X$ is restricted to $\mathbb{L}^2(\sigma_X)$, i.e., square-integrable random variables measurable w.r.t. the σ-algebra generated by the inputs X (see, Section 1.2).

When it comes to a vector of random elements, especially when the very particular case of mutual independence is not assumed, one needs to restrict the inputs explicitly to avoid trivial situations (e.g., , constant a.s. inputs or redundancy). In the case of the framework presented in Section 1.2, two standard assumptions are assumed to hold, and the results presented in this chapter are to be understood in this context:

1. For every $i \in D$, $\sigma_\emptyset \subset \sigma_i$, i.e., the $\mathbb{P}$-trivial σ-algebra (see, Definition A.6) is strictly contained in the σ-algebras generated by individual inputs. In other words, this entails that *none of the inputs are constant a.s.;*
2. For every $A, B \in \mathcal{P}_D$ such that $B \subset A$, $\sigma_B \subset \sigma_A$, i.e., the σ-algebra generated by a subset of inputs is necessarily strictly contained in the σ-algebras generated by a bigger subset of inputs. In other words, *adding an input to a subset of inputs necessarily adds "more information"*.

The first assumption is standard in many probabilistic theoretical frameworks (see, e.g., [203]). The last assumption is, however, less standard but remains reasonable. For instance, consider an example with three inputs. The fact that $\sigma_1 \subseteq \sigma_{12}$ comes naturally from the definition of generated σ-algebras. However, if one lets $\sigma_1 = \sigma_{12}$, that would entail that *every* σ_{12}-measurable function $f(X_1, X_2)$ can in fact be written as a function of only X_1, hence making the subset (X_1, X_2) "redundant" w.r.t. to X_1. This assumption is automatically respected whenever the inputs are mutually independent but does not necessarily hold in general.

4.2.1 The Lebesgue space $\mathbb{L}^2(\sigma_X)$ and its subspaces

The Lebesgue spaces of square-integrable random variables, as defined in Definition 2.4, show some intrinsic properties concerning the sub-σ-algebras they are defined on. Two of these classical results, which are of interest, are recalled.

Theorem 4.1. *For two sub σ-algebras $\mathcal{B}_1$ and $\mathcal{B}_2$ of $\mathcal{F}$, the following assertions hold.*

1. *If $\mathcal{B}_1 \subseteq \mathcal{B}_2$, then $\mathbb{L}^2(\mathcal{B}_1) \subseteq \mathbb{L}^2(\mathcal{B}_2)$.*
2. *$\mathbb{L}^2(\mathcal{B}_1) \cap \mathbb{L}^2(\mathcal{B}_2) = \mathbb{L}^2(\mathcal{B}_1 \cap \mathcal{B}_2)$.*

Proof: See, [203], Theorem 2.

When interpreting σ-algebras as "information", this result is rather intuitive: the set of random variables with less information is necessarily contained in the set of random variables with more information. Additionally, Theorem 4.1 shows that random variables in the intersection of two Lebesgue spaces are necessarily measurable w.r.t. to the two generating σ-algebras, which is also intuitive.

Hence, when it comes to $\mathbb{L}^2(\sigma_X)$, Theorem 4.1 implies that the set of subspaces $\left\{\mathbb{L}^2(\sigma_A)\right\}_{A \in \mathcal{P}_D}$ display some ordering structure w.r.t. the inclusion binary relation.

Lemma 4.1. *Let $A, B \in \mathcal{P}_D$, such that $B \subseteq A$. Then*

$$\mathbb{L}^2(\sigma_B) \subseteq \mathbb{L}^2(\sigma_A).$$

Notice that, by definition, σ_A is a σ-algebra that contains $\cup_{i\in B}\sigma_i$ since $B \subseteq A$. Since σ_B is the smallest σ-algebra containing $\cup_{i\in B}\sigma_i$, then necessarily $\sigma_B \subseteq \sigma_A$. Applying in turn Theorem 4.1 (1.) leads to the result.

Lemma 4.1 can be understood as the fact that the Lebesgue space of *random outputs* w.r.t. a subset of inputs X_B *is included* in the Lebesgue space of *random outputs* w.r.t. a bigger subset of inputs X_A (i.e., provided $B \subseteq A$). For instance, for two inputs $X = (X_1, X_2)$, Lemma 4.1 entails that the set of random outputs which are only functions of X_1 is included in the set of random outputs which are functions of both X_1 *and* X_2. This behavior is rather intuitive due to the intrinsic definition of measurability, i.e., a random element measurable w.r.t. σ_B is necessarily measurable w.r.t. σ_A provided $\sigma_B \subseteq \sigma_A$. Additionally, notice that $\mathbb{L}^2(\sigma_\emptyset)$ is necessarily comprised of constant a.s. random variables, thanks to Lemma A.1.

Remark 4.1. The space $\mathbb{L}^2(\sigma_X)$ thus contains each element of the set of subspaces $\left\{\mathbb{L}^2(\sigma_A)\right\}_{A\in\mathcal{P}_D}$, where, $\sigma_D := \sigma_X$. In addition, thanks to Lemma 4.1, one can notice that $\left\{\mathbb{L}^2(\sigma_A)\right\}_{A\in\mathcal{P}_D}$ is endowed with some algebraic structure w.r.t. to the inclusion operator. In fact, **the order is preserved between** $(\mathcal{P}_D, \subseteq)$ and $\left(\left\{\mathbb{L}^2(\sigma_A)\right\}_{A\in\mathcal{P}_D}, \subseteq\right)$, algebraically speaking.

A real Banach space is a complete normed space, usually defined as a tuple $(\mathcal{M}, \|\cdot\|)$, where $\mathcal{M}$ is a vector space over the reals (or more generally, over a field) and $\|\cdot\| : \mathcal{M} \to \mathbb{R}$ is a norm, with the added property that the limit of every converging sequence of elements of $\mathcal{M}$ (i.e., Cauchy sequences) is in $\mathcal{M}$ itself. Whenever the norm $\|\cdot\|$ stems from an inner product $\langle\cdot,\cdot\rangle$, the resulting space is called a *Hilbert space* (see, e.g., [45], Definition 1.6). Hence, every Hilbert space is a Banach space [195].

Regarding Lebesgue spaces of square-integrable random variables, they are, in fact, Hilbert space.

Theorem 4.2. *Let $\mathcal{B} \subseteq \mathcal{F}$ be a sub-σ-algebra. The Lebesgue space $\mathbb{L}^2(\mathcal{B})$ is then a Hilbert space with inner product defined, for any $Z_1, Z_2 \in \mathbb{L}^2(\sigma_B)$, as:*

$$\mathbb{E}[XY] = \int_\Omega Z_1(\omega)Z_2(\omega)d\mathbb{P}(\omega).$$

Proof: See, [150] Theorem 9.4.1.

Hence, the set of Lebesgue spaces $\left\{\mathbb{L}^2(\sigma_A)\right\}_{\mathcal{P}_D}$ is comprised of Hilbert spaces, and more notably, for any $A \subset D$, since $\mathbb{L}^2(\sigma_A)$ is a Hilbert space, it can be seen as a *closed subspace of* $\mathbb{L}^2(\sigma_X)$ (since it is complete). Additionally, each of these Hilbert spaces is *infinite-dimensional* since no further restriction is put on the input space $(E, \mathcal{E})$. When studying the *relationships* between closed subspaces of an infinite-dimensional Hilbert space, the notions of *angles between subspaces* offer relevant tools. They are the main topic of the next section.

4.2.2 Angles between closed subspaces of a Hilbert space

As highlighted in the previous section, the subspaces $\left\{\mathbb{L}^2(\sigma_A)\right\}_{A\in\mathcal{P}_D}$ of $\mathbb{L}^2(\sigma_X)$ present a particular algebraic form, which is related to the power-set. The next step is exploring the *relationships* between these subspaces. For instance, one can wonder if (or rather *when*) these subspaces are all orthogonal. To that extent, the notion of *angles between subspaces of Hilbert spaces* has been introduced in functional analysis. In particular, two angles introduced by Dixmier and Friedrichs are of interest to decompose outputs. These angles, initially introduced for the general analysis of abstract infinite-dimensional Hilbert spaces, have also been introduced in probability theory, as highlighted in the remainder of this section.

Dixmier's angle Dixmier's angle [60] can be understood as the *minimal angle* between two closed subspaces of a Hilbert space. Its cosine is defined as follows.

Definition 4.1 (Dixmier's angle). Let H and K be closed subspaces of a Hilbert space $\mathcal{H}$ with inner product $\langle\cdot,\cdot\rangle$ and norm $\|\cdot\|$. The cosine of Dixmier's angle is defined as

$$c_0(H, K) := \sup\{|\langle x, y\rangle| : x \in H, \|x\| \leq 1, \quad y \in K, \|y\| \leq 1\}.$$

Loosely speaking, Dixmier's angle can be understood as the smallest angle between two elements of the two closed subspaces (or limits of converging sequences of these elements). In probability theory, when applied to two generated Lebesgue spaces, this angle is directly linked to the notion of *maximal correlation* between random elements, as a dependence measure between random elements (e.g., vectors) [87].

Definition 4.2 (Maximal correlation). Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let Z_1 and Z_2 be two random elements, and denote σ_{Z_1} and σ_{Z_2} their generated σ-algebra. The *maximal correlation* between Z_1 and Z_2 is Dixmier's angle between $\mathbb{L}^2(\sigma_{Z_1})$ and $\mathbb{L}^2(\sigma_{Z_2})$, i.e., $c_0\left(\mathbb{L}^2(\sigma_{Z_1}), \mathbb{L}^2(\sigma_{Z_2})\right)$.

The maximal correlation has been extensively studied as a dependence measure (see, e.g., [183, 129, 55, 50]), or as a means to quantify the dependence between generated σ-algebras for studying the mixing properties of stochastic processes [61].

The maximal correlation is particularly suitable for studying the independence of random elements. For instance, let Z_1 and Z_2 be two random elements, and let $\mathbb{L}_0^2(\sigma_{Z_1})$ and $\mathbb{L}_0^2(\sigma_{Z_2})$ be their respective induced Lebesgue space of *centered random variables*. Then, the following equivalence holds:

$$c_0\left(\mathbb{L}_0^2(\sigma_{Z_1}), \mathbb{L}_0^2(\sigma_{Z_2})\right) = 0 \iff \mathbb{L}_0^2(\sigma_{Z_1}) \perp \mathbb{L}_0^2(\sigma_{Z_2}) \iff Z_1 \perp\!\!\!\perp Z_2,$$

where the independence is to be understood w.r.t. $\mathbb{P}$. In other words, Z_1 and Z_2 are independent *if and only if* $\mathbb{L}^2(\sigma_{Z_1})$ and $\mathbb{L}^2(\sigma_{Z_2})$ are orthogonal, which happens *if and only if* the maximal correlation between Z_1 and Z_2 is equal to zero (see, [150], Chapter 3).

Friedrichs' angle Friedrichs' angle [79] differs from Dixmier's angle in one way: the supremum is taken outside the intersection of the two subspaces. It is defined as follows.

Definition 4.3 (Friedrichs' angle). Let H and K be closed subspaces of a Hilbert space $\mathcal{H}$ with inner product $\langle \cdot, \cdot \rangle$ and norm $\|\cdot\|$. The cosine of Friedrichs' angle is defined as

$$c(H, K) := \sup \left\{ |\langle x, y \rangle| : \begin{cases} x \in H \cap (H \cap K)^{\perp}, \|x\| \leq 1 \\ y \in K \cap (H \cap K)^{\perp}, \|y\| \leq 1 \end{cases} \right\},$$

where the orthogonal complement is taken w.r.t. to $\mathcal{H}$.

In probability theory, this quantity is known as the maximal partial (or relative) correlation [35, 36, 51] between two random elements.

Definition 4.4 (Maximal partial correlation). Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let Z_1 and Z_2 be two *random elements*, and denote σ_{Z_1} and σ_{Z_2} their generated σ-algebra. The *maximal partial correlation* between Z_1 and Z_2 is Friedrichs' angle between $\mathbb{L}^2(\sigma_{Z_1})$ and $\mathbb{L}^2(\sigma_{Z_2})$, i.e., $c\left(\mathbb{L}^2(\sigma_{Z_1}), \mathbb{L}^2(\sigma_{Z_2})\right)$.

The maximal partial correlation is suitable for deciphering conditional independence between σ-algebras generated by random elements and whether the conditional expectations w.r.t. to those σ-algebras commute. For a sub-sigma algebra $\mathcal{G} \subset \mathcal{F}$, denote $\mathbb{E}_{\mathcal{G}}$ the conditional expectation operator w.r.t. $\mathcal{G}$ and $\underset{\mathcal{G}}{\perp\!\!\!\perp}$ denotes the conditional independence relation w.r.t. $\mathcal{G}$ (see, [126], Chapter 8). One then has the following equivalence

$$c\left(\mathbb{L}^2(\sigma_{Z_1}), \mathbb{L}^2(\sigma_{Z_2})\right) = 0 \iff \sigma_{Z_1} \underset{\sigma_{Z_1} \cap \sigma_{Z_2}}{\perp\!\!\!\perp} \sigma_{Z_2} \iff \mathbb{E}_{\sigma_{Z_1}} \circ \mathbb{E}_{\sigma_{Z_2}} = \mathbb{E}_{\sigma_{Z_2}} \circ \mathbb{E}_{\sigma_{Z_2}} = \mathbb{E}_{\sigma_{Z_1} \cap \sigma_{Z_2}}, \quad (4.2)$$

In other words, the conditional expectations w.r.t. Z_1 and Z_2 commute, *if and only if* the maximal partial correlation between Z_1 and Z_2 is equal to zero (see, [126], Theorems 8.13 and 8.14).

Remark 4.2. In the remainder of this chapter, any reference to Friedrichs' or Dixmier's angle refers to the *cosine of the angle* (in $[0, 1]$) instead of the angle itself (in $[0, \pi/2]$).

Properties of Friedrichs' and Diximier's angles Outside of their intrinsic links with the notions of independence and conditional independence, these angles are better known in the functional analysis literature as tools to assess if the sum of closed subspaces of Hilbert spaces is closed. Some properties

relevant to proving our result are presented. The interested reader is referred to [59] for a more complete overview.

Theorem 4.3 (Properties of Dixmier's angle). *Let H, K be closed subspaces of a Hilbert space $\mathcal{H}$. Then, one has that $0 \leq c_0(H,K) = c_0(K,H) \leq 1$, and for any $x \in H$, and $y \in K$:*

$$|\langle x, y \rangle| \leq c_0(H,K) \|x\| \|y\|, \tag{4.3}$$

and for a proper closed subspace $\widetilde{H} \subset H$,

$$c_0\left(\widetilde{H}, K\right) \leq c_0(H,K).$$

Moreover, the following statements are equivalent.

1. $c_0(H,K) < 1$;
2. $H \cap K = \{0\}$ *and* $H + K$ *is closed in* $\mathcal{H}$.

Proof: See, [59], Lemmas 2.3 and 2.10, Theorem 2.12.

The previous result can be understood as follows. First, Dixmier's angle allows *to sharpen the Cauchy-Schwarz inequality* thanks to Eq. (4.3). For probabilistic considerations, this entails that the minimal angle between Lebesgue spaces (i.e., the maximal correlation) allows controlling the magnitude of the covariances between their elements. Moreover, whenever the angle is strictly less than 1, the two subspaces of interest are *in a direct-sum*, and their sum is *closed*, which is central to the following developments.

Theorem 4.4 (Properties of Friedrichs' angle). *Let H, K be closed subspaces of a Hilbert space $\mathcal{H}$. Then, one has that*

$$0 \leq c(H,K) = c(K,H) \leq 1.$$

Notice that if $H \subseteq K$, then $c(H,K) = 0$. Moreover, the following statements are equivalent.

1. $c(H,K) < 1$;
2. $H + K$ *is closed in* $\mathcal{H}$.

Proof: See, [59], Lemmas 2.3 and 2.10, Theorem 2.13.

For the purposes of this manuscript, one property of Friedrichs' angle is of interest: whenever its value is strictly less than 1, the sum of the two subspaces is *closed*. The closure of sums of subspaces plays a central part in the following developments. Moreover, these two angles are related, as the following result highlights.

Lemma 4.2 (Relation between the two angles). *Let H, K be closed subspaces of a Hilbert space $\mathcal{H}$. Then, one has that*

$$0 \leq c(H,K) \leq c_0(H,K) \leq 1.$$

Moreover, the following equality holds

$$c(H,K) = c_0\left(H \cap (H \cap K)^{\perp}, K\right) = c_0\left(H, K \cap (H \cap K)^{\perp}\right),$$

and if $H \cap K = \{0\}$, then $c(H,K) = c_0(H,K)$.

Proof: See, [59], Lemmas 2.3 and 2.10.

Dixmier's and Friedrichs' angles are two tools to control the relationships between two closed subspaces of an abstract infinite-dimensional Hilbert space. They admit probabilistic counterparts as generalized dependence measures between random elements, with deep links with the notion of independence and conditional independence. In the present section, the main goal is to describe the overall relationships (i.e., through these angles) between the set of subspaces $\left\{\mathbb{L}^2(\sigma_A)\right\}_{A \in \mathcal{P}_D}$ of $\mathbb{L}^2(\sigma_X)$, but *not only pairwise but also globally*. The following section introduces a particular matrix involving Friedrichs' angle.

Feshchenko matrix As illustrated above, the maximal and partial correlation are good candidates to control the dependence structure of random elements. They can be understood as a generalization of the correlation and partial correlation of random variables. Hence, they offer a natural avenue for generalizing to the notion of *covariance and precision matrices*. In particular, precision matrices (i.e., inverses of covariance matrices) can be written using partial correlations (see, e.g., [136] p.129). This idea is by no means new and has already been introduced in the study of *graphical models* [147], where generalized covariance and precision matrices have been used to study particular algebraic structures of σ-algebras. However, using Friedrichs' angle as a generalized partial correlation in such a setting seems to have yet to be done in the probability theory literature.

A novel generalization of precision matrices is introduced and named the *maximal coalitional precision matrix*. It can be loosely understood as follows:

- Each element of this matrix allows comparing two *subsets of inputs*. Hence, it is of size $\left(2^d \times 2^d\right)$, and is *indexed by the elements of* $\mathcal{P}_D$;
- For each pair of subsets of inputs, the corresponding element of this matrix contains the negative of the Friedrichs' angle between their respective generated Lebesgue spaces. In other words, in the standard definition of precision matrices, the partial correlation is replaced with the maximal partial correlation.

Formally, the maximal coalitional precision matrix can be defined as follows.

Definition 4.5 (Maximal coalitional precision matrix). Let $X = (X_1, \ldots, X_d)$ be random inputs (i.e., a vector of random elements). The maximal coalitional precision matrix of X is the $\left(2^d \times 2^d\right)$ symmetric, set-indexed matrix Δ, defined entry-wise, for any $A, B \in \mathcal{P}_D$, by

$$\Delta(A,B) = \begin{cases} 1 & \text{if } A = B; \\ -c\left(\mathbb{L}^2\left(\sigma_A\right), \mathbb{L}^2\left(\sigma_B\right)\right) & \text{otherwise.} \end{cases}$$

Furthermore, denote $\Delta|_A$ the principal $\left(2^{|A|} - 1 \times 2^{|A|} - 1\right)$ submatrix of Δ relative to the proper subsets of $A \in \mathcal{P}_D$, i.e., $\forall B, C \in \mathcal{P}_A, B \neq A, C \neq A$

$$\Delta|_A(B,C) = \Delta(B,C).$$

In the field of functional analysis, a similar type of matrix is used to derive a sufficient condition for sums of closed subspaces of an infinite-dimensional Hilbert space to be closed, following the pioneering work of Ivan Feshchenko [73, 72] on this question. Since the maximal coalitional precision matrix is ultimately used for that purpose in the following developments, and for the sake of conciseness, the matrix Δ defined in Definition 4.5 is referred to as the *Feshchenko matrix* in the remainder of this manuscript.

4.2.3 Direct sums, complemented subspaces and projections

As stated in the introduction of this chapter, one can see the HDMR of $G(X)$ with dependent inputs as a direct-sum decomposition of the Hilbert space $\mathbb{L}^2\left(\sigma_X\right)$, which involves the subspaces $\left\{\mathbb{L}^2\left(\sigma_A\right)\right\}_{A \in \mathcal{P}_D}$. To that extent, the notions of internal direct-sums and direct-sum decomposition are introduced, as well as the notion of complement when it comes to infinite-dimensional Hilbert spaces. Finally, the projection operators of Hilbert spaces are formally introduced.

(Internal) Direct sums and direct sum decomposition An internal direct-sum decomposition of a vector space entails expressing this vector space as a particular sum of subspaces. The adjective internal refers to the fact that one sums *subspaces* of an ambient vector space and differs from *external direct-sums* (see, Appendix D.1.1). However, the adjective internal is omitted in the following for conciseness, except to avoid any confusion in the developments. Direct-sum decompositions of vector spaces can be formally defined as follows.

Theorem 4.5 (Direct-sum decomposition). *Let W be a vector space, and for a positive integer n, let $W_1, \ldots, W_n$ be proper subspaces of W (i.e., $W_i \subset W$ for every $i = 1, \ldots, n$). Then, the following statements are equivalent:*

1. *Any $w \in W$ can be written uniquely as $w = \sum_{i=1}^{n} w_i$ where $w_i \in W_i$ for $i = 1, \ldots, n$;*
2. *For $i = 1, \ldots, n$, one has that $W_i \cap (W_1 + W_2 + \cdots + W_{i-1} + W_{i+1} + \cdots + W_n) = \{0\}$;*
3. *$W = +_{i=1}^{n} W_i$ and additionally, for any $w = \sum_{i=1}^{n} w_i \in W$, where $w_i \in W_i$ one has that*

$$w = 0 \implies w_i = 0, \quad i = 1, \ldots, n.$$

If any of these three conditions are met, then W is said to admit a direct sum decomposition*, which is denoted*

$$W = \bigoplus_{i=1}^{n} W_i.$$

Proof: See, [12] Definition 1.40 and Proposition 1.44.

One can notice the resemblance between the sought-after decomposition in Eq. (4.1) and the one defined in Theorem 4.5 (1.). Since the random output $G(X)$ belongs to the (vector) Hilbert space $\mathbb{L}^2(\sigma_X)$, the problem of HDMR of random outputs can be seen as finding suitable subspaces of $\mathbb{L}^2(\sigma_X)$, where each subspace would be related to every Lebesgue space $\{\mathbb{L}^2(\sigma_A)\}_{A \in \mathcal{P}_D}$, and such that they form a direct-sum decomposition of $\mathbb{L}^2(\sigma_X)$. To that extent, the notion of *complement of a subspace of a Hilbert space* is central.

Closure and complement of a subspace When dealing with infinite-dimensional Hilbert spaces, particularly its subspaces, particular attention must be paid to their closure. When the ambient (i.e., initial space) Hilbert space is finite-dimensional, every (linear vector) subspace is automatically closed, but this is not the case for infinite-dimensional ambient spaces. The notion of closedness is intrinsically similar to the notion of completeness. Formally, let $(\mathcal{H}, \|\cdot\|)$ be an infinite-dimensional Hilbert space, and let $H \subset \mathcal{H}$ be a proper subspace of $\mathcal{H}$. H is said to be closed in $\mathcal{H}$ if the limit of every converging sequence of elements of H is in H as well. Hence, if H is a closed subspace of $\mathcal{H}$, $(H, \|\cdot\|)$ is itself a Hilbert space.

More importantly, a closed proper subspace H of a Hilbert space $\mathcal{H}$ is always *complemented*, i.e., there exist some subspace K of $\mathcal{H}$ such that $\mathcal{H}$ admits the direct-sum decomposition:

$$\mathcal{H} = H \oplus K.$$

For instance, as a consequence of the Hilbert projection theorem, the *orthogonal complement* $H^{\perp}$ of H in $\mathcal{H}$, defined as

$$H^{\perp} := \{x \in \mathcal{H} : \forall y \in H, \quad \langle x, y \rangle = 0\},$$

is an example of such complement (see, e.g., [190], Theorem 12.4), as long as H is closed. Orthogonal complements are always closed. It is also important to note that many complements may exist for a single closed subspace. However, the orthogonal complement is uniquely defined (i.e., other complements are thus not orthogonal to the subspace). Hence, finding complements of subspaces is inherently linked with direct-sum decompositions, as they can be interpreted as "the remainder of the ambient space".

For the developments in this chapter, the orthogonal complements of subspaces of $\{\mathbb{L}^2(\sigma_A)\}_{A \in \mathcal{P}_D}$ are formally introduced.

Definition 4.6. Let $B \in \mathcal{P}_D$ and let H be a subspace of $\mathbb{L}^2(\sigma_B)$. For any $A \in \mathcal{P}_D$ such that $B \subseteq A$, denote

$$H^{\perp_A} = \left\{ f(X_A) \in \mathbb{L}^2(\sigma_A) : \int_{E_A} f(x_A) g(x_B) dP_{X_A}(x_A) = 0, \quad \forall g(X_B) \in H \right\},$$

i.e., the orthogonal complement of $H \subseteq \mathbb{L}^2(\sigma_B)$ in $\mathbb{L}^2(\sigma_A)$, and, in particular, denote by $\perp = \perp_D$ the orthogonal complement in $\mathbb{L}^2(\sigma_X)$.

These particular orthogonal complements have an interesting property, as described below.

Lemma 4.3. *Let $A, B \in \mathcal{P}_D$, such that $B \subseteq A$, and let H be a subspace of $\mathbb{L}^2(\sigma_B)$. Then*

$$H^{\perp_B} \subseteq H^{\perp_A}.$$

Proof: From Lemma 4.1, one has that $\mathbb{L}^2(\sigma_B) \subseteq \mathbb{L}^2(\sigma_A)$, and the proof is a direct consequence of the definition of the orthogonal complements.

In other words, the orthogonal complement w.r.t. a smaller subset is included in the orthogonal complement of a bigger subspace.

Projection operators For two Banach spaces $(\mathcal{M}_1, \|\cdot\|_1)$ and $(\mathcal{M}_2, \|\cdot\|_2)$, and a linear operator $T : \mathcal{M}_1 \to \mathcal{M}_2$ denote the range of T as

$$\mathrm{Ran}(T) := \{T(x) : x \in \mathcal{M}_1\} \subseteq \mathcal{M}_2,$$

and its nullspace as

$$\mathrm{Ker}(T) := \{x \in \mathcal{M}_1 : T(x) = 0\} \subseteq \mathcal{M}_1.$$

Let $\mathcal{H}$ be a Hilbert space and $P : \mathcal{H} \to \mathcal{H}$ be a bounded linear operator. If P is idempotent and bounded operator (i.e., $P \circ P = P$), then $\mathcal{H}$ admits the direct sum decomposition $\mathcal{H} = \mathrm{Ran}(P) \oplus \mathrm{Ker}(P)$ (see, [45], Proposition 3.2). P is then called the projector on $\mathrm{Ran}(P)$ parallel to $\mathrm{Ker}(P)$ and is defined as

$$\begin{aligned} P : \mathcal{H} = \mathrm{Ran}(P) \oplus \mathrm{Ker}(P) &\to \mathcal{H} \\ x = x_R + x_K &\mapsto x_R \end{aligned}$$

where $x_R \in \mathrm{Ran}(P)$ and $x_K \in \mathrm{Ker}(P)$. In this case, the operator $I - P$ is the projection on $\mathrm{Ker}(P)$, parallel to $\mathrm{Ran}(P)$. On the other hand, if there are two closed subspaces M and N of a Hilbert space $\mathcal{H}$ such that $\mathcal{H} = M \oplus N$, then there exists a continuous idempotent operator (i.e., a projector) P with range $\mathrm{Ran}(P) = M$ and $\mathrm{Ker}(P) = N$ (see, [84] Theorem 7.90). In this case, P is said to be *canonical* (w.r.t. to the direct sum decomposition $\mathcal{H} = M \oplus N$).

In the case where $\mathrm{Ker}(P) = \mathrm{Ran}(P)^{\perp}$, then the projection is said to be *orthogonal*, which is equivalent to P being self-adjoint (see, [84] Theorem 7.71). Hence, in this framework, for every $A \in \mathcal{P}_D$, one can see the conditional expectation operators $\mathbb{E}_A[\cdot]$ as the *orthogonal projectors* of elements of $\mathbb{L}^2(\sigma_X)$ onto $\mathbb{L}^2(\sigma_A)$, parallel to $\mathbb{L}^2(\sigma_A)^{\perp}$.

Hence, the projectors in Hilbert spaces are intrinsically linked to direct-sum decompositions. The oblique projectors, built on direct-sum decompositions, are usually called "canonical projectors". These operators and their evaluations are central in characterizing suitable model-centric influence and value measures.

4.3 Coalitional output decomposition with dependent inputs

As eluded in the previous section, being able to define an HDMR of a random output $G(X) \in \mathbb{L}^2(\sigma_X)$ can be seen as finding some "coalitional direct-sum decomposition" of $\mathbb{L}^2(\sigma_X)$. In other words, it amounts to finding, for every $A \in \mathcal{P}_D$, some subspace $V_A \subseteq \mathbb{L}^2(\sigma_A)$, such that:

$$\mathbb{L}^2(\sigma_X) = \bigoplus_{A \in \mathcal{P}_D} V_A.$$

Coming from Theorem 4.5, that would entail that any $G(X) \in \mathbb{L}^2(\sigma_X)$ can be uniquely written as

$$G(X) = \sum_{A \in \mathcal{P}_D} G_A(X_A),$$

where $G_A(X_A) \in V_A$, which is reminiscent Eq. (4.1).

This section shows that such a direct-sum decomposition is achievable under two reasonable assumptions. In addition to the result, its geometric interpretation is also discussed, and the particular case of mutually independent inputs is showcased.

4.3.1 Two reasonable assumptions

Non-perfect functional dependence The first assumption can be understood as a condition on the σ-algebra generated by the subsets of inputs when considered *as functions*. More precisely, we put a particular restriction on the intersection of their pre-images.

Assumption 2 (Non-perfect functional dependence). *For any $A, B \in \mathcal{P}_D$,*

$$\sigma_A \cap \sigma_B = \sigma_{A \cap B}$$

While mutual independence of X (see, Section 2.3.1) implies that Assumption 2 hold (see, Section 4.3.3), it is essential to note that this assumption is less restrictive. In a nutshell, it can be understood as the restriction that "the subsets of inputs cannot be expressed as a function of other subsets". This interpretation comes from the following result.

Proposition 4.1. *Let $X = (X_1, \ldots, X_d)$ be inputs, and suppose that Assumption 2 hold. Then, for any $A, B \in \mathcal{P}_D$ such that $A \cap B \notin \{A, B\}$ (i.e., the sets cannot be subsets of each other), there is no mapping $T : E_A \to E_B$ such that $X_B = T(X_A)$ a.s.*

Proof: Suppose that there exists a mapping $T : E_A \to E_B$ such that $X_B = T(X_A)$ a.s. Then, one has that $\sigma_B \subseteq \sigma_A$, which in turn implies that $\sigma_A \cap \sigma_B = \sigma_B$. Notice that necessarily $A \cap B \subset B$ and in the present framework $\sigma_{A \cap B} \subset \sigma_B$. Thus $\sigma_A \cap \sigma_B$ is necessarily different than $\sigma_{A \cap B}$, and thus Assumption 2 cannot hold. The result follows by taking the opposite implication.

Non-degenerate stochastic dependence While the first assumption considers the inputs *functionally*, the second assumption directly restricts their distribution. It can be seen as a restriction of the *inner product of the Lebesgue space* $\mathbb{L}^2(\sigma_X)$, and more precisely, it controls the angles between the subspaces $\mathbb{L}^2(\sigma_A)$, $A \in \mathcal{P}_D$ through the Feshchenko matrix Δ (see, Definition 4.5) of the inputs X. It is relatively straightforward.

Assumption 3 (Non-degenerate stochastic dependence). *The Feshchenko Δ of X is positive definite.*

Since Δ can be seen as a generalized precision matrix, this assumption is relatively reasonable since standard precision matrices (inverse of positive definite covariance matrices) are often assumed to be positive definite. One can notice that under this assumption, for any $A \in \mathcal{P}_D$, the matrices $\Delta|_A$ are also positive since they are principal submatrices of Δ. This assumption entails an interesting consequence regarding Friedrichs' angle between generated Lebesgue spaces.

Lemma 4.4. *Suppose that Assumption 3 hold. Then, for any $A, B \in \mathcal{P}_D$ such that $A \neq B$,*

$$c\left(\mathbb{L}^2(\sigma_A), \mathbb{L}^2(\sigma_B)\right) < 1.$$

Proof: Suppose that Assumption 3 hold. Then, in particular, the principal submatrix of Δ

$$\begin{pmatrix} 1 & -c\left(\mathbb{L}^2(\sigma_A), \mathbb{L}^2(\sigma_B)\right) \\ -c\left(\mathbb{L}^2(\sigma_A), \mathbb{L}^2(\sigma_B)\right) & 1 \end{pmatrix}$$

is positive definite as well, and thus,

$$2 - 2c\left(\mathbb{L}^2(\sigma_A), \mathbb{L}^2(\sigma_B)\right) > 0 \quad \Longleftrightarrow \quad c\left(\mathbb{L}^2(\sigma_A), \mathbb{L}^2(\sigma_B)\right) < 1.$$

Thus, having a definite positive Feshchenko matrix entails that the maximal partial correlation between X_A and X_B is strictly less than 1 (i.e., the angle itself must be greater than zero). Hence, it can be interpreted by the fact that the subspaces $\mathbb{L}^2(\sigma_A)$ and $\mathbb{L}^2(\sigma_B)$ *must have distinct elements*.

Moreover, when it comes to input and vector exogeneity (see, Definition 3.2), assuming Assumption 2 implies that Assumption 1 from Chapter 3 hold.

Proposition 4.2. *Let X be random inputs, and $G(X) \in \mathbb{L}^2(\sigma_X)$ be a random output. Let $E \subseteq D$, and suppose that, for every $i \in E$, X_i is an exogenous input (see, Definition 3.2). Then, if Assumption 2 is*

supposed to hold, X_E is an exogenous vector.

Proof: If for every $i \in E$, X_i is exogenous, this implies that:

$$G(X) \in \bigcap_{i \in E} \mathbb{L}^2(\sigma_{-i}).$$

However, notice that, thanks to Theorem 4.1 and under Assumption 2,

$$\begin{aligned}\bigcap_{i \in E} \mathbb{L}^2(\sigma_{-i}) &= \mathbb{L}^2\left(\bigcap_{i \in E} \sigma_{-i}\right) \\ &= \mathbb{L}^2\left(\sigma_{\bigcap_{i \in E} -i}\right) = \mathbb{L}^2(\sigma_{-E}).\end{aligned}$$

Hence $G(X) \in \mathbb{L}^2(\sigma_{-E})$ and from the Doob-Dynkin lemma (see, Lemma A.2), there exists some $f(X_{-E})$ such that $G(X) = f(X_{-E})$ a.s., and thus X_E is an exogenous vector.

4.3.2 Output decomposition and geometric interpretation

The main result of this chapter is stated below.

Theorem 4.6 (Direct-sum decomposition of $\mathbb{L}^2(\sigma_X)$). *For every $A \in \mathcal{P}_D$, let $V_\emptyset = \mathbb{L}^2(\sigma_\emptyset)$ and for every $B \in \mathcal{P}_A$, let*

$$V_B = \left[\dotplus_{C \in \mathcal{P}_{-B}} V_C\right]^{\perp_B}.$$

If Assumptions 2 and 3 hold, then for every $A \in \mathcal{P}_D$, one has that

$$\mathbb{L}^2(\sigma_A) = \bigoplus_{B \in \mathcal{P}_A} V_B.$$

Proof of Theorem 4.6 on page p. 128.

The model-centric random output decomposition follows as a corollary.

Corollary 4.1 (Orthocanonical decomposition). *Let $X = (X_1, \ldots, X_d)$ be random inputs. Suppose that Assumptions 2 and 3 hold. Then, for any $G : E \to \mathbb{R}$ such that $G(X) \in \mathbb{L}^2(\sigma_X)$, $G(X)$ can be uniquely decomposed as*

$$G(X) = \sum_{A \in \mathcal{P}_D} G_A(X_A),$$

where each $G_A(X_A) \in V_A$.

Proof: It is a direct consequence of Theorem 4.6 and Theorem 4.5.

In addition, this coalitional decomposition is *properly gradual*.

Proposition 4.3. *Let $X = (X_1, \ldots, X_d)$ be random inputs, $G(X) \in \mathbb{L}^2(\sigma_X)$ be a random output and suppose that Assumptions 2 and 3 hold. Then, each summand of the orthocanonical decomposition of $G(X)$ is a proper representant (i.e., Definition 2.2) of X_A.*

Proof: Let $A \in \mathcal{P}_D$ and notice that $G_A(X_A) \in V_A \subset \mathbb{L}^2(\sigma_A)$, and thus the σ-algebra generated by $G_A(X_A)$ is included in σ_A and it is a representant of X_A. Now let any $B \subset A$ and suppose that the σ-algebra of $G_A(X_A)$ is included in σ_B. In this case, one would have that $G_A(X_A) \in \mathbb{L}^2(\sigma_B)$. However, by construction, one has that $V_A \perp \mathbb{L}^2(\sigma_B)$, and thus, necessarily, $G_A(X_A) = 0$. Hence, if $G_A(X_A)$ is σ_B-measurable, it is necessarily constant, and its generated σ-algebra is necessarily contained in $\sigma_\emptyset$. Thus, $G_A(X_A)$ is a proper representant of X_A.

Since the output coalitional decomposition is properly gradual, it entails that $\forall A \subseteq D$, the summands $G_A(X_A) \in V_A$ in Corollary 4.1 are *exactly functions of* X_A, in the sense that if they were σ_B-measurable for any $B \subset A$ (and hence functions of X_B), they would necessarily be equal to 0.

Despite the somewhat formal nature of Theorem 4.6, its interpretation is rather intuitive. Given a univariate function $G_1(X_1) \in \mathbb{L}^2(\sigma_1)$, it is well known that it can always be decomposed as

$$G_1(X_1) = \mathbb{E}[G_1(X_1)] + [G_1(X_1) - \mathbb{E}[G_1(X_1)]]. \tag{4.4}$$

In other words, a random variable can always be decomposed as its expectation plus its centered version. The first step of the result formalizes this idea. $V_\emptyset = \mathbb{L}^2(\sigma_\emptyset)$ is comprised of constant a.e. random variables and is a closed subspace of $\mathbb{L}^2(\sigma_1)$. Thus $V_\emptyset$ is complemented in $\mathbb{L}^2(\sigma_1)$, and, in particular, it is complemented by V_1, its orthogonal complement. V_1 is thus comprised of every function of $\mathbb{L}^2(\sigma_1)$ which are orthogonal to the constants (i.e., they are centered). Thus, since $\mathbb{L}^2(\sigma_1) = V_\emptyset \oplus V_1$, one recovers the relation in (4.4).

For two inputs X_1 and X_2, Assumption 2 ensures that the subspaces $\mathbb{L}^2(\sigma_1)$ and $\mathbb{L}^2(\sigma_2)$ of $\mathbb{L}^2(\sigma_{12})$ are not comprised of the same random variables, due to a *functional relation between* X_1 *and* X_2. On the other hand, Assumption 3 ensures that these subspaces are not the same due to a *degenerate stochastic relation*. Under those two assumptions, the sum $\mathbb{L}^2(\sigma_1) + \mathbb{L}^2(\sigma_2) = V_\emptyset + V_1 + V_2$ is a closed subspace of $\mathbb{L}^2(\sigma_{12})$, and thus, is complemented by V_{12} which is none other than its orthogonal complement. Notice that V_1 and V_2 are *not necessarily orthogonal*, but both are orthogonal to $V_\emptyset$ and V_{12}.

The same reasoning can be applied with three inputs. The two assumptions ensure that $\mathbb{L}^2(\sigma_{12})$, $\mathbb{L}^2(\sigma_{23})$, and $\mathbb{L}^2(\sigma_{13})$ are not pairwise equal due to either a functional or a stochastic relation. In this case, their sum is a closed subspace in $\mathbb{L}^2(\sigma_{123})$, and thus, it is complemented by V_{123} (i.e., the orthogonal complement of $\mathbb{L}^2(\sigma_{12}) + \mathbb{L}^2(\sigma_{23}) + \mathbb{L}^2(\sigma_{13})$). However, notice that neither V_{12}, V_{13} and V_{23} are pairwise orthogonal, nor V_1, V_2 and V_3. The same mechanism can be continued for any number of inputs.

Hence, the subspaces $(V_A)_{A \in \mathcal{P}_D}$ in Theorem 4.6 can be interpreted as the subspaces of functions of X which, for any $A \in \mathcal{P}_D$, are σ_A-measurable (i.e., are functions of X_A), but are orthogonal to the linear combinations of functions in $(V_B)_{B \in \mathcal{P}_{-A}}$. In other words, the elements of V_A can only contain proper representants, which can be understood as multivariate non-linear functions of exactly X_A. For instance, for two inputs X_1 and X_2, V_{12} can be seen as the space of functions of X_1 and X_2, that "are not" (in the sense of being the complement of) linear combinations of functions of X_1 and X_2. Given this construction, a natural interpretation of V_A would be the space of "interactions" between the inputs X_A.

One can additionally notice some structure in the construction depicted above. In particular, some of the subspaces in $(V_A)_{A \in \mathcal{P}_D}$ are pairwise orthogonal, while others are not necessarily. It is known as a *hierarchical orthogonality structure*, which is further discussed in the following.

Hierarchical orthogonality The set of subspaces $(V_A)_{A \in \mathcal{P}_D}$ presents a particular orthogonality structure, namely *hierarchical orthogonality*, reminiscent of the one described in [38]. However, in our framework, this structure arises naturally rather than by construction.

Proposition 4.4 (Hierarchical orthogonality). *We place ourselves in the framework of Theorem 4.6. For any $A \in \mathcal{P}_D$, and any $B \subset A$*

$$V_A \perp V_B.$$

Proof: It is a direct consequence of the definition of V_A.

This particular structure can be illustrated using the Boolean lattice described in Section 2.2.3. This structure can be illustrated using a Hasse diagram, as in Figure 4.1 a). One can notice that $(V_A)_{A \in \mathcal{P}_D}$ endowed with the binary relation $\perp$ (i.e., the relation "is in the orthogonal complement of"), then the algebraic structure is preserved, as illustrated in Figure 4.1 b). In order to formally differentiate between the structurally hierarchical subspaces and those that are not necessarily orthogonal, two different sets related to this structure are introduced. For any $A \in \mathcal{P}_D$, the first one is the set of *comparables* (i.e., the elements of $\mathcal{P}_D$ that are subsets of A or such that A is a subset of), denoted

$$\mathcal{C}_A = \mathcal{P}_A \cup \{B \in \mathcal{P}_D : A \subseteq D\},$$

and notice that, for any $B \in \mathcal{C}_A$, $V_B \perp V_A$. Then, we define the set of *uncomparables* of A as

$$\mathcal{U}_A = \mathcal{P}_D \setminus \mathcal{C}_A,$$

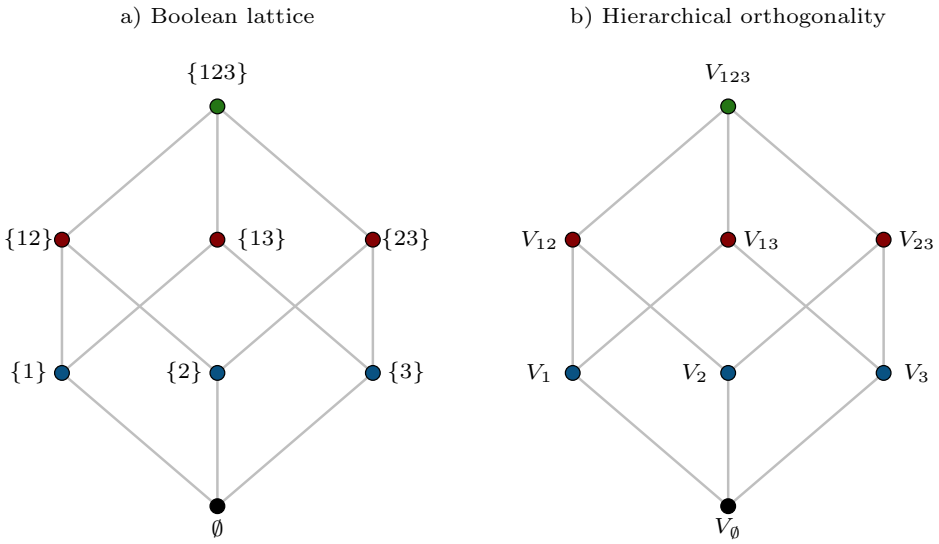

Figure 4.1: Illustration of the hierarchical orthogonality structure for three inputs. These Hasse diagrams are meant to be read from the bottom to the top. If an edge joins two elements, the binary relation to the above element links the bottom element. On a), the binary relation is $\subset$, while on b) the binary relation is $\perp$.

and notice that, in general, for every $B \in \mathcal{U}_A$, V_A is not necessarily orthogonal to V_B. And notice additionally that, for any $A \in \mathcal{P}_D$

$$\mathcal{P}_D = \mathcal{C}_A \cup \mathcal{U}_A.$$

Remark 4.3. It is important to note that the hierarchical orthogonality of the subspaces $(V_A)_{A\in\mathcal{P}_D}$ is a consequence of the *choice of inductively choosing orthogonal complements* in Theorem 4.6. Other complements, i.e., not necessarily orthogonal, could have been chosen, leading to a different structure. This is why the output decomposition in Corollary 4.1 is called "orthocanonical". A different choice of complements could lead to a different decomposition.

Canonical projections First, assuming that Theorem 4.6 holds, two different projectors onto the subspaces V_A, for every $A \in \mathcal{P}_D$, can be defined. Let A be any element of $\mathcal{P}_D$. Denote by P_A the orthogonal projector onto V_A, i.e., ,

$$P_A : \mathbb{L}^2(\sigma_X) \to \mathbb{L}^2(\sigma_X), \quad \text{such that } \mathrm{Ran}(P_A) = V_A \text{ and } \mathrm{Ker}(P_A) = V_A^\perp.$$

Since V_A is a closed subspace of $\mathbb{L}^2(\sigma_X)$, the orthogonal projector P_A exists and is uniquely defined. Additionally, for every $A \in \mathcal{P}_D$, denote the following subspaces of $\mathbb{L}^2(\sigma_X)$

$$W_A = \bigoplus_{B\in\mathcal{P}_D : B\neq A} V_B,$$

and the operators

$$\begin{aligned} Q_A : \mathbb{L}^2(\sigma_X) &\to \mathbb{L}^2(\sigma_X) \\ G(X) = \sum_{B\in\mathcal{P}_D} G_B(X_B) &\mapsto G_A(X_A) \end{aligned}$$

and notice that Q_A is the projector onto V_A parallel to W_A, which is well-defined thanks to the direct-sum decomposition of Theorem 4.6 (see, [181] Theorem 3.4). For any $A \in \mathcal{P}_D$, the operators $P_A(\cdot)$ and $Q_A(\cdot)$ are both projectors onto V_A, but their nullspaces differ.

The projectors onto the subspaces $\mathbb{L}^2(\sigma_A)$, for every $A \in \mathcal{P}_D$ are now defined. First, the orthogonal projector onto $\mathbb{L}^2(\sigma_A)$ is defined as

$$\mathbb{E}_A : \mathbb{L}^2(\sigma_X) \to \mathbb{L}^2(\sigma_X), \quad \text{such that } \mathrm{Ran}(\mathbb{E}_A) = \mathbb{L}^2(\sigma_A) \text{ and } \mathrm{Ker}(P_A) = \mathbb{L}^2(\sigma_A)^{\perp},$$

and notice that it is the conditional expectation operator of $H(X)$ given X_A (see, [126], Chapter 8). Additionally, denote the subspace

$$\widetilde{W}_A = \bigoplus_{B \in \mathcal{P}_D, B \notin \mathcal{P}_A} V_B$$

and the operator

$$\begin{aligned} \mathbb{M}_A : \mathbb{L}^2(\sigma_X) &\to \mathbb{L}^2(\sigma_X) \\ G(X) = \sum_{B \in \mathcal{P}_D} G_B(X_B) &\mapsto \sum_{B \in \mathcal{P}_A} G_B(X_B) \end{aligned}$$

and, thanks to Theorem 4.6, notice that $\mathbb{M}_A$ is the projection onto $\mathrm{Ran}(\mathbb{M}_A) = \mathbb{L}^2(\sigma_A)$ parallel to $\mathrm{Ker}(\mathbb{M}_A) = \widetilde{W}_A$. Hence, for any $A \in \mathcal{P}_D$, the operators $\mathbb{E}_A[\cdot]$ and $\mathbb{M}_A[\cdot]$ are two projections onto $\mathbb{L}^2(\sigma_A)$, but with different nullspaces. While $\mathbb{E}_A[\cdot]$ represents the well-known conditional expectation operator in probability theory, the operator $\mathbb{M}_A[\cdot]$ does not seem to have been extensively studied in this literature.

The first result is a particular consequence of the hierarchical orthogonality structure. It is known as the *annihilating property* (see, e.g., [105] Lemma 1, or [134]), which has been well-documented in the case of mutually independent inputs. This property admits a somewhat surprising generalization in the framework of Theorem 4.6.

Proposition 4.5 (Annihilating property). *We place ourselves in the framework of Theorem 4.6 and Corollary 4.1. For any $A \in \mathcal{P}_D$ and any $B \subset A$*

$$P_B(Q_A(G(X))) = P_B(G_A(X_A)) = 0.$$

Proof: From Proposition 4.4, for every $B \subset A$, one has that $V_B \perp V_A$, and thus $G_A(X_A) \in V_A \subset V_B^{\perp}$.

Another interesting result is the fact that the oblique projections $(Q_A)_{A \in \mathcal{P}_D}$ onto the V_A can be expressed in terms of the oblique projections $(\mathbb{M}_A)_{A \in \mathcal{P}_D}$ onto the generated Lebesgue spaces.

Proposition 4.6 (Formula for oblique projections). *We place ourselves in the framework of Theorem 4.6 and Corollary 4.1. One has that, for any $G(X) \in \mathbb{L}^2(\sigma_X)$, and for any $A \in \mathcal{P}_D$*

$$Q_A(G(X)) = \sum_{B \in \mathcal{P}_A} (-1)^{|A|-|B|} \mathbb{M}_A(G(X)),$$

where $|.|$ denotes the cardinality of sets.

Proof: By definition of $\mathbb{M}_A$, one has that

$$\forall A \in \mathcal{P}_D, \quad \mathbb{M}_A(G(X)) = \sum_{B \in \mathcal{P}_A} Q_A(G(X)),$$

which, thanks to Rota's generalization of the Möbius inversion formula in Corollary 2.1, is equivalent to

$$\forall A \in \mathcal{P}_D, \quad Q_A(G(X)) = \sum_{B \in \mathcal{P}_A} (-1)^{|A|-|B|} \mathbb{M}_A(G(X)).$$

Proposition 4.6 highlights the fact that Theorem 4.6 is indeed a *model-centric* properly gradual decomposition of the random output $G(X)$. It is interesting to note that, in this particular case, the value measure is defined as being the canonical oblique projection $\mathbb{M}_A(G(X))$, leading to the influence measure $Q_A(G(X)) = G_A(X_A)$.

In order to better visualize how the decomposition of Theorem 4.6 can be understood in terms of projections, one can take an example with two inputs X_1 and X_2, and a centered random output

$G(X_1, X_2) \in \mathbb{L}^2(\sigma_{12})$. Figure 4.2 illustrates this situation. $G(X_1, X_2)$ can be written as a sum of three elements, $G_1(X_1) \in V_1$, $G_2(X_2) \in V_2$ and $G_{12}(X_1, X_2) \in V_{12}$. $G_{12}(X_1, X_2)$ is none other than the orthogonal projection of G onto V_{12}, due to the fact that V_{12} is the orthogonal complement of $V_1 + V_2$ and, naturally, $G_1(X_1) + G_2(X_2) = [I - P_{12}](G(X))$. Now, recall that since V_1 and V_2 are not necessarily orthogonal (which is represented as the angle α (which is non-zero, thanks to the Assumptions 2 and 3) in Figure 4.2), $G_1(X_1)$ (resp. $G_2(X_2)$) is none other than the oblique projection of $G(X)$ onto V_1 parallel to V_2 (resp. onto V_2 parallel to V_1).

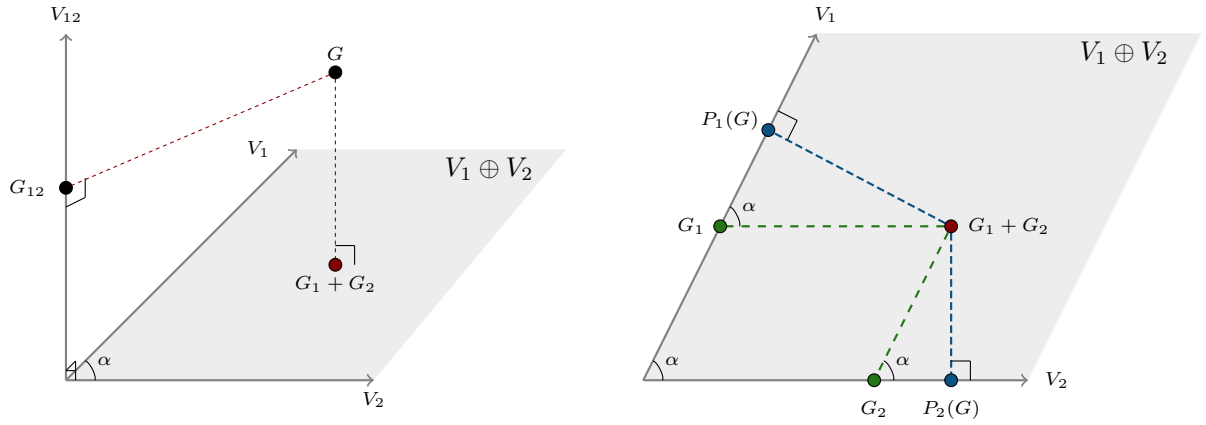

Figure 4.2: Illustration of a centered function decomposition with two dependent inputs.

Remark 4.4. It is essential to note that Figure 4.2 is a mere illustration and should not be treated as a rigorous representation. In fact, the subspaces V_1, V_2 and V_{12} are **infinite-dimensional**.

4.3.3 Mutual independence and Hoeffding's decomposition

It is well-known that the independence of two random elements (w.r.t. to $\mathbb{P}$) is linked to the independence of the σ-algebras they generate, which can be characterized by the orthogonality of the centered generated Lebesgue spaces. More precisely, two sub σ-algebras $\mathcal{A}_1$ and $\mathcal{A}_2$ of a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ are said to be independent if $\mathbb{L}^2(\mathcal{A}_1)$ and $\mathbb{L}^2(\mathcal{A}_2)$ are orthogonal on the constant functions (see, [150], Chapter IV, Definition 3.0.1). More precisely,

$$\mathcal{A}_1 \perp\!\!\!\perp \mathcal{A}_2 \iff c_0\left(\mathbb{L}_0^2(\mathcal{A}_1), \mathbb{L}_0^2(\mathcal{A}_2)\right) = 0,$$

where $\perp$ is defined relative to the inner product on $\mathbb{L}^2(\mathcal{F})$. Additionally, two random elements Z_1, Z_2 defined on $(\Omega, \mathcal{F}, \mathbb{P})$ are considered independent if their generated σ-algebras are independent.

When dealing with a vector of random elements $X = (X_1, \ldots, X_d)$, mutual independence can be defined w.r.t. the independence of their generated σ-algebras. More precisely, X is said to be mutually independent if

$$\forall A \in \mathcal{P}_D, \quad \sigma_A \perp\!\!\!\perp \sigma_{D\setminus A} \iff c_0\left(\mathbb{L}_0^2(\sigma_A), \mathbb{L}_0^2(\sigma_{D\setminus A})\right) = 0.$$

Proposition 4.7. *Let X be a vector of random elements. If X is mutually independent, then Assumption 2 hold.*

Proof: From [150], note that for two σ-algebras $\mathcal{B}_1$ and $\mathcal{B}_2$,

$$\mathcal{B}_1 \perp\!\!\!\perp \mathcal{B}_2 \quad \Longrightarrow \quad \mathcal{B}_1 \cap \mathcal{B}_2 = \sigma_\emptyset.$$

Suppose that Assumption 2 does not hold. Hence, in particular, for any $A \in \mathcal{P}_D$,

$$\sigma_A \cap \sigma_{D\setminus A} \neq \sigma_\emptyset.$$

It implies that σ_A and $\sigma_{D\setminus A}$ cannot be independent. Hence, since this holds for any $A \in \mathcal{P}_D$, X cannot be mutually independent. The result is proven by taking the opposite implication.

Proposition 4.8. *Let X be a vector of random elements and suppose that Assumption 2 holds. X is mutually independent if and only if $\forall A, B \in \mathcal{P}_D$, $A \neq B$,*

$$c\left(\mathbb{L}^2\left(\sigma_A\right), \mathbb{L}^2\left(\sigma_B\right)\right) = 0.$$

Proof: Notice that, in the general case, if $B \subset A$, then $c\left(\mathbb{L}^2\left(\sigma_A\right), \mathbb{L}^2\left(\sigma_B\right)\right)$ is necessarily equal to zero. Thus, we focus on the case where $A \cap B = C \notin \{A, B\}$. Now, suppose that for any $A, B \in \mathcal{P}_D$, $c\left(\mathbb{L}^2\left(\sigma_A\right), \mathbb{L}^2\left(\sigma_B\right)\right) = 0$. Hence, in particular, under Assumption 2, notice that for every $A \in \mathcal{P}_D$

$$\begin{aligned} c\left(\mathbb{L}^2\left(\sigma_A\right), \mathbb{L}^2\left(\sigma_{D\setminus A}\right)\right) &= c_0\left(\mathbb{L}^2\left(\sigma_A\right) \cap \mathbb{L}^2\left(\sigma_\emptyset\right)^\perp, \mathbb{L}^2\left(\sigma_{D\setminus A}\right) \cap \mathbb{L}^2\left(\sigma_\emptyset\right)^\perp\right) \\ &= c_0\left(\mathbb{L}_0^2\left(\sigma_A\right), \mathbb{L}_0^2\left(\sigma_{D\setminus A}\right)\right). \end{aligned}$$

Thus, for every $A \in \mathcal{P}_D$,

$$c_0\left(\mathbb{L}_0^2\left(\sigma_A\right), \mathbb{L}_0^2\left(\sigma_{D\setminus A}\right)\right) = 0 \iff \sigma_A \perp\!\!\!\perp \sigma_{D\setminus A},$$

which is equivalent to X being mutually independent.

Suppose that X is mutually independent, and thus, $P_X = \bigtimes_{i\in D} P_{X_i}$, which implies that, for any $A, B \in \mathcal{P}_D$, with $A \cap B = C \notin \{A, B\}$,

$$\mathbb{E}_A \circ \mathbb{E}_B = \mathbb{E}_B \circ \mathbb{E}_A = \mathbb{E}_C,$$

Thus, the orthogonal projections onto $\mathbb{L}^2\left(\sigma_A\right)$ and $\mathbb{L}^2\left(\sigma_B\right)$ commute, which is equivalent to (see, (4.2))

$$c\left(\mathbb{L}^2\left(\sigma_A\right), \mathbb{L}^2\left(\sigma_B\right)\right) = 0.$$

Corollary 4.2. *Let X be a vector of random elements and suppose that Assumption 2 holds. X is mutually independent if and only if its Feshchenko matrix Δ is the identity.*

Proof: It is a direct consequence of Proposition 4.8, by definition of Δ.

Hence, if the inputs are mutually independent, both Assumption 2 and Assumption 3 hold and lead to the very particular case of Δ being the identity matrix. One has the following result when it comes to the resulting decomposition of $\mathbb{L}^2\left(\sigma_X\right)$.

Proposition 4.9. *Let X be random inputs and suppose that Assumption 2 holds. X is mutually independent if and only if*

$$\forall A, B \in \mathcal{P}_D, B \neq A \quad V_A \perp V_B.$$

Proof: Notice that, in general, if Assumption 2 hold, one has that for any $A, B \in \mathcal{P}_D$, $B \neq A$

$$c_0\left(V_A, V_B\right) \leq c\left(\mathbb{L}^2\left(\sigma_A\right), \mathbb{L}^2\left(\sigma_B\right)\right).$$

Note that, from Proposition 4.7, Assumption 2 holds for a mutually independent X. Moreover, notice from Proposition 4.8 that X is mutually independent if and only if, $\forall A, B \in \mathcal{P}_D$, $A \neq B$, $c\left(\mathbb{L}^2\left(\sigma_A\right), \mathbb{L}^2\left(\sigma_B\right)\right) = 0$, thus necessarily $c_0\left(V_A, V_B\right) = 0$, which is equivalent to $V_A \perp V_B$.

Proposition 4.9 is, in fact, equivalent to the Hoeffding functional decomposition for mutually independent inputs (see, Theorem 2.1), which can be seen as a very particular case of Theorem 4.6 where X admits a Feshchenko matrix equal to the identity. In this very particular case, the subspaces V_A are all pairwise orthogonal, and the *orthogonal and oblique projectors are equal*, leading to Eq. (2.1) which can be seen as a particular case of Proposition 4.6.

In a nutshell, Theorem 4.6 generalizes Hoeffding's decomposition in Theorem 2.1 to Feshchenko matrix *that are different from the identity*, or in other words, for not-necessarily mutually independent inputs. It allows characterizing properly gradual output decompositions, which can be, in turn, used for defining influence measures in the same fashion as the FANOVA in Corollary 2.2. The following section is dedicated to exploring and studying such influence measures.

4.4 Model-centric influence measures

This section is dedicated to introducing the notion of *orthocanonical decompositions of QoIs*, which can be understood as influence measures relying on the (fairly general) model-centric decomposition offered by Theorem 4.6. In particular, two QoI decompositions are presented: the decomposition of an evaluation of the model and the decomposition of its variance.

4.4.1 Orthocanonical evaluation decomposition

For $\omega \in \Omega$, denote $x = X(\omega) \in E$ an observation of X. Subsequently, denote $G(x) \in \mathbb{R}$ the evaluation on x of a random output $G(X) \in \mathbb{L}^2(\sigma_X)$. In this case, the QoI is the observation $G(x)$ itself. Thanks to Theorem 4.6, one can define the *orthocanonical decomposition of* $G(x)$.

Definition 4.7 (Orthocanonical decomposition of an evaluation). Let $X = (X_1, \dots, X_d)$ be a vector of random elements, let $G(X) \in \mathbb{L}^2(\sigma_X)$ be a random output and assume that Assumptions 2 and 3 hold. For any $\omega \in \Omega$, denote $x = X(\omega)$ and for every $A \in \mathcal{P}_D$, denote $x_A = X_A(\omega)$. From Theorem 4.6 notice that:

$$G(x) = \sum_{A \in \mathcal{P}_D} G_A(x_A).$$

where $G_A(X_A) \in V_A$. The *orthocanonical decomposition of the evaluation* $G(x)$ is the properly gradual influence measure $\phi : \mathcal{P}_D \to \mathbb{R}$ defined as:

$$\phi(A) = Q_A(G(x)) = G_A(x_A) = \sum_{B \in \mathcal{P}_A} (-1)^{|A|-|B|} \mathbb{M}_B\left(G(x)\right)$$

where Q_A is the projection onto V_A parallel to W_A and $\mathbb{M}_A$ is the projection onto $\mathbb{L}^2(\sigma_A)$ parallel to $\breve{W}_A$.

From an input-centric standpoint (see, Chapter 3), where one begins with the choice of a value measure $v : \mathcal{P}_D \to \mathbb{R}$ in order to derive a coalitional decomposition of $G(x)$, the only choice of value measure leading to the orthocanonical evaluation decomposition is relatively straightforward: one must choose the oblique projections $\mathbb{M}$. Moreover, the usual choice of the conditional expectation, as done in [149], is suitable if and only if the inputs are mutually independent, as highlighted by the following result.

Proposition 4.10. *Let $X = (X_1, \dots, X_d)$ be random inputs, $G(X) \in \mathbb{L}^2(\sigma_X)$ be a random output, and assume that Assumptions 2 and 3 hold. Then, the orthocanonical decomposition ϕ of $G(x)$ is equal to*

$$\phi(A) = \sum_{B \in \mathcal{P}_A} (-1)^{|A|-|B|} \mathbb{E}_B\left(G(x)\right), \quad \forall A \in \mathcal{P}_D$$

if and only if X is mutually independent.

Proof: First, notice that $\mathbb{M}_A = \mathbb{E}_A$ if and only if $\breve{W}_A$ is the orthogonal complement of $\mathbb{L}^2(\sigma_A)$. One can notice that, $\breve{W}_A$ is a complement of $\mathbb{L}^2(\sigma_A)$ in $\mathbb{L}^2(\sigma_X)$, and from Proposition 4.9, one has that

$$\mathbb{L}^2(\sigma_A) = \bigoplus_{B \in \mathcal{P}_A} V_B \perp \breve{W}_A = \bigoplus_{B \in \mathcal{P}_D, B \notin \mathcal{P}_A} V_B,$$

hold for every $A \in \mathcal{P}_D$ if and only if X is mutually independent. In this case, $\breve{W}_A$ is an orthogonal complement of $\mathbb{L}^2(\sigma_A)$, and by unicity, $\breve{W}_A = \mathbb{L}^2(\sigma_A)^\perp$, and thus $\mathbb{M}_A = \mathbb{E}_A$.

Hence, choosing the conditional expectations for the input-centric evaluation decomposition to be orthocanonical is *suitable if and only if X is mutually independent*. Moreover, as highlighted in Chapter 3, the allocations in the Harsanyi set can be seen as aggregations of a coalitional decomposition. In particular, one can define the orthocanonical Shapley attribution scheme, as an aggregation of the orthocanonical evaluation decomposition.

Definition 4.8 (Orthocanonical Shapley attribution for dependent inputs). Let $X = (X_1, \dots, X_d)$ be random inputs, $G(X) \in \mathbb{L}^2(\sigma_X)$ be a random output and suppose that Assumptions 2 and 3 hold. The canonical Shapley attribution of an evaluation $G(x)$ is the allocation C-Sh$D \to \mathbb{R}$ given, for every $i \in D$, by

$$\text{C-Sh}_i = \sum_{A \in \mathcal{P}_D : i \in A} \frac{Q_A(G(x))}{|A|} = \sum_{A \in \mathcal{P}_D : i \in A} \frac{\sum_{B \in \mathcal{P}_A} (-1)^{|A|-|B|} \mathbb{M}_A[G(x)]}{|A|}.$$

Hence, the orthocanonical Shapley attribution is the Shapley values of the cooperative game (D, v) where the value function v is given by

$$v(A) = \mathbb{M}_A[G(X)] = \sum_{B \in \mathcal{P}_A} Q_A(G(X)), \quad \forall A \in \mathcal{P}_D,$$

and the subsequent Harsanyi dividends of (D, v) are none other than the orthocanonical evaluation decomposition of $G(x)$, i.e.,

$$\mathcal{D}_v(A) = Q_A(G(X)) = \sum_{B \in \mathcal{P}_A} (-1)^{|A|-|B|} \mathbb{M}_A[G(x)], \quad \forall A \in \mathcal{P}_D.$$

While these indices rely on the natural decomposition of $\mathbb{L}^2(\sigma_X)$ in the context of dependent inputs, they remain an aggregation of the canonical decomposition of $G(X)$. However, in this case, the choice of value function can be justified, and the dividends can be geometrically interpreted.

4.4.2 Variance decomposition

For the case of importance quantification, the QoI is $\mathbb{V}(G(X))$. Two ways to approach the problem of decomposing $\mathbb{V}(G(X))$ are proposed: The *orthocanonical variance decomposition* relies on the orthocanonical decomposition of $G(X)$ (see, Corollary 4.1). In contrast, the *organic variance decomposition* aims at defining and disentangling *pure interaction effects* from *dependence effects*.

Canonical variance decomposition In light of Corollary 4.1, the orthocanonical variance decomposition of $G(X)$ is rather intuitive. It relies on the following rationale:

$$\begin{aligned}
\mathbb{V}(G(X)) &= \text{Cov}(G(X), G(X)) \\
&= \sum_{A \in \mathcal{P}_D} \text{Cov}(G_A(X_A), G(X)) \\
&= \sum_{A \in \mathcal{P}_D} \left[\mathbb{V}(G_A(X_A)) + \sum_{B \in \mathcal{U}_A} \text{Cov}(G_A(X_A), G_B(X_B)) \right].
\end{aligned}$$

reminiscent of the well-known "covariance decomposition" [209, 38, 96, 48]. Two indices arise from this decomposition.

Definition 4.9 (Orthocanonical variance decomposition). We place ourselves in the framework of Theorem 4.6. For any $A \in \mathcal{P}_D$, let

$$S_A^U = \mathbb{V}(G_A(X_A)),$$

defines the *structural contribution of* X_A *to* $G(X)$, while

$$S_A^C = \sum_{B \in \mathcal{U}_A} \text{Cov}(G_A(X_A), G_B(X_B)),$$

represents the *correlative contribution of* X_A *to* $G(X)$.

Remark 4.5. It is important to note that both the magnitude of S_A^U and S_A^C varies w.r.t. the dependence structure of the inputs (i.e., the angles between the subspaces $(V_A)_{A \in \mathcal{P}_D}$). Hence, S_A^C cannot be understood as a pure quantification of "dependence effects" and S_A^U cannot quantify "pure interaction".

The canonical decomposition of $\mathbb{V}(G(X))$ is suitable in practice if the dependence structure of X is assumed to be inherent in the modeling of the studied phenomenon. In other words, if one aims to understand the global relationship between X and $G(X)$. Moreover, these two indices can be computed using the oblique projections $\mathbb{M}$.

Proposition 4.11. *Suppose that Theorem 4.6 hold. Then, for any $A \in \mathcal{P}_D$*

$$S_A^C = \sum_{B \in \mathcal{P}_A} (-1)^{|A|-|B|} Cov\left(\mathbb{M}_B(G(X)), [I - \mathbb{M}_A](G(X))\right).$$

Proof of Proposition 4.11 on p.132.

Proposition 4.12. *Suppose that Theorem 4.6 hold. Then, for any $A \in \mathcal{P}_D$*

$$S_A^U = \sum_{B \in \mathcal{P}_A} (-1)^{|A|-|B|} \left[\mathbb{V}\left(\mathbb{M}_B(G(X))\right) - Cov\left(\mathbb{M}_B(G(X)), [I - \mathbb{M}_A](G(X))\right)\right].$$

Proof of Proposition 4.12 on p.133.

Organic variance decomposition The goal of the *organic variance decomposition* is to separate "pure interaction effects" to "dependence effects". Pure interaction can be seen as studying the functional relation between the inputs X and the random output $G(X)$ without considering the dependence structure of X. Hence, it amounts to performing a canonical variance decomposition of $\mathbb{V}(G(X))$ *under mutual independence of* X. Formally, let $X = (X_1, \ldots, X_d)$ be a vector of random elements. The induced probability measure P_X is not necessarily the product measure $\bigtimes_{i \in D} P_{X_i}$. Now, denote $\widetilde{X} = (\widetilde{X}_1, \ldots, \widetilde{X}_d)$ the vector of random elements such that

$$\forall i \in D, \quad \widetilde{X}_i \text{ and } X_i \text{ have the same distribution.} \quad \text{and} \quad P_{\widetilde{X}} := \bigtimes_{i \in D} P_{X_i}.$$

In other words, X and $\widetilde{X}$ have the same univariate marginals, but $\widetilde{X}$ is the mutual independent version of X and, for any $A \in \mathcal{P}_D$, denote $\widetilde{X}_A$ its marginals. Suppose that $G(X) \in \mathbb{L}^2(\sigma_X)$ and $G(\widetilde{X}) \in \mathbb{L}^2(\sigma_{\widetilde{X}})$, and, for any $H(\widetilde{X})$ denote

$$\mathbb{E}^{\perp\!\!\!\perp}\left[H(\widetilde{X})\right] := \int_E H(x) \prod_{i \in D} dP_{X_i}(x_i), \quad \text{and } \mathbb{V}^{\perp\!\!\!\perp}(H(X)) := \mathbb{E}^{\perp\!\!\!\perp}\left[\left(H(\widetilde{X}) - \mathbb{E}^{\perp\!\!\!\perp}\left[H(\widetilde{X})\right]\right)^2\right].$$

Notice that, since $\widetilde{X}$ is mutually independent, it respects both Assumptions 2 and 3, and hence, one can perform the following canonical decomposition in $\mathbb{L}^2(\sigma_{\widetilde{X}})$

$$G(\widetilde{X}) = \sum_{A \in \mathcal{P}_D} \widetilde{G}_A(\widetilde{X}_A),$$

where the $\widetilde{G}_A(\widetilde{X}_A)$ are all pairwise orthogonal (see, Section 4.3.3), and hence

$$\mathbb{V}^{\perp\!\!\!\perp}\left(G(\widetilde{X})\right) = \sum_{A \in \mathcal{P}_D} \mathbb{V}^{\perp\!\!\!\perp}\left(\widetilde{G}_A(\widetilde{X}_A)\right)$$

The following influence measures are proposed.

Definition 4.10 (Pure interaction effect). Suppose that Theorem 4.6 hold. For any $A \in \mathcal{P}_D$, let

$$S_A = \frac{\mathbb{V}^{\perp\!\!\!\perp}\left(\widetilde{G}_A\right)}{\mathbb{V}^{\perp\!\!\!\perp}\left(G(\widetilde{X})\right)} \mathbb{V}(G(X))$$

define the *pure interaction indices*.

These indices are, in fact, the Sobol' indices of $G(\widetilde{X})$ [204], which are known in the literature as quantifying pure interaction [48]. These indices can also be expressed as functions of the orthogonal projections onto the subspaces $\mathbb{L}^2\left(\sigma_{\widetilde{X}_A}\right)$ as follows

$$S_A = \sum_{B \in \mathcal{P}_A} (-1)^{|A|-|B|} \mathbb{E}_B^{\perp\!\!\!\perp}\left(G(\widetilde{X})\right),$$

where, $\forall A \in \mathcal{P}_D$

$$\mathbb{E}_A^{\perp\!\!\!\perp}\left(G(\widetilde{X})\right) = \int_{E_{D\setminus A}} G(\widetilde{X}_A, x_{D\setminus A}) \prod_{i \in D\setminus A} dP_{X_i}(x_i).$$

For further considerations about these indices, the interested reader is referred to [48].

Remark 4.6. In certain situations, when the measure induced by X is part of a particular family of random vectors, it is possible to find a mapping $T : E \to E$ such that

$$\widetilde{X} = T(X).$$

For instance, if P_X is in the family of elliptical distribution, it amounts to performing a Nataf transform of the inputs [138, 137].

One desirability criterion can be brought forward when defining dependence effects: the set of indices must all be equal to zero if and only if X is mutually independent. Formally, denote $(\phi_A)_{A\in\mathcal{P}_D}$ an abstract set of dependence effects. One must have that

$$\forall A \in \mathcal{P}_D, \quad \phi_A = 0, \quad \Longleftrightarrow \quad X \text{ is mutually independent.}$$

Thanks to the geometric interpretation of the canonical decomposition of $\mathbb{L}^2(\sigma_X)$, many quantities that respect this property can be defined. However, we focus on one particular quantity, which can be easily interpreted.

Lemma 4.5. *Suppose that Theorem 4.6 hold. Let $G(X) \in \mathbb{L}^2(\sigma_X)$. Then,*

$$Q_A(G(X)) = P_A(G(X)) \text{ a.s.}, \quad \forall A \in \mathcal{P}_D \quad \Longleftrightarrow \quad X \text{ is mutually independent.}$$

Proof of Lemma 4.5 on p.133.

In other words, Lemma 4.5 states that the oblique projections Q_A are orthogonal if and only if X is mutually independent. Hence, a relatively intuitive index would quantify the distance between these two projections.

Definition 4.11 (Dependence effects). Suppose that Theorem 4.6 hold. For any $A \in \mathcal{P}_D$, let

$$S_A^D = \mathbb{V}\left(Q_A(G(X)) - P_A(G(X))\right) = \mathbb{E}\left[\left(Q_A(G(X)) - P_A(G(X))\right)^2\right]$$

define the *dependence effect of X_A*.

Furthermore, these indices are naturally all zero if and only if X is mutually independent.

Proposition 4.13.

$$S_A^D = 0, \forall A \in \mathcal{P}_D \quad \Longleftrightarrow \quad X \text{ is mutually independent.}$$

Proof: This is a direct consequence of Lemma 4.5, coupled with the fact that the expected squared distance is a distance.

Links between the canonical and organic indices The second link entails that the correlative effects sum up to the sum of the differences between the structural and pure interaction effects.

Proposition 4.14. *Suppose that Theorem 4.6 hold. One has that*

$$\sum_{A\in\mathcal{P}_D} S_A^C = \sum_{A\in\mathcal{P}_D} \left[S_A - S_A^U\right].$$

Proof: Notice that

$$\sum_{A\in\mathcal{P}_D} S_A = \mathbb{V}\left(G(X)\right) = \sum_{A\in\mathcal{P}_D} S_A^U + S_A^C$$

and thus

$$\sum_{A\in\mathcal{P}_D} \left[S_A - S_A^U\right] = \sum_{A\in\mathcal{P}_D} S_A^C.$$

Hence, the sum of the correlative indices can be interpreted as the difference between the sum of the pure interaction effects and the structural effects.

4.5 Analytical example: two Bernoulli inputs

In order to illustrate Theorem 4.6, one can take an interest in the following illustration: X is defined as two Bernoulli random variables (here, $E = \{0,1\}^2$) X_1 and X_2, with success probability q_1 and q_2 respectively. The joint law of X can be fully expressed using three parameters: q_1, q_2, and $\rho = \mathbb{E}\left[X_1 X_2\right]$. More precisely, one has that:

$$\begin{cases} p_{00} = 1 - q_1 - q_2 + \rho \\ p_{01} = q_2 - \rho \\ p_{10} = q_1 - \rho \\ p_{11} = \rho \end{cases}$$

where, for $i,j \in \{0,1\}$, one denotes $p_{ij} = \mathbb{P}\left(\{X_1 = i\} \cap \{X_2 = j\}\right)$. Denote the (4×4) diagonal matrix $P = \mathrm{diag}(p_{00}, p_{01}, p_{10}, p_{11})$. Any function $G : \{0,1\}^2 \to \mathbb{R}$ can be represented as a vector in $\mathbb{R}^4$, where each element represents a value that G can take w.r.t. the values taken by X. For $i,j \in \{0,1\}$, denote $G_{ij} = G(i,j)$, and thus

$$G = \begin{pmatrix} G_{00} \\ G_{01} \\ G_{10} \\ G_{11} \end{pmatrix},$$

where each G_{ij} can be observed with probability p_{ij}.

4.5.1 Orthocanonical decomposition as solving equations

In this particular case, one can analytically compute the decomposition of G related to Theorem 4.6. It can be performed by finding suitable unit-norm vectors in $\mathbb{R}^4$

$$v_\emptyset = \begin{pmatrix} c \\ c \\ c \\ c \end{pmatrix}, v_1 = \begin{pmatrix} g_0 \\ g_0 \\ g_1 \\ g_1 \end{pmatrix}, v_2 = \begin{pmatrix} h_0 \\ h_1 \\ h_0 \\ h_1 \end{pmatrix}, v_{12} = \begin{pmatrix} k_{00} \\ k_{01} \\ k_{10} \\ k_{11} \end{pmatrix}$$

such that

$$\begin{cases} v_\emptyset^\top P v_1 = 0 \\ v_\emptyset^\top P v_2 = 0 \\ v_\emptyset^\top P v_{12} = 0 \\ v_{12}^\top P v_1 = 0 \\ v_{12}^\top P v_2 = 0 \end{cases}, \quad \text{and}, \quad \begin{cases} v_\emptyset^\top P v_\emptyset = 1 \\ v_1^\top P v_1 = 1 \\ v_2^\top P v_2 = 1 \\ v_{12}^\top P v_{12} = 1 \end{cases} \tag{4.5}$$

which results in a system of nine equations with nine real unknown parameters (i.e., c for $v_\emptyset$, h_0, h_1 for v_1, g_0, g_1 for v_2, and $k_{00}, k_{01}, k_{10}, k_{11}$ for v_{12}). Given these vectors, one has that any function G can be written as

$$G = e v_\emptyset + \alpha v_1 + \beta v_2 + \delta v_{12},$$

resulting in four additional equations with four unknown parameters. These 13 equations and 13 parameters can be computed analytically. The symbolic programming package `sympy` is used to perform these calculations [154]. The interested reader is referred to the accompanying GitHub repository[1], or to Appendix D.4 for details about the computations, and the actual analytical values.

For the purposes of this manuscript, one particular observation is discussed in the following section.

4.5.2 Angle, comonotonicity and definite positiveness of Δ

Dixmier's angle between V_1 and V_2 can be analytically computed regarding this illustration. It is equal to:

$$c\left(\mathbb{L}^2\left(\sigma_1\right), \mathbb{L}^2\left(\sigma_2\right)\right) = c_0\left(V_1, V_2\right) = \left|v_1^\top P v_2\right| = \left|\frac{-q_1 q_2 + \rho}{\sqrt{q_1}\sqrt{q_2}\sqrt{1-q_1}\sqrt{1-q_2}}\right|.$$

Hence, for the Feshchenko matrix Δ to be definite positive (and thus for Assumption 3 to hold), it entails that

$$\left|\frac{-q_1 q_2 + \rho}{\sqrt{q_1}\sqrt{q_2}\sqrt{1-q_1}\sqrt{1-q_2}}\right| < 1$$

which entails that ρ must be bounded by

$$B_0 := \max\left\{0, q_1 q_2\left(1 - \sqrt{\frac{\left(q_1-1\right)\left(q_2-1\right)}{q_1 q_2}}\right)\right\} < \rho < \min\left(1, q_1 q_2\left(1 + \sqrt{\frac{\left\{q_1-1\right\}\left(q_2-1\right)}{q_1 q_2}}\right)\right) := B_1.$$

However, the classical Fréchet bounds for ρ for bivariate Bernoulli random variables (see, [123], p.210) are equal to

$$H_0 := \max\left(0, q_1 + q_2 - 1\right) \leq \rho \leq \min\left(q1, q2\right) := H_1,$$

and notice that these bounds are attained if and only if X is counter-comonotonic or comonotonic. However, attaining these bounds violates Assumption 2 (and in particular Proposition 4.1). However, one can notice that

$$B_0 \leq H_0, \text{ and }, H_1 \leq B_1,$$

which entails that if X is not either counter-comonotonic or comonotonic (and thus Assumption 2 holds), and ρ is strictly contained in the Fréchet bounds, then Δ is will always be definite-positive, and Assumption 3 will hold. In other words, Assumptions 2 and 3 *always hold* for any copula between two Bernoulli random variables, as long as they are strictly contained in the Fréchet-Hoeffding bounds. Hence, the assumptions required for Theorem 4.6 *will virtually always hold* in this particular case.

4.6 Conclusion

In this chapter, a model-centric approach is presented in order to define influence measures. This approach requires the ability to decompose the random output $G(X)$. In the literature, the Hoeffding decomposition [102] allows such a decomposition under the assumption of mutual independence of the inputs. A novel framework has been proposed, relying on tools from probability theory, functional analysis, and combinatorics, which ultimately allowed generalizing Hoeffding's result under two reasonable assumptions. This novel properly gradual decomposition can be expressed using oblique projections of the random output on some particular subspaces, which obey some underlying structure: hierarchical orthogonality. This *orthocanonical decomposition* defines model-centric influence measures for two QoIs: an evaluation (i.e., observation) of the random output and its variance. Aside from the definition of these interpretability tools, some of their properties are studied, and an emphasis is put on their geometric interpretation. Finally, a particular case is studied, where the inputs are composed of two Bernoulli random variables for which the decompositions can be computed analytically.

The first main challenge towards adopting the indices is estimation. While many methods exist to estimate conditional expectations (i.e., the orthogonal projections onto the Lebesgue spaces generated by subsets of inputs), the literature is rather scarce when it comes to the estimation of such oblique projections. Many of these schemes related to conditional expectation estimation rely on the variational problem offered by Hilbert's projection theorem (i.e., orthogonal projections as a distance-minimizing

[1] https://github.com/milidris/phdThesis

problem). A first idea would be to express oblique projections as a distance-minimizing optimization problem under constraints. A second idea would be to take advantage of the particular expression of oblique projections (see, e.g., [2, 46]), which, in our case, would translate, in particular, for every $A \in \mathcal{P}_D$, to

$$\mathbb{M}_A = P_{\mathbb{L}^2(\sigma_A)} \circ \left(P_{\mathbb{L}^2(\sigma_A)} + P_{\widetilde{W}_A} - P_{\widetilde{W}_A} \circ P_{\mathbb{L}^2(\sigma_A)} \right)^{-1},$$

where for a subspace $V \subset \mathbb{L}^2(\sigma_X)$, P_V is the orthogonal projection on V. However, this approach involves estimating the inverse of an operator, which is a challenging feat. A final idea is to find suitable bases for each $(V_A)_{A \in \mathcal{P}_D}$ to project $G(X)$ onto. However, it remains relatively complicated since these subspaces are infinite-dimensional (i.e., the bases are most likely Schauder). Non-orthogonal polynomial bases would be a great start to study this problem whenever X is endowed with a multivariate Gaussian probabilistic structure. When estimating the pure interaction effects, a perspective would be to take inspiration from *importance sampling schemes*, and in particular on copula densities. In the presented framework, copula densities can be used to define an isometric mapping between $\mathbb{L}^2(\sigma_X)$ and $\mathbb{L}^2(\sigma_{\widetilde{X}})$, which would enable to go from the Lebesgue space of X to the Lebesgue space generated by the mutually independent version on X.

The second main challenge is understanding the extent of such an approach. Aside from the uncertainty quantification this framework offers, it is a step towards a more global treatment of dependencies in (non-linear) multivariate statistics. As one can notice, this framework offers a (somewhat surprisingly) linear approach to possibly highly non-linear problems (due to the function G and to the stochastic dependence on X), where Assumptions 2 and 3 will play a pivotal role going forward. The question of the closure of subspaces generated by subsets of inputs is not new (the interested reader is referred to the work of Ivan Feshchenko, see, e.g., [73, 72]), but the approach taken in this chapter does not seem to have been explored for multivariate statistics purposes. Aside from the presented results, the point of view presented in this chapter uncovers an exciting path towards a more complete overview of non-linear multivariate statistics. However, many aspects remain to be mastered, implications to be discovered, and links with existing literature to unveil. Moreover, concerning Assumptions 2 and 3, practical tools need to be developed to statistically assess whether they are respected.

Finally, one can note the importance of the Boolean lattice algebraic structure, which is intrinsically part of the presented developments. Highlighting its role in the presented developments is essential in building a path toward studying different algebraic structures for different analyses. Rota's result is very general (see, Theorem B.1) and does not only apply to Boolean lattices (i.e., powersets). It holds for any (finite) partially ordered set. Studying other algebraic structures can pave the way for more complex analysis, where the relationship between the inputs may differ (e.g., causality). For example, one can think of hierarchical structures (e.g., to represent physical causality) or the presence of trigger variables [174], which would result in a different algebraic structure, but still be partially ordered. More generally, there seems to be a deep link with the statistical field of *graphical models* [136] that needs to be unveiled, where Feshchenko matrices will probably play an important role.

Chapter 5

Robustness to input perturbations

Contents

5.1 Assessing robustness by perturbing inputs . . . 79
5.2 Perturbing quantiles . . . 80
5.2.1 Quantile perturbation classes . . . 80
5.2.2 Two sets of interpretable quantile perturbation classes . . . 82
5.2.3 Copula preservation and marginal perturbations . . . 84
5.3 Wasserstein projections . . . 85
5.3.1 Motivations . . . 85
5.3.2 Marginal quantile constrained Wasserstein projections . . . 86
5.3.3 Relaxed projection problem . . . 86
5.4 Computing the perturbed distributions . . . 87
5.4.1 Analytical solution for the relaxed problem with no smoothing . . . 87
5.4.2 Isotonic piece-wise interpolating polynomial smoothing . . . 87
5.5 Illustration on use-cases . . . 90
5.5.1 Acoustic fire extinguisher: Airflow perturbation . . . 91
5.5.2 River water level: surrogate model validation . . . 94
5.5.3 Conclusions . . . 97
5.6 Discussion . . . 97

Abstract (English).

Robustness studies of black-box models are recognized as necessary for numerical models based on structural equations and predictive models learned from data. These studies must assess the model's robustness to possible misspecification regarding its inputs (e.g., covariate shift). The study of black-box models, through the prism of uncertainty quantification (UQ), is often based on sensitivity analysis involving a probabilistic structure imposed on the inputs. In contrast, ML models are solely constructed from observed data. This work aims to unify the UQ and ML interpretability approaches by providing relevant and easy-to-use tools for both paradigms. To provide a generic and understandable framework for robustness studies, perturbations of the inputs are defined by constraining their quantiles. The Wasserstein distance between probability measures is used to solve the problem while preserving the inputs' dependence structure. It is demonstrated that this perturbation problem can be analytically solved. Ensuring regularity constraints through isotonic polynomial approximations leads to smoother perturbations, which can be more suitable in practice. Numerical experiments on real case studies from the UQ and ML fields highlight the computational feasibility of such studies and provide local and global insights on the robustness of black-box models to input perturbations.

Abstract (Français).

Les études de robustesse des modèles boîte-noire sont utiles dans la validation des modèles numériques basés sur des équations structurelles et les modèles prédictifs appris à partir de données. Ces études doivent évaluer la robustesse du modèle face à une possible mauvaise spécification de ses entrées (par exemple, un décalage covariable). L'analyse des modèles boîte-noire, à travers le prisme de la quantification de l'incertitude (UQ), repose souvent sur une analyse de sensibilité impliquant une structure probabiliste imposée aux entrées, tandis que les modèles d'apprentissage automatique sont construits uniquement à partir de données observées. Ce travail vise à unifier les approches de la UQ et de l'interprétabilité des modèles d'apprentissage automatique en fournissant des outils pertinents et faciles à utiliser pour ces deux paradigmes. Afin d'établir un cadre générique et compréhensible pour les études de robustesse, on propose de perturber les entrées (supposées aléatoires), en s'appuyant sur des contraintes sur leurs quantiles. La distance de Wasserstein entre les mesures de probabilité est ensuite minimisée afin de résoudre ce problème, tout en préservant la structure de dépendance des entrées. Il est démontré que ce problème de perturbation peut être résolu analytiquement. En assurant des contraintes de régularité par le biais d'approximations polynomiales isotones, cela conduit à des perturbations plus lisses, qui peuvent être plus adaptées en pratique. Des expériences numériques sur des études de cas réelles, provenant des domaines de la UQ et de l'apprentissage automatique, mettent en évidence la faisabilité computationnelle de telles études et fournissent des informations locales et globales sur la robustesse des modèles boîte-noire fâce aux perturbations des leurs entrées.

Keywords .

Wasserstein distance • Optimal transport • Probability measure projection • Robustness • Epistemic Uncertainty • Quantiles • Isotonic Polynomials.

This chapter expands on the following contribution.

Pre-prints:

M. Il Idrissi, N. Bousquet, F. Gamboa, B. Iooss, and J. M. Loubes. Quantile-constrained Wasserstein projections for robust interpretability of numerical and machine learning models. Preprint, 2023. URL: https://hal.science/hal-03784768

Conferences:

M. Il Idrissi, N. Bousquet, F. Gamboa, B. Iooss, and J. M. Loubes. Projection de mesures de probabilité sous contraintes de quantile par distance de Wasserstein et approximation monotone polynomiale. In *53èmes Journées de Statistique de la Société Française de Statistique (SFdS)*, Lyon, France, 2022. URL: https://hal.science/hal-03597059

M. Il Idrissi, N. Bousquet, F. Gamboa, B. Iooss, and J. M. Loubes. Robustness assessment of black-box models using quantile-constrained wasserstein projections. In *2022 Annual Meeting of MASCOT-NUM Research Group*, Clermont-Ferrand, France, 2022. URL: https://mascotnum2022.sciencesconf.org/. (Poster)

5.1 Assessing robustness by perturbing inputs

This chapter is dedicated to the *robustness assessment* of a random output through controlled input perturbations. As highlighted in Section 1.3.1, these interpretability methods aim at defining precise perturbations *on the distribution of the inputs*, which in-turn characterize *perturbed inputs*. Then, studying the behavior of the output's QoIs under these perturbations allows answering conundrums of the type:

What are the differences of a black-box model's QoIs induced by a given perturbation of its inputs?

This conundrum entails uncovering a causal link (in the physical sense) between a perturbation and the behavior of the black-box model. Thus, particular care must be put on *the definition of the perturbations*, in order to ensure that the perturbation does not uncontrollably modify the initial distribution (e.g., perturbing the mean of an input implies changing the dependence structure between all the inputs).

The problem of input perturbations is analogous to many frameworks in both the ML field (e.g., domain adaptation [34], covariate shift [93, 213], adversarial attacks [10]) and SA (e.g., distributional sensitivity analysis [5, 162], distributional robustness [141, 86]) or distributional modifications to understand the fairness of algorithms [53, 54]. In the field of financial mathematics, it can also be linked to the notion of *distortion functions*, which are paramount to defining distortion risk measures [15].

In the literature, many methods have been proposed in order to define relevant perturbations (e.g., via geodesics on Fréchet manifolds [86, 127], adversarially [157], using empirical quantiles [13]). However, while generic and automatic, these methods often disregard the physical meaning of these perturbations. To that extent, four desirability criteria are proposed to ensure that the perturbations are *meaningful* to the eyes of domain experts and decision-makers. For instance, perturbations can be used as proxies for epistemic uncertainty, leading to exploratory studies on the behavior of a model induced by a lack of knowledge. Another example would be prospectively designing perturbations to anticipate future changes in the inputs (e.g., climate change). Finally, suppose a gap between some observed data and domain experts' opinions is proven. In that case, perturbations can be modeled to enforce this knowledge while keeping some empirical information gathered on the field.

Formally, given initial inputs X with induced probability measure P_X, and a black-box model $G : E \to Y$, the *input perturbation methodology* can be broken down into three steps:

1. First, define perturbations on the distribution P of the inputs X;
2. Next, find the the perturbed distribution $P_{\widetilde{X}}$ that respect these perturbations, leading to perturbed inputs $\widetilde{X} \sim P_{\widetilde{X}}$;
3. Finally, compare QoIs of the random outputs $G(X)$ and $G(\widetilde{X})$.

Finding suitable perturbed distribution $P_{\widetilde{X}}$ can be generically modeled as the following optimization problem:

$$\begin{aligned} P_{\widetilde{X}} \in \underset{P}{\operatorname{argmin}} \quad & \mathcal{D}\left(P_X, P\right) \\ \text{s.t.} \quad & P \in \mathcal{C}. \end{aligned} \tag{5.1}$$

where $\mathcal{D}$ is a discrepancy between probability measures, and $\mathcal{C}$ is a *perturbation class*, i.e., a particular subset of probability measures that *respect certain constraints*. The framework presented in Section 1.2 is restricted according to the following assumptions.

Special case . In this chapter, the following is assumed concerning the inputs:

- For every $i \in D$, $E_i \subseteq \mathbb{R}$ and thus $E \subseteq \mathbb{R}^d$, i.e., the random inputs are a random vector;
- For every $i \in D$, X_i has a finite variance, i.e.,

$$\mathbb{E}\left[X_i^2\right] = \int_{E_i} x_i^2 dP_{X_i}(x_i) < \infty.$$

In the context of this manuscript, four desirability criteria are of interest and are detailed as follows:

Interpretability The perturbations should be meaningful to domain experts and decision-makers. It ensures that well-understood phenomena induce the uncovered variations in the model's behavior. Hence, designing perturbations should be done with practitioners and precisely reflect a domain-specific question. In-fine, perturbation interpretability ensures that the (physical) causal link one aims to draw of a perturbation on the behavior of a model is insightful on the question at stake.

Genericity The perturbations should be generic because they should not depend on restrictive properties assumed to hold for G (e.g., continuity, derivability) or P_X (e.g., absolute continuity). Genericity ensures the proposed methodology is *post-hoc* [16]. To emphasize the duality between SA and ML interpretability [182, 119], generic perturbation ensures that the proposed methodology is usable in both settings.

Proximity The perturbed distribution should be "close" to the initial distribution P_X. Proximity ensures that the perturbed distribution remains somewhat similar to the initial, where similarity needs to be measured through a discrepancy. For instance, closeness in the KL divergence sense entails similar information, whereas closeness in the Wasserstein distance sense has a more geometric meaning. Either way, the initial distribution, be it empirical or chosen, bears some information on the behavior of the input, which needs to be partially preserved.

Exploration The perturbation scheme should allow for exploring unobserved or low probability regions of $\mathcal{X}$. This criterion ensures that "out of distribution" scenarios can be reached. Hence, the model's behavior can be assessed on "unusual" (for P_X) evaluations, which is crucial when testing for robustness.

The choices of discrepancy and the definition of perturbation classes explored in this chapter, and motivated by the desirability criteria listed above, are the following:

- The 2-Wasserstein distance as a discrepancy between probability measures to ensure genericity and exploration;
- Perturbation classes $\mathcal{C}$ based on three types of constraints:
  - Interpolation constraints on generalized quantile functions to ensure interpretability and genericity;
  - Smoothness of the generalized quantile functions to ensure exploration;
  - Copula-preservation to ensure interpretability.

Some preliminary notations In the following, denote by $\mathcal{P}(\mathbb{R}^d)$ the set of probability measures defined on $\mathbb{R}^d$ and, for a positive integer p, denote by $\mathcal{P}_p(\mathbb{R})$ the set of probability measures defined on $\mathbb{R}$ with finite p-th moment. For every univariate input $X_i, i \in D$, denote by P_i its induces probability measure, and let $\Omega_{X_i} \subset \mathbb{R}$ be its *application domain*. It represents the range in which X_i is intended to vary in practice [189] (see, Appendix E.1.1). Additionally, denote by $F_{P_i}(t)$ the *cumulative distribution function* (cdf) of X_i, and by $F_{P_i}^{\leftarrow}$ and $F_{P_i}^{\rightarrow}$ its *generalized quantile function* (gqf) and the *right-continuous generalized inverse of* F_{P_i} (see, Definition E.3). More generally, denote by $\mathcal{F}$ the space of distribution functions (see, Definition E.2), and $\mathcal{F}^{\rightarrow}$ the space of generalized quantile functions (see, Appendix E.1.2).

5.2 Perturbing quantiles

5.2.1 Quantile perturbation classes

Motivations First, for any univariate probability measure $P_i \in \mathcal{P}(\mathbb{R})$ induced by an input X_i, *its gqf* $F_{P_i}^{\leftarrow}$ *always exists*. Hence, perturbing marginal quantiles do not require additional assumptions on the initial probability measure P_X or the shape of the target perturbed probability measure $P_{\tilde{X}}$. It ensures that the proposed methodology is generic, in contrast to the one proposed in [140] based on generalized moments.

Second, *quantiles are interpretable*. In many applied problems, quantile specifications are often key to studying the influence of input variables on a decision-making output. Beyond the fact that quantiles have a decision-theoretical sense through pinball cost functions [41], numerous applications dealing with economic stress tests or risk mitigation against natural hazards use quantiles as influential inputs of decision-helping models. For instance, in the drought risk studies in [65], the association between soil wetness, climatic, seismic, and socioeconomic variables is often carried out using marginal quantiles that are features for predictive cost models. Input variations of daily value-at-risk percentiles, computed from legacy data, were recently required by the European Banking Authority for generating macroeconomic scenarios used for EU-wide stress tests [11]. Reverse SA studies for financial risk management, such as those conducted in [176], are primarily based on moving values-at-risk, which are quantiles.

The following examples offer additional concrete illustrations of using quantiles for influence analysis. They highlight two quantile perturbation schemes: quantile shifting and application domain dilatation.

Example 5.1 (Economic stress test (Inspired by [27])). Assume that an economic shock happens in an abstract country. Prospective analyses announce a \$200 drop in the population median wage. Before the shock, the population wage distribution P_X is known (or observed), thanks to some annual census data. This distribution has a median wage of \$2000. The new population wage distribution is unknown due to the lack of recent data. The economists want to know if they can be confident in their predictive macro-economic model f w.r.t. this sudden change. A way to answer this problem would be assessing the behavior of the model f on a distribution $P_{\widetilde{X}}$, such that:

$$F^{\leftarrow}_{P_{\widetilde{X}}}(0.5) = 1800.$$

Example 5.2 (River water level). This example is inspired from [120] and more deeply studied in Section 5.5.2. The safety of an industrial site located near a river is studied through the prediction of the water level $G(X)$ where G is a numerical hydrodynamic model, and X gathers the physical features of the river. A key dimension of X is the Strickler roughness coefficient K_s for the upstream water level [82], which is modeled as a truncated Gaussian distribution on $\Omega_X = [15, 55]$. However, this application domain is tainted with epistemic uncertainties on the actual nature of the riverbed (e.g., more or less subject to shrubby vegetation). The practical use of G would require assessing its predictive power under a wider interval $\Omega_X = [5, 65]$. A way to express this prospective study is to assess the model's behavior on a distribution $P_{\widetilde{X}}$, such that:

$$F^{\rightarrow}_{K_s}(0) = 5, \quad F^{\leftarrow}_{K_s}(1) = 65.$$

Formal Definition Since, for a fixed $\alpha \in [0, 1]$, α-quantiles are not necessarily unique, equality constraints on quantile functions seem somewhat arbitrary (see, Appendix E.1.2). It would amount to constraining the infimum of the set of α-quantiles. Arguably, given a desired α-quantile value of $b \in \mathbb{R}$, a more reasonable constraint would be for b to be in the *set of α-quantiles* of the perturbed distribution. Formally, it amounts to for constraints of the type:

$$F^{\leftarrow}(\alpha) \geq b \geq F^{\leftarrow}(\alpha^+) = F^{\rightarrow}(\alpha), \tag{5.2}$$

In other words, the perturbed univariate distribution should have b as one of its α-quantile values. In the case where the perturbed cdf is invertible, it becomes a traditional equality constraint as α-quantiles become uniquely defined (i.e., $F^{\leftarrow}(\alpha) = F^{\rightarrow}(\alpha)$). In the following, the inequality constraints defined in Eq. (5.2) are referred to as *quantile constraints*.

Definition 5.1 (Quantile perturbation class). Let K be a positive integer, and let $\alpha = (\alpha_1, \ldots, \alpha_K)^\top \in [0, 1]^K$ and $b = (b_1, \ldots, b_K)^\top \in \mathbb{R}^K$. The *quantile perturbation class* $\mathcal{Q}(\alpha, b) \subseteq \mathcal{P}(\mathbb{R})$ is the set of probability measures defined as:

$$\mathcal{Q}(\alpha, b) = \left\{ Q \in \mathcal{P}(\mathbb{R}) \ : \ F^{\leftarrow}_Q(\alpha_i) \leq b_i \leq F^{\rightarrow}_Q(\alpha_i), \quad i = 1, \ldots, K \right\}.$$

An equivalent characterization, thanks to the uniqueness of gqfs, is:

$$\mathcal{Q}(\alpha, b) = \left\{ Q \in \mathcal{P}(\mathbb{R}) \ : \ F^{\leftarrow}_Q = L \in \mathcal{F}^{\leftarrow}, L(\alpha_i) \leq b_i \leq L(\alpha_i^+), \quad i = 1, \ldots, K \right\}.$$

It is possible to derive sufficient conditions on α and b in order for $\mathcal{Q}(\alpha, b)$ to be non-empty:

Lemma 5.1. *Let $\alpha \in [0,1]^K$ and $b \in \mathbb{R}^K$, which are assumed to be ordered without loss of generality. If*

$$0 \leq \alpha_1 < \cdots < \alpha_K \leq 1, \quad \text{and} \quad b_1 < \cdots < b_K, \tag{5.3}$$

then $\mathcal{Q}(\alpha, b)$ is non-empty.

Proof of Lemma 5.1 on p.152.

Quantile perturbation classes can contain probability measures with discontinuous gqfs. Ensuring smooth perturbed gqfs can be of practical interest. It entails further restricting the gqfs of the probability measures in a quantile perturbation class to respect some smoothness conditions. They can be formally defined as follows.

Definition 5.2 (Smooth quantile perturbation class). Let K be a positive integer, $\alpha = (\alpha_1, \ldots, \alpha_K)^\top \in [0,1]^K$, $b = (b_1, \ldots, b_K)^\top \in \mathbb{R}^K$ and let $\mathcal{V} \subseteq \mathcal{F}^{\leftarrow}$ be a given set of smooth non-decreasing functions. The *smooth quantile perturbation class* $\mathcal{Q}_{\mathcal{V}}(\alpha, b) \subseteq \mathcal{P}(\mathbb{R})$ is the set of probability measures defined as:

$$\mathcal{Q}_{\mathcal{V}}(\alpha, b) = \left\{ Q \in \mathcal{P}(\mathbb{R}) \ : \ F_Q^{\leftarrow} \in \mathcal{V}, F_Q^{\leftarrow}(\alpha_i) \leq b_i \leq F_Q^{\rightarrow}(\alpha_i), \quad i = 1, \ldots, K \right\}.$$

Note that smooth perturbation classes generalize perturbation classes since $\mathcal{Q} = \mathcal{Q}_{\mathcal{F}^{\leftarrow}}$.

5.2.2 Two sets of interpretable quantile perturbation classes

Two sets of quantile perturbation classes are introduced: quantile shifts and application domain dilatation.

Quantile shifts. Quantile shifts are defined by constraining an initial α-quantile to take values in a pre-determined range. Formally, given a quantile level $\alpha \in [0,1]$, and an initial α-quantile $p_\alpha = F_P^{\leftarrow}(\alpha)$, quantile shifts defines a set of quantile perturbations classes of probability measures having their α-quantiles ranging over a compact interval $[\eta_0, \eta_1] \subseteq \Omega_X$ such that $\eta_0 < p_\alpha < \eta_1$. In other words, for each $b_\alpha \in [\eta_0, \eta_1]$, a quantile perturbation class $\mathcal{Q}_{\mathcal{V}}(\alpha, b_\alpha)$ can be constructed. This particular type of set of quantile perturbation classes can be described using a *perturbation intensity* $\theta \in [-1, 1]$:

Lemma 5.2. *Let $\Theta = [-1, 1]$ and denote $\boldsymbol{\eta} = (\eta_0, \eta_1)$ with $\eta_0 < p_\alpha < \eta_1$. For $\theta \in \Theta$, let,*

$$b_\alpha(\boldsymbol{\eta}, \theta) = \begin{cases} p_\alpha(1+\theta) - \theta\eta_0 & \text{if } -1 \leq \theta < 0, \\ p_\alpha & \text{if } \theta = 0, \\ p_\alpha(1-\theta) + \theta\eta_1 & \text{if } 0 < \theta \leq 1. \end{cases}$$

Then, for any gqf $F^{\leftarrow} \in \mathcal{F}^{\leftarrow}$ such that

$$F^{\leftarrow}(\alpha) \geq b_\alpha(\boldsymbol{\eta}, \theta) \geq F^{\rightarrow}(\alpha),$$

one has that,

$$\begin{aligned} \theta = -1 &\Leftrightarrow b_\alpha(\boldsymbol{\eta}, \theta) = \eta_0, \\ \theta = 0 &\Leftrightarrow b_\alpha(\boldsymbol{\eta}, \theta) = p_\alpha, \\ \theta = 1 &\Leftrightarrow b_\alpha(\boldsymbol{\eta}, \theta) = \eta_1, \end{aligned} \tag{5.4}$$

and for any $-1 \leq \theta_1 < \theta_2 \leq 1$,

$$b_\alpha(\boldsymbol{\eta}, \theta_1) < b_\alpha(\boldsymbol{\eta}, \theta_2).$$

Proof of Lemma 5.2 on p.152.

In other words, $b_\alpha(\boldsymbol{\eta}, \theta) \in [\eta_0, \eta_1]$ is a strictly increasing function of θ and $\theta = 0$ indicates that p_α must remain untouched (i.e., the quantile value is not perturbed). Figure 5.1 (a.) illustrates this perturbation scheme. Quantile shifts are formally defined as the collection of perturbation classes $\{\mathcal{T}(\boldsymbol{\eta}, \theta)\}_{\theta \in [-1,1]}$

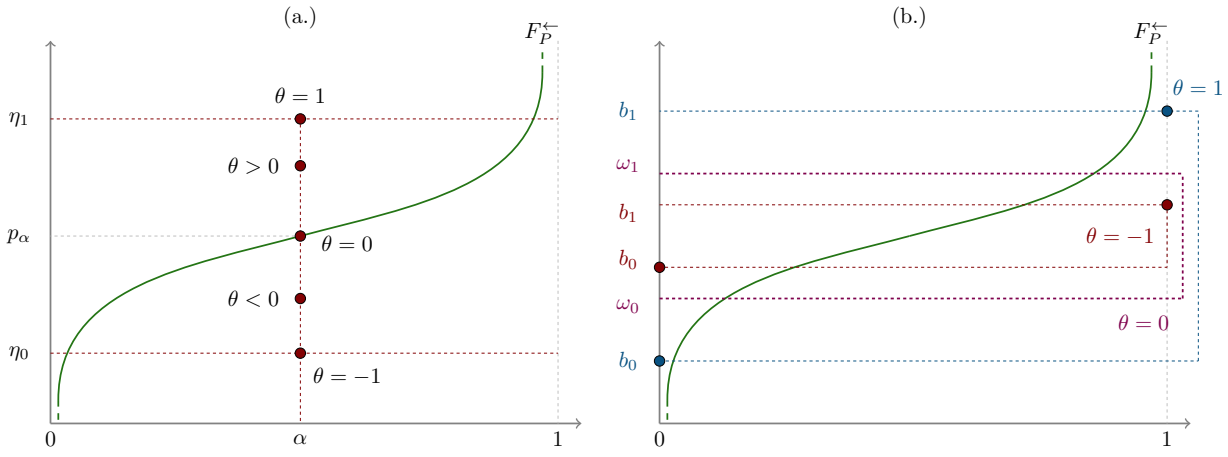

Figure 5.1: Quantile shift (a.) and application domain dilatation (b.) perturbation schemes. The initial quantile function is displayed in green. On the left, red points indicate different quantile shifting constraints between η_0 and η_1, leading to different intensity values θ. On the right, the application domain's width (in magenta) is up to doubled (blue points) or down to halved (red points), according to an intensity parameter $\theta \in [-1, 1]$.

where,

$$\begin{aligned}\mathcal{T}(\boldsymbol{\eta}, \theta) &= \left\{Q \in \mathcal{P}(\mathbb{R}) \ : \ F_Q^{\leftarrow}(\alpha) \leq b_\alpha(\boldsymbol{\eta}, \theta) \leq F_Q^{\rightarrow}(\alpha)\right\} \\ &= \mathcal{Q}\left(\alpha, b_\alpha(\boldsymbol{\eta}, \theta)\right)\end{aligned} \tag{5.5}$$

Application domain dilatations. *Application domain dilatations* consists of perturbing the bounds of the application domain of a marginal input. For a univariate $X \sim P$ with $\Omega_X = [\omega_0, \omega_1]$, the dilatation process amounts to widening or narrowing the width (or diameter $\text{diam}(\Omega_X)$) of Ω_X. It amounts constraining the extreme quantiles (i.e., $\alpha \in \{0, 1\}$) while preserving the midpoint of Ω_X. The dilatation is characterized by a parameter $\eta > 1$ controlling the rescaling magnitude of Ω_X. In other words, one aims at finding a perturbed distribution $P_{\widetilde{X}}$ with support $\text{Supp}\left(P_{\widetilde{X}}\right) = [b_0, b_1]$ for $b_0, b_1 \in \mathbb{R}$, $b_0 < b_1$, where the midpoint of $[b_0, b_1]$ is equal to the midpoint of Ω_X, but such that $\text{diam}\left(P_{\widetilde{X}}\right) := \text{diam}\left(\text{Supp}\left(P_{\widetilde{X}}\right)\right)$ is rescaled compared to $\text{diam}\left(\Omega_X\right)$. Similarly to quantile shift, the next lemma formalizes expressions for these two bounds as a function of a perturbation intensity $\theta \in [-1, 1]$.

Lemma 5.3. *Let $\eta > 1$. For $\theta \in [-1, 1]$, let:*

$$b_0(\eta, \theta) = \begin{cases} \frac{1}{2}\left(\omega_0\big(2 - \theta(\eta^{-1} - 1)\big) + \theta\omega_1(\eta^{-1} - 1)\right) & \text{if } -1 \leq \theta < 0, \\ \omega_0 & \text{if } \theta = 0, \\ \frac{1}{2}\left(\omega_0\big(2 + \theta(\eta - 1)\big) - \theta\omega_1(\eta - 1)\right) & \text{if } 0 < \theta \leq 1, \end{cases}$$

$$b_1(\eta, \theta) = \begin{cases} \frac{1}{2}\left(\omega_1\big(2 - \theta(\eta^{-1} - 1)\big) + \theta\omega_0(\eta^{-1} - 1)\right) & \text{if } -1 \leq \theta < 0, \\ \omega_1 & \text{if } \theta = 0, \\ \frac{1}{2}\left(\omega_1\big(2 + \theta(\eta - 1)\big) - \theta\omega_0(\eta - 1)\right) & \text{if } 0 < \theta \leq 1. \end{cases}$$

Then, for every $(\theta, \eta) \in [-1, 1] \times [1, \infty)$,

$$b_0(\eta, \theta) + b_1(\eta, \theta) \quad = \quad \omega_0 + \omega_1 \quad \textit{(midpoints equality)}.$$

Denote $\boldsymbol{b}(\eta, \theta) = [b_0(\eta, \theta), b_1(\eta, \theta)]$, and notice that

$$\begin{aligned} \theta = -1 &\Leftrightarrow \textit{diam}\left(\boldsymbol{b}(\eta, \theta)\right) = \frac{\textit{diam}(\Omega_X)}{\eta}, \\ \theta = 0 &\Leftrightarrow \textit{diam}\left(\boldsymbol{b}(\eta, \theta)\right) = \textit{diam}(\Omega_X), \\ \theta = 1 &\Leftrightarrow \textit{diam}\left(\boldsymbol{b}(\eta, \theta)\right) = \eta\textit{diam}(\Omega_X), \end{aligned} \tag{5.6}$$

and for any $-1 \leq \theta_1 < \theta_2 \leq 1$,

$$\textit{diam}\left(\boldsymbol{b}(\eta, \theta_1)\right) < \textit{diam}\left(\boldsymbol{b}(\eta, \theta_2)\right).$$

Proof of Lemma 5.3 on p.152.

In other words, $\text{diam}\left(\boldsymbol{b}(\eta, \theta)\right) \in \left[\eta^{-1}\text{diam}(\Omega_X), \eta\text{diam}(\Omega_X)\right]$ is a strictly increasing function of θ, and for $\theta = 0$, one has that $\boldsymbol{b}(\eta, \theta) = \Omega_X$, i.e., the application domain is not perturbed.

Figure 5.1 (b.) illustrates this perturbation scheme. The initial application domain is displayed in magenta and is subject to a dilatation of parameter $\eta = 2$. The red constraints halve its width, and the blue constraints double it. One can additionally check that in both cases, the midpoint of the original validity domain is preserved. Application domain dilatations are formally defined as the collection of perturbation classes $\{\mathcal{T}(\eta, \theta)\}_{\theta \in [-1,1]}$ where,

$$\begin{aligned} \mathcal{T}(\boldsymbol{\eta}, \theta) &= \left\{Q \in \mathcal{P}(\mathbb{R}) \ : \ F_Q^{\leftarrow}(m) \leq b_m(\eta, \theta) \leq F_Q^{\rightarrow}(m), m \in \{0, 1\}\right\} \\ &= \mathcal{Q}\left((0,1)^\top, (b_0(\eta, \theta), b_1(\eta, \theta))^\top\right) \end{aligned} \tag{5.7}$$

Many perturbation settings can be defined by combining quantile shifts and domain dilatations. However, for the sake of simplicity, quantile shifts and domain dilatations are studied independently in Section 5.5.

5.2.3 Copula preservation and marginal perturbations

Motivations Regarding multivariate perturbations in general, independence assumptions are often required [141]. While it facilitates mathematical calculations, it is questionable in practice. One of the main challenges in ML interpretability and SA is to account for the potential dependence structure between the inputs (or features) [177].

Dependencies provide helpful information on the global behavior of the inputs. In SA, the dependence structure is often chosen after extensive studies [57], and expresses the physical relationship between the uncertainties on the inputs. In ML, it can be argued that preserving dependencies avoids creating meaningless patterns [22] and is critical in some practical studies [145, 173]. Dependencies between random variables are usually modeled using copula-based representations [164].

From the *interpretability* standpoint, in practice, the intricacies of multivariate insights due to stochastic dependence are much more complicated to grasp. Moreover, many of the properties presented above do not hold regarding multivariate quantile functions: The definition of multivariate quantile functions is a highly non-trivial task. Many interesting approaches have been recently proposed [40, 94]. However, they lack the broad adoption of their univariate counterpart in practice, which makes them less interpretable.

Thus, in order to ensure the *interpretability*, the proposed perturbation methodology is restricted to:

- Quantile perturbations on marginal inputs.
- Perturbed probability measures $P_{\widetilde{X}}$ having the same copula as the initial probability measure P_X.

Marginal perturbation maps and copula preservation Let $X \sim P$ and for $i = 1, \ldots, d$, let each marginal input $X_i \sim P_i$ and $(F_i^{\leftarrow})_{i=1,\ldots,d}$ be a collection of quantile functions in $\mathcal{F}^{\leftarrow}$. A *marginal perturbation map* is a mapping:

$$\begin{array}{rccc} T: & \mathcal{X} & \to & \mathcal{X} \\ & \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix} & \mapsto & \begin{pmatrix} T_1(x_1) \\ \vdots \\ T_d(x_d) \end{pmatrix} \end{array} \tag{5.8}$$

where

$$T_j = \left[F_j^{\leftarrow} \circ F_{P_j}\right], \quad j = 1, \ldots, d.$$

The perturbed inputs can be expressed as $\widetilde{X} := T(X)$.

Lemma 5.4. *Suppose that each $F_i^{\leftarrow}$, $i = 1, \dots, d$ is strictly increasing:*

(i) If P is an empirical measure then X and $\widetilde{X}$ have the same empirical copula.

(ii) If P is atomless then X and $\widetilde{X}$ have the same copula.

Proof of Lemma 5.4 on p.153.

Hence, perturbation maps composed of compositions of marginal cdfs and strictly increasing quantile functions preserve the copula. For instance, if P is an empirical measure related to an observed dataset, applying T to every observation results in a perturbed dataset with the same Spearman correlation matrix.

Copula-preserving multivariate perturbation classes Combining quantile perturbation classes with marginal perturbation maps allows for defining multivariate perturbation classes. Let $X \sim P$, and for $i = 1, \dots, d$, let $\theta_i \in [0,1]^K \times \mathbb{R}^K$ and $\theta = (\theta_1, \dots, \theta_d)$. Finally, let $\mathcal{Q}^{(i)} := \mathcal{Q}(\theta_i)$ be the perturbation class associated with the input X_i. For $Q \in \mathcal{P}(\mathbb{R}^d)$, and denote $Q_1, \dots, Q_d$ its marginal distributions. Denote the set:

$$\mathcal{Q}_d(\theta) = \left\{ Q \in \mathcal{P}(\mathbb{R}^d) \ : \ Q_i \in \mathcal{Q}^{(i)} \right\},$$

and for any $Q \in \mathcal{P}(\mathbb{R}^d)$, denote T_Q the marginal perturbation map defined as:

$$\begin{array}{rccc} T_Q : & \mathcal{X} & \to & \mathcal{X} \\ & \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix} & \mapsto & \begin{pmatrix} \left[F_{Q_1}^{\leftarrow} \circ F_{P_1}\right](x_1) \\ \vdots \\ \left[F_{Q_d}^{\leftarrow} \circ F_{P_d}\right](x_d) \end{pmatrix} \end{array} \tag{5.9}$$

Marginal quantile perturbation classes are defined as the set:

$$\mathcal{Z}(P, \theta) = \{ Q \in \mathcal{Q}_d(\theta) \ : \ T_Q(X) \sim Q, X \sim P \},$$

and, from Lemma 5.4, *copula-preserving marginal quantile perturbation classes* are defined as:

$$\widetilde{\mathcal{Z}}(P, \theta) = \left\{ Q \in \mathcal{Z}(P, \theta) \ : \ F_{Q_i}^{\leftarrow} \text{ is strictly increasing}, i = 1, \dots, d \right\}.$$

5.3 Wasserstein projections

5.3.1 Motivations

The Wasserstein distance is deeply rooted in optimal transportation theory [219] and has been used successfully in many ML and deep learning applications [80, 8]. It has also been extensively studied as a tool for guaranteeing distributional robustness to adversarial attacks in ML [62]. It has been used in SA to produce novel sensitivity indices [75, 28].

First, the 2-Wasserstein distance is *interpretable*. The choice of transportation cost as the squared distance is intrinsically linked to notions of the L^2 norms, which can be interpreted as lengths, analogous to the well-known Euclidean geometry [219]. It becomes natural and intuitive to quantify transportation costs as distances between points. It becomes even more natural in one dimension since the 2-Wasserstein distance can be interpreted as the squared difference in areas between two quantile functions. Hence, *proximity* between two univariate probability measures, in the 2-Wasserstein sense, is rather natural.

Moreover, the 2-Wasserstein distance ensures *genericity*. The only requirement for two probability measures to be comparable is the finiteness of the variance of the random variable they induce. This assumption is classical in SA and ML interpretability. Compared to the KL divergence, which requires the absolute continuity of one probability measure versus the other and the existence of logarithmic moments, it appears less restrictive. In practice, it allows for more flexible perturbations: if P is an empirical measure (i.e., purely atomic), then Q is not restricted to be purely atomic; conversely, if P admits a density, then it does not restrict Q to admit a density. These benefits are key in unifying the frameworks of SA and ML interpretability: the flexibility of the 2-Wasserstein distance allows for greater explicit

control (e.g., through smoothing restriction) on the resulting perturbed measure Q, independently of the properties of P.

Additionally, the 2-Wasserstein distance allows for *exploration*. Optimal transport maps between two probability measures w.r.t. the 2-Wasserstein distance are (usually) not linear [194]. In other words, perturbed solutions are not limited to the support of the initial probability measures: atoms can be added and ranges with 0 probability can be made relevant.

5.3.2 Marginal quantile constrained Wasserstein projections

The problem of finding a probability measure Q closest to P, but $Q \in \widetilde{Z}(P,\theta)$ can be formalized as follows:

$$Q = \underset{G \in \mathcal{P}_2(\mathbb{R}^d)}{\operatorname{argmin}} W_2^2(P,G) \quad \text{s.t.} \quad G \in \widetilde{Z}(P,\theta) \tag{5.10}$$

However, since the set of probability measures in $\widetilde{Z}(P,\theta)$ share the same copula as P, this problem can be simplified:

Lemma 5.5. *The perturbation map $T : \mathbb{R}^d \to \mathbb{R}^d$ that minimizes (5.10) is defined, for any $x = (x_1,\dots,x_d)^\top \in \mathbb{R}^d$, as:*

$$T(x) = \begin{pmatrix} \left[F^{\leftarrow}_{Q_1} \circ F_{P_1}\right](x_1) \\ \vdots \\ \left[F^{\leftarrow}_{Q_d} \circ F_{P_d}\right](x_d) \end{pmatrix}$$

where, for $i = 1,\dots,d$:

$$\begin{aligned} F^{\leftarrow}_{Q_i} = \underset{L \in L^2([0,1])}{\operatorname{argmin}} & \left\{ \int_0^1 \left(L(x) - F^{\rightarrow}_{P_i}(x)\right)^2 dx \right\} \\ \text{s.t.} \quad & L(\alpha_j) \le b_j \le L\left(\alpha_j^+\right), \quad i = 1,\dots,K, \\ & L \text{ is strictly increasing.} \end{aligned} \tag{5.11}$$

where for $\alpha = (\alpha_1,\dots,\alpha_k)^\top$, $b = (b_1,\dots,b_k)^\top$, $\theta_i = (\alpha, b)$.

Proof of Lemma 5.5 on p.153.

Hence, solving the projection problem in (5.10) is equivalent to solving the d problems of the form of (5.11).

5.3.3 Relaxed projection problem

Imposing that the resulting optimally perturbed marginal gqf be strictly increasing guarantees preserving the initial copula of the inputs. However, such constraints can lead to the non-existence of an optimum of (5.11) due to the non-closure of the set of strictly increasing functions [25]. To that extent, this work focuses on a relaxation of the problem in (5.11) to increasing functions, namely:

$$\begin{aligned} F^{\leftarrow} = \underset{L \in L^2([0,1])}{\operatorname{argmin}} & \left\{ \int_0^1 \left(L(x) - F^{\rightarrow}_{P_i}(x)\right)^2 dx \right\} \\ \text{s.t.} \quad & L(\alpha_j) \le b_j \le L\left(\alpha_j^+\right), \quad i = 1,\dots,K, \\ & L \in \mathcal{V} \subseteq \mathcal{F}^{\leftarrow}. \end{aligned} \tag{5.12}$$

where $\mathcal{V}$ can be understood as a set of "smooth quantile functions" (see, Definition 5.2). Notice that this problem is indeed a relaxation of the initial problem. Indeed, if $\mathcal{V}$ is chosen as the set of strictly increasing functions, this problem becomes equivalent to Eq. (5.11).

Remark 5.1. In practice, the relaxed problem (5.12) is numerically more straightforward to solve and can still lead to strictly increasing solutions.

5.4 Computing the perturbed distributions

5.4.1 Analytical solution for the relaxed problem with no smoothing

The following proposition provides a convenient way to solve the perturbation problem (5.12) in the particular case of $\mathcal{V} = \mathcal{F}^{\leftarrow}$.

Proposition 5.1. *Let P be a probability measure in $\mathcal{P}_2(\mathbb{R})$. Let $\alpha \in [0,1]^K$ and $b \in \mathbb{R}^k$, such that $\alpha_1 < \cdots < \alpha_K$ and $b_1 < \cdots < b_K$, and $\mathcal{Q}(\alpha, b)$ the associated quantile perturbation class. For $i = 1, \ldots, K$, let $\beta_i = F_P(b_i)$. Define the intervals $A_i = (c_i, d_i]$ for $i = 1, \ldots, K$, such that:*

$$c_1 = \min(\beta_1, \alpha_1), \quad c_i = \min\Big[\max(\alpha_{i-1}, \beta_i), \alpha_i\Big], i = 2, \ldots, K,$$
$$d_K = \max(\beta_K, \alpha_K), \quad d_j = \max\Big[\min(\beta_j, \alpha_{j+1}), \alpha_j\Big], j = 1, \ldots, K-1.$$

Let $A = \bigcup_{i=1}^K A_i$ and $\overline{A} = [0,1] \setminus A$. Then the problem (5.12) where $\mathcal{V} = \mathcal{F}^{\leftarrow}$ has a unique solution which can be written as, for any $y \in [0,1]$:

$$F_Q^{\leftarrow}(y) = \begin{cases} F_P^{\rightarrow}(y) & \text{if } y \in \overline{A}, \\ b_i & \text{if } y \in A_i, \quad i = 1, \ldots, K. \end{cases} \tag{5.13}$$

Proof of Proposition 5.1 on p.153.

In order to interpret this result, illustrated in Figure 5.2, let us recall that when a quantile function is constant on an interval, it implies that the associated probability measure admits an atom at the value taken by the gqf on this interval. Moreover, the mass allocated to this atom is equal to the length of the interval. Additionally, each jump of the quantile function induces an interval with no mass. The solution displayed in (5.13) shows that both initial and perturbed quantile functions are equal on $\overline{A}$. However, they differ on every interval A_i in the following fashion:

- Q have atoms at each constraint point b_i, $i = 1, \ldots, K$;
- Each of these atoms have mass $Q(\{b_i\}) = d_i - c_i$, for $i = 1, \ldots, K$;
- Each open interval $I_i \subset \mathbb{R}$ defined as

$$I_i = \begin{cases} \Big(\max(F_P^{\leftarrow}(\alpha_i), b_{i-1}), b_i\Big), & \text{when } b_i > F_P^{\leftarrow}(\alpha_i), \\ \Big(b_i, \min\left(b_{i+1}, F_P^{\leftarrow}(\alpha_i)\right)\Big), & \text{when } b_i < F_P^{\leftarrow}(\alpha_i) \end{cases} \tag{5.14}$$

  with, by convention, $b_0 = -\infty$ and $b_{K+1} = \infty$, has no mass. To put it briefly, $Q(I_i) = 0$ for every $i = 1, \ldots, K$.

In other words, whenever an α-quantile p_α is shifted up to a value b, the perturbation entails sending every possible value in the range (p_α, b) to b. Hence, every value in (p_α, b) cannot be sampled according to Q. Moreover, the singleton $\{b\}$ now admits a probability of being observed equal to the initial probability of this interval, i.e., $Q(\{b\}) = P\big((p_\alpha, b)\big)$. When an α-quantile is shifted to b, the interval becomes (b, p_α), and the same reasoning can be done.

The analytical result of Proposition 5.1 is rather intuitive. Indeed, the Wasserstein distance quantifies the amount of *work* needed to transform a probability measure into another one [194]. When using W_2, the amount of work is quantified using the Euclidean distance, i.e., transporting a point x_0 to x_1 requires $(x_0 - x_1)^2$ work units. This intrinsic "point-wise way of quantifying similarities" can be recovered in Proposition 5.1: perturbing an α-quantile entails giving the initial mass of an interval adjacent to b to the singleton $\{b\}$ in order to satisfy the constraint.

5.4.2 Isotonic piece-wise interpolating polynomial smoothing

The analytical solution provided in Proposition 5.1 presents a significant drawback: part of the application domain of the perturbed input receives no mass, which hurts the *exploration* criteria. This is because

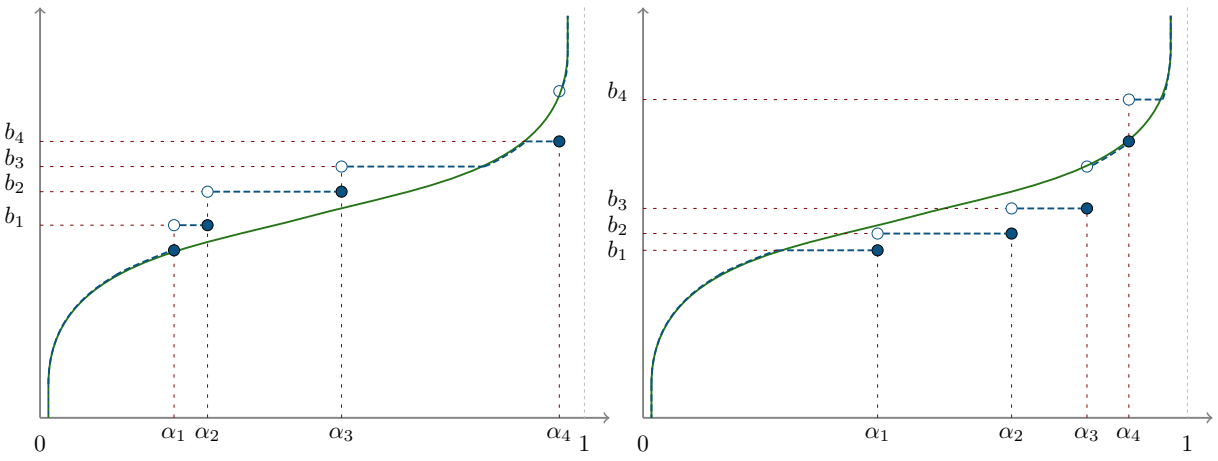

Figure 5.2: Characterizing quantile function of the solution of the perturbation problem (dashed blue). The initial quantile function (i.e., $F_P^{\leftarrow}$) is displayed in green, and dashed red lines identify the quantile constraints. (a.) and (b.) illustrate different possible perturbation configurations, increasing or decreasing several initial quantile values.

$\mathcal{F}^{\leftarrow}$ contains discontinuous functions. Ensuring continuity through a smooth perturbation class $\mathcal{Q}_{\mathcal{V}}$ where $\mathcal{V}$ is a set of continuous, non-decreasing functions can remove this issue.

This section studies the projection of the gqf $F_P^{\leftarrow}$ of a univariate input onto a space of piece-wise continuous polynomials. It implies that the support of Q must be bounded. These bounds are made explicit using extremal quantile constraints (i.e., $F_Q^{\leftarrow}(0)$ and $F_Q^{\leftarrow}(1)$ are constrained to take finite values). Formally, the goal is to find a piece-wise polynomial of the form

$$G(x) = \begin{cases} G_0(x) & \text{if } \alpha_0 := 0 \leq x < \alpha_1, \\ \vdots & \\ G_i(x) & \text{if } \alpha_i \leq x < \alpha_{i+1}, \\ \vdots & \\ G_K(x) & \text{if } \alpha_K \leq x \leq 1 =: \alpha_{K+1}. \end{cases} \tag{5.15}$$

under the continuity constraints at each knot on the grid $\alpha_1 < \cdots < \alpha_K$, i.e.,

$$G_i(\alpha_{i+1}) = G_{i+1}(\alpha_{i+1}), \quad i = 0, \ldots, K-1.$$

Here, each $G_j \in \mathbb{R}[x]_{\leq p}$, for $j = 0, \ldots, K$, where $\mathbb{R}[x]_{\leq p}$ denotes the set of all real polynomials of degree at most equal to a nonnegative integer p. Let $\mathcal{S}_p$ denote the space of functions defined by Eq. (5.15). Restricting the solution of the perturbation problem in Eq. (5.12) leads to the following optimization problem

$$\begin{aligned} F_Q^{\leftarrow} = \underset{L \in L^2([0,1])}{\operatorname{argmin}} & \left\{ \int_0^1 \left(L(x) - F_P^{\rightarrow}(x)\right)^2 dx \right\} \\ \text{s.t.} \quad & L(\alpha_i) = b_i, \quad i = 1, \ldots, K, \\ & L \in \mathcal{F}^{\leftarrow} \cap \mathcal{S}_p. \end{aligned} \tag{5.16}$$

or, in other words, $\mathcal{V} = \mathcal{F}^{\leftarrow} \cap \mathcal{S}_p$ in the initial relaxed problem. Due to the piece-wise nature of polynomials in $\mathcal{S}_p$ defined on the $\alpha_0 < \alpha_1 < \cdots < \alpha_K < \alpha_{K+1} = 1$, solving the problem in Eq. (5.16) reduces to solve sub-problems on each sub-interval $[\alpha_i, \alpha_{i+1}], i = 0, \ldots, K$ of $[0, 1]$. Eq. (5.16) is indeed separable into $K + 1$ independent optimization sub-problems. Each defines an optimal component G_i of the piece-wise polynomial G as defined in Eq. (5.15).

Any of these problems can be formulated generically as follows. Let $[t_0, t_1] \subset [0, 1]$, and $z_0, z_1 \in \mathbb{R}$ be interpolation values at t_0 and t_1 respectively. The goal is to find the solution to the optimization

sub-problem

$$\begin{aligned} S = \underset{L \in \mathbb{R}[x]_{\leq p}}{\operatorname{argmin}} & \left\{ \int_{t_0}^{t_1} (F_P^{\leftarrow}(x) - L(x))^2 dx \right\} \\ \text{s.t. } & L(t_0) = z_0, L(t_1) = z_1, \\ & L'(x) \geq 0, \quad \forall x \in [t_0, t_1]. \end{aligned} \tag{5.17}$$

This optimization sub-problem is nothing more than the L^2 isotonic (i.e., monotonic, in this case non-decreasing) polynomial approximation on a compact interval [161], with interpolation constraints at the boundaries. The interpolating polynomials have been extensively studied in the literature [77], as well as isotonic polynomial regression and approximation [196, 221]. However, this specific optimization problem does not seem to have been thoroughly studied.

A strategy for solving (5.17) is to use the *sum-of-squares* (SOS) [135] representation of nonnegative polynomials. These SOS representations can then be characterized using semi-definite positive (SDP) matrices [171, 172, 205]. A similar characterization of isotonic polynomials has been proposed in [205]. The following result shows that this optimization problem fits into the category of strictly convex programs: the solution of Eq. (5.20) is unique [25].

Theorem 5.1. *Let $[t_0, t_1] \subset [0, 1]$. Let M be the symmetric positive definite $((d+1) \times (d+1))$ moment matrix of the Lebesgue measure on $[t_0, t_1]$, i.e. for $i, j = 1, \ldots, d+1$,*

$$M_{ij} = \int_{t_0}^{t_1} x^{i+j-2} dx = \frac{(t_1)^{i+j-1} - (t_0)^{i+j-1}}{i+j-1}, \tag{5.18}$$

and denote $r \in \mathbb{R}^{d+1}$ the moment vector of $F_P^{\rightarrow}(x)$, i.e., for $i = 0, \ldots, d$

$$r_i = \int_{t_0}^{t_1} x^i F_P^{\rightarrow}(x) dx. \tag{5.19}$$

Then, the vector $s^* = (s_0, \ldots, s_d)^\top \in \mathbb{R}^{d+1}$ of coefficients characterizing the polynomial S in (5.17) is the solution of the following convex constrained quadratic program

$$\begin{aligned} s^* = \underset{s \in \mathbb{R}^{p+1}}{\operatorname{argmin}} & \; s^\top M s - 2 s^\top r \\ \text{s.t. } & s \in \mathcal{K}, \end{aligned} \tag{5.20}$$

where $\mathcal{K}$ is an identifiable closed convex subset of $\mathbb{R}^{p+1}$ (for the sake of conciseness, $\mathcal{K}$ is characterized within the proof).

Proof: see, Section E.5 on p.154.

Remark 5.2. Constraining the polynomials in (5.17) to be strictly increasing (i.e., $L'(x) > 0$) would ensure copula preservation. However, the set $\mathcal{K}$ in Theorem 5.1 would be open, and the existence of an optimal solution would not be guaranteed.

As solving for s^* in Eq. (5.20) is a convex-constrained quadratic program, it can be addressed efficiently using devoted solvers. The problem in Eq. (5.16) amounts to solving $K+1$ optimization problems of the form Eq. (5.20). Furthermore, computations can be done in parallel. The problem in Eq. (5.20) can be formulated and solved using `CVXR`, an `R` package for disciplined convex programming [81]. The optimization scheme is illustrated in Algorithm 1.

While computing the Lebesgue moment matrix M on each sub-interval of $[0, 1]$ is straightforward, computing strategies for r, the moment vector of $F_P^{\leftarrow}$, can vary depending on whether P is empirical or not. Additional computational details are given in Appendix E.3.

To provide a frame of reference for the practical usage of this method, the empirical computational time of solving one element of G, w.r.t. the polynomial degree is studied. Values $t_0, t_1 \in [0, 1]$, and $z_0, z_1 \in \Omega_X$ are randomly selected, and an isotonic interpolating piece-wise continuous polynomial is fitted (i.e., solving Eq. (5.20)). Polynomials of degrees ranging from 2 to 50 are fitted for each experiment,

Algorithm 1 Isotonic interpolating piece-wise continuous polynomial optimization strategy

Require: α, b, $F_P^{\rightarrow}$, p

1: **for** $i = 0, \ldots, K$ **do** (in parallel)
2: Compute M on $[\alpha_i, \alpha_{i+1}]$ (5.18).
3: Compute r on $[\alpha_i, \alpha_{i+1}]$ (5.19).
4: Setup `CVXR` constraints.
5: $s^{(i)} \leftarrow$ Solve (5.20).
6: $G_i(x) \leftarrow \sum_{j=0}^{p} s_j^{(i)} x^j$
7: **end for**
8: **return** $G(x) \leftarrow \sum_{i=0}^{K} G_i(x) \mathbb{1}_{[\alpha_i, \alpha_{i+1}]}(x)$

repeated 150 times. The execution time[1] has been recorded and is displayed in Figure 5.3. The mean computational time seems to be linear w.r.t. the polynomial degree. However, the higher the degree, the wider the 90% time coverage seems to be, which may be caused by the complexity of the underlying optimization problem. In our limited testing, further numerical experiments showed that small polynomial degrees (≤ 7) often appear sufficient to obtain good approximations. Moreover, the approximation error tends to stabilize, w.r.t. the polynomial degree, rather rapidly.

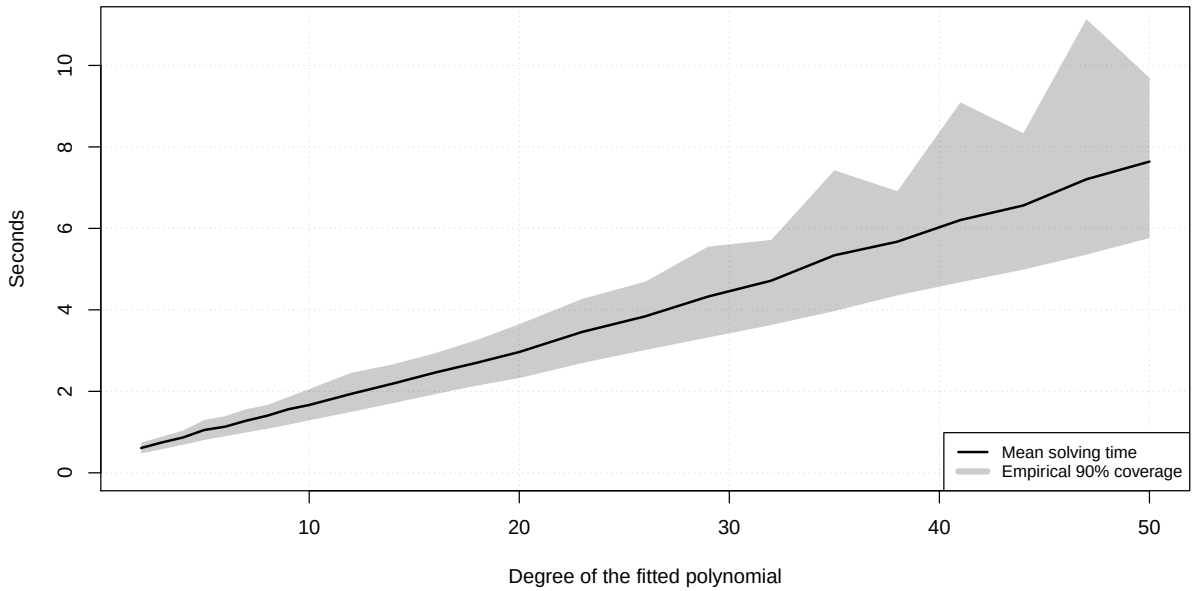

Figure 5.3: Computational solving time in seconds of the optimization problem in Eq. (5.20) using `CVXR`, w.r.t. the chosen degree of the polynomial. Computations have not been executed in parallel.

Remark 5.3. The numerical solver used is `SCS V3.2.1` [166]. The quantile functions have been mapped to take values between $[-1, 1]$ to improve numerical stability. All the figures and all obtained optimal perturbations have been computed by performing this pre-processing step first. The interested reader is referred to the accompanying GitHub repository[a].

[a] https://github.com/milidris/phdThesis

5.5 Illustration on use-cases

The perturbation method is applied to two use cases to illustrate the robustness insights one can gather regarding black-box models. First, the robustness to feature perturbations of a classification model (i.e., a one-layer neural network) trained on an acoustic fire extinguisher dataset is studied. Local and global

[1] Using an AMD Ryzen 7 4750U 8-core processor.

diagnostics are showcased, leading to tangible insights. The second use case deals with a numerical hydrological model from the UQ literature. The perturbation methodology allows going beyond classical metrics for surrogate model validation.

Remark 5.4. The following applications apply optimal perturbations using an isotonic polynomial smoothing with an arbitrarily high degree. The degree is chosen based on an empirical inspection of the solutions and ensuring that the approximation error remains relatively the same w.r.t. higher degrees.

Particular attention has been put on copula preservation. Even though the relaxed problem in Eq. (5.12) is solved in the following applications, the solutions are composed of strictly increasing marginally perturbed quantile functions.

5.5.1 Acoustic fire extinguisher: Airflow perturbation

Perturbation strategy A straightforward perturbation strategy is proposed for the Airflow feature. The perturbation is composed of the $K = 14$ constraints:

- The application domain of the feature is preserved by setting both the 0 and 1-quantiles to the dataset's minimum and maximum observed value.
- The left tail of the distribution is preserved by constraining every quantile of level 10% to 60% with a step of 5% to interpolate the empirical quantile function of the feature.
- A quantile shift perturbation is put on the 80%-quantile of the feature, with an initial value of $F_P^{\leftarrow}(0.8) = 12$, being shifted between 9.5 ($\theta = -1$) and 14.5 ($\theta = 1$).

In addition to these perturbations, piece-wise continuous isotonic polynomials smoothing is enforced. The degree of each increasing polynomial has been arbitrarily chosen to be up to 9. The constraints and the resulting quantile-constrained Wasserstein projections are illustrated in Figure 5.4 for intensity values -1, 0, and 1.

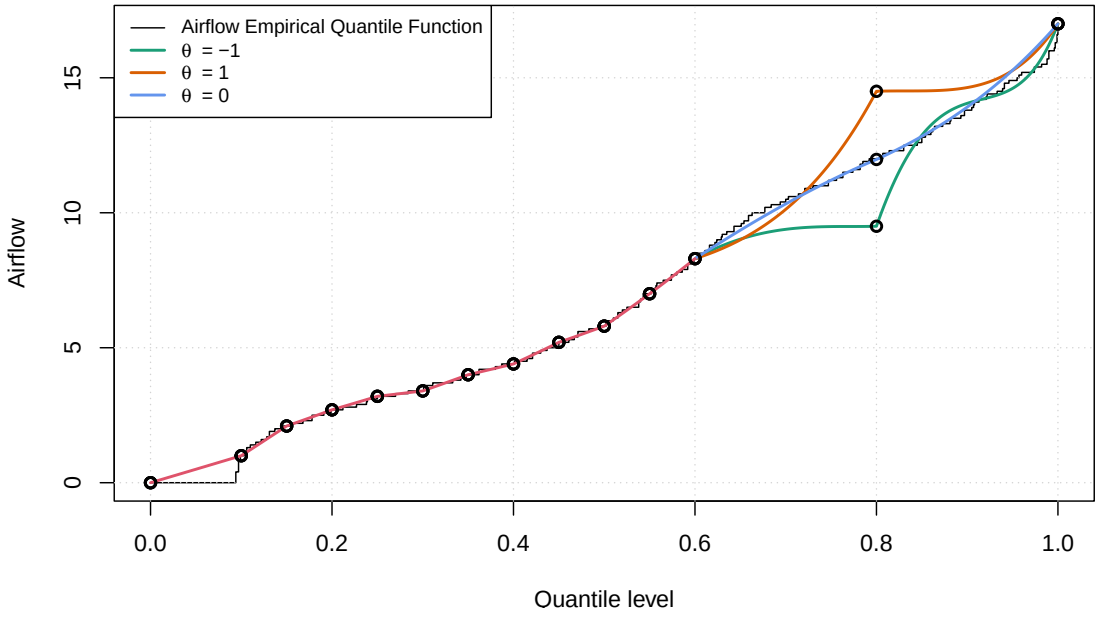

Figure 5.4: Quantile functions of the optimally perturbed Airflow feature, with a chosen polynomial degree equal to 9. The red line represents the preserved tail; meanwhile, the green, blue, and yellow lines represent various quantile shift intensity levels ($\theta = -1$, $\theta = 0$, and $\theta = 1$, respectively).

The perturbed quantile level has been chosen with the model's decision boundary in mind: no observation in the initial dataset with an Airflow value exceeding 12.3m/s is classified by the model as not extinguishing the fire, regardless of the values taken by the other features. Perturbing the 80%-quantile of the Airflow variable allows for exploring the model's behavior in regions close to this decision boundary. More importantly, it allows for assessing the predictive robustness of the neural network in this region

under perturbations of varying magnitude. Generally, this quantile shift regime can be understood as a perturbation on the right tail of the initial distribution, i.e., on values higher than the 60%-quantile.

Model robustness assessment First, global robustness insights are highlighted. The left plot of Figure 5.5 presents the proportion of perturbed observations with predictions of 1 w.r.t. to the intensity of the perturbation. Notice that the proportion is increasing, along with θ. Hence, decreasing the value of the initial 80%-quantile tends to result in a lower number of predicted put-out fires, and increasing its value results in an increasing number of predicted put-out fires. This interpretation is relatively intuitive: all other things being equal, a higher Airflow value entails a higher chance of predicting $Y = 1$. The right plot of Figure 5.5 presents the proportion of prediction shift w.r.t. θ. Notice that the higher the magnitude of the perturbation (positively or negatively), the more predictions tend to change, and the closest θ is to 0, the fewer predictions shift. This observation informs on the predictive stability in the vicinity of the decision boundary of the model: small perturbations tend to result in fewer prediction shifts than bigger perturbations.

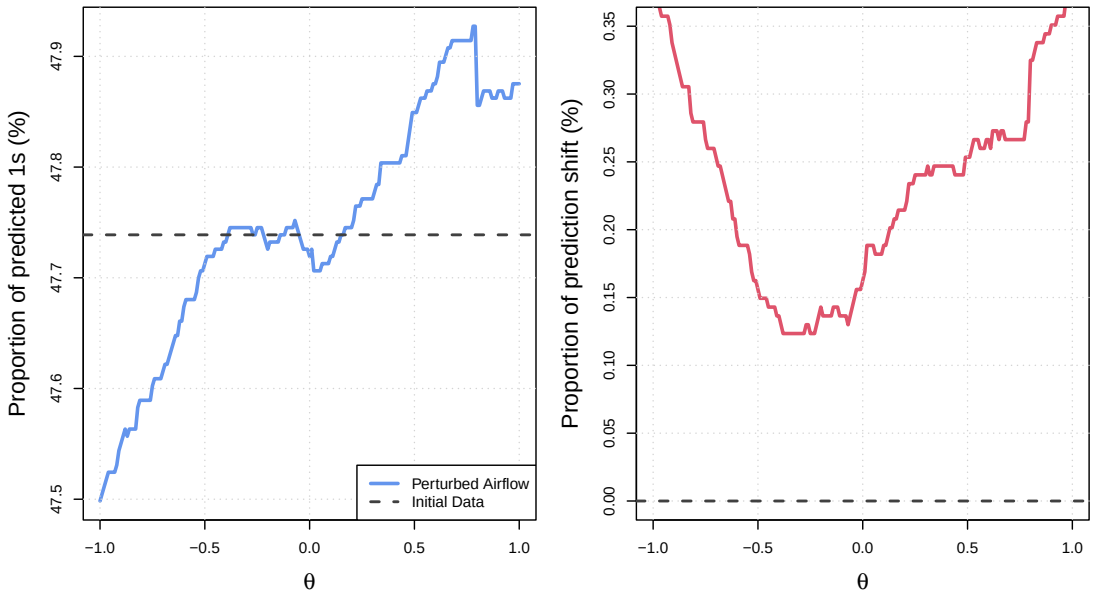

Figure 5.5: Proportion of predictions $Y = 1$ (left) and proportion of classification prediction shift (right) compared to the initial data, w.r.t. the perturbation intensity parameter θ.

Figure 5.6 presents the behavior of the Shapley effects (see, Chapter 3) w.r.t. the quantile shift intensity parameter θ. These indices have been computed using the nearest-neighbor (KNN) approach proposed in [33] (with an arbitrarily chosen number of neighbors equal to 6). Studying the behavior of importance measures informs on the stability of this diagnostic (i.e., feature importance order) w.r.t. input perturbation, i.e., if the importance hierarchy between the inputs changes due to perturbations around the model's decision boundary. The left barplot presents the initial target Shapley effects, computed on the model's prediction on the observed data, and the right plot presents their behavior under the airflow perturbation. One can notice that the importance indices remain stable w.r.t. θ. This result indicates that the global SA of the neural network is relatively robust to the distributional perturbations driven by θ. Hence, one can be confident in those diagnostics under uncertainties in the region near the model's decision boundary.

Finally, the robustness of the neural network can also be assessed locally. Figure 5.7 allows visualizing whether a prediction has shifted w.r.t. to the effective magnitude of the perturbation. The black line indicates no perturbation change: the airflow value of an observation has been mapped to itself. For a fixed initial airflow datapoint, its vertical distance to the black line indicates the (signed) magnitude of the applied perturbation. Red points indicate that the prediction has shifted w.r.t. the initial dataset, and blue points indicate no predictive change. One can note the presence of red dots close to the black line around the prediction boundary of the model. Small perturbations for observations with airflow values around 12, all other features being equal, can lead to a prediction change. Hence, the confidence in predictions on observations in this region can be questioned. However, notice the lack of red dots near the black line for airflow values on the interval $[13, 17]$ and on the interval $[7, 10]$. Hence, one can be

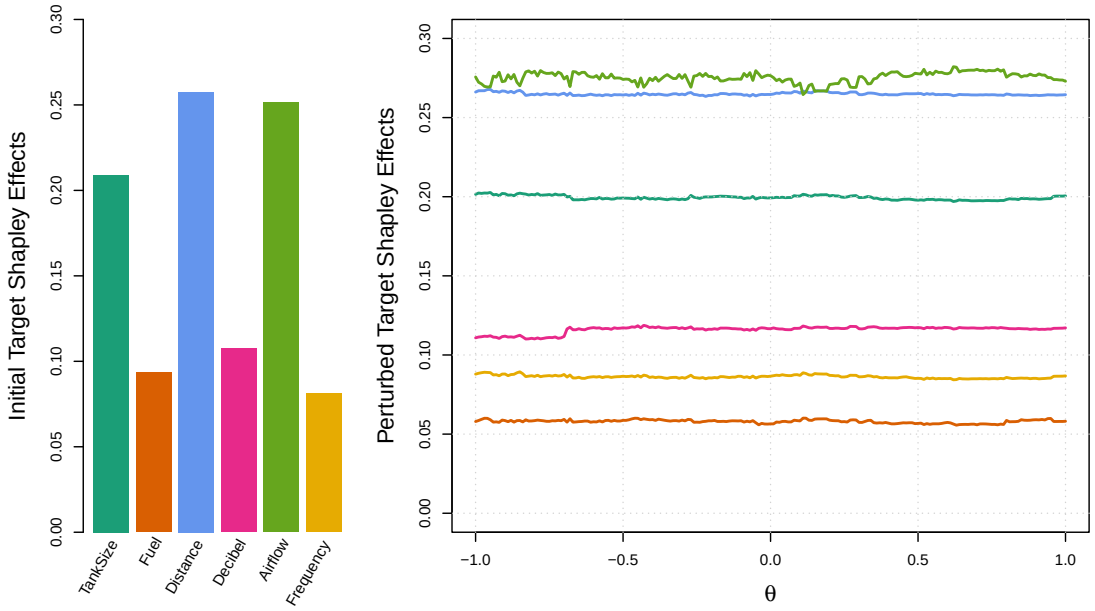

Figure 5.6: Initial (left) and perturbed (right) target Shapley effects, w.r.t. the intensity parameter θ, using the same color panel.

confident in the model's predictions for Airflow values on these intervals, which seem robust w.r.t. the quantile shift.

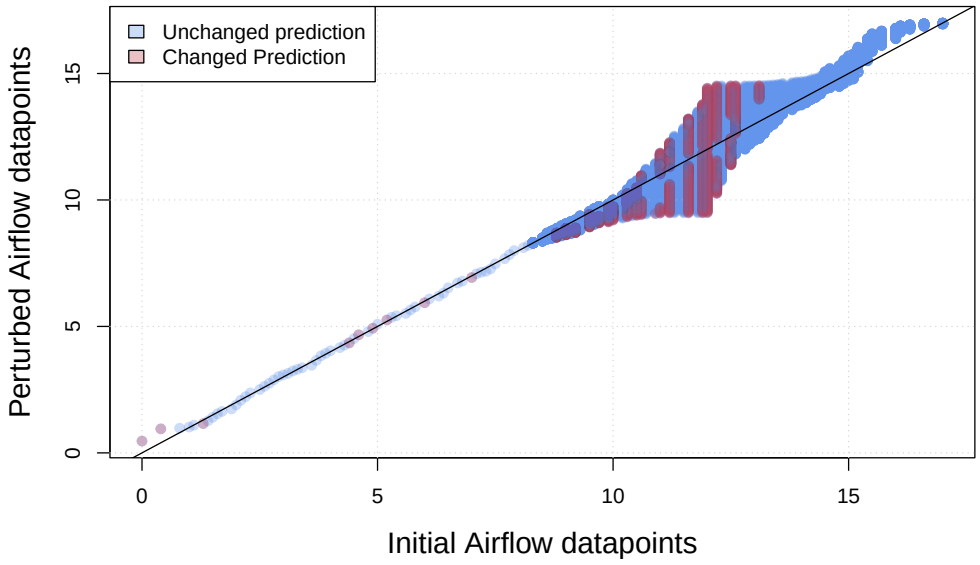

Figure 5.7: Perturbed datapoints w.r.t. their initial values. The black line represents no perturbation. The red and blue dots represent either a classification shift due to the perturbation or no classification shift.

One may notice the presence of small perturbations resulting in prediction changes for Airflow values around $[0, 5]$. However, since the perturbation scheme focuses on exploring the model's behavior around the decision boundary, their interpretation is voluntarily omitted: a different perturbation scheme involving perturbing the left tail of the airflow distribution would be advised.

In summary, besides its good prediction accuracy, the model is globally robust to distributional perturbation focused around the decision boundary of its Airflow feature. Moreover, one can be confident in the feature importance indices since they remain relatively similar under perturbation. Locally, the model prediction seems stable w.r.t. small perturbations, except on a small interval around its decision boundary (a behavior generally expected in ML applications). In conclusion, this robust interpretability analysis further assesses the model's behavior beyond classical accuracy metrics and provides additional arguments for its validation.

5.5.2 River water level: surrogate model validation

Perturbation strategy In this use case, the three following inputs are perturbed. The river's maximum annual water flow rate Q, the river length L, and the upstream river level Z_m are subject to the following punctual quantile constraints:

- Quantile perturbations on Q:
  - Shift of the application domain from $[500, 3000]$ to $[500, 3200]$;
  - Preserve the median of the distribution;
  - Increase the initial 15%-quantile by 75;
  - Decrease the initial 75%-quantile by 125;
- Quantile perturbations on L:
  - Shift the application domain from $[4990, 5010]$ to $[4988, 5012]$;
  - Preserve the median of the distribution;
- Quantile perturbations on Z_m:
  - Preserve the application domain and the median of the initial distribution;
  - Increase the 80% and 90%-quantiles by 0.1;
  - Decrease the 25%-quantile by 0.05.

The initial input distributions, their application domain, and the optimally perturbed results are illustrated in Figure 5.8. These constraints are mainly enforced to illustrate that multiple inputs can be perturbed simultaneously while preserving their dependence structure. They can be interpreted, for instance, as domain experts' knowledge injection into the initial probabilistic structure of the inputs (e.g., to study a specific river arm).

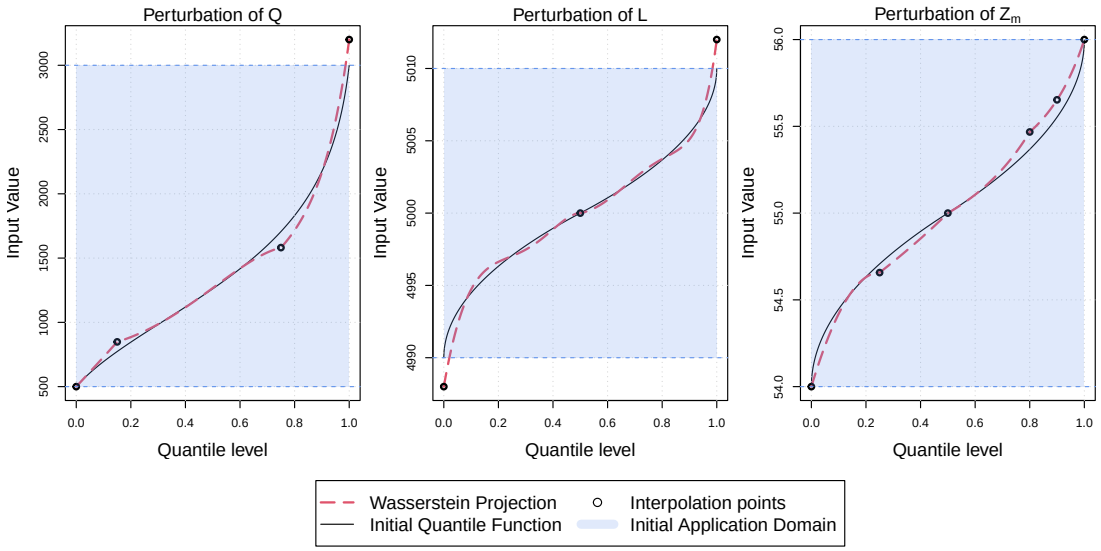

Figure 5.8: Initial quantile functions, application domains, and corresponding optimally perturbed quantile functions of the Q, L, and Z_m inputs.

In addition to these constraints, the Strickler coefficient K_s is subject to an application domain dilatation perturbation, with a scaling parameter $\eta = 1.5$. Each perturbation intensity represents a degree of uncertainty on the type of riverbed roughness. When $\theta = -1$, the width of the initial application domain is reduced, i.e., from $[15, 55]$ to $[21.66, 48.33]$, which can be interpreted in a situation where the epistemic uncertainty on the riverbed roughness is narrower, between a slow winding natural river, up to a plain river without shrub vegetation. When $\theta = 1$, the epistemic uncertainty on the riverbed is much wider. The application domain equals $[5, 65]$, depicting a wider range of possible riverbed roughness, from

proliferating algae to smooth concrete. Figure 5.9 illustrates the initial K_s distribution and the optimally perturbed quantile functions for θ equal to -1 and 1. Hence, θ can be interpreted as a proxy for the "amount" of epistemic uncertainty on the riverbed roughness.

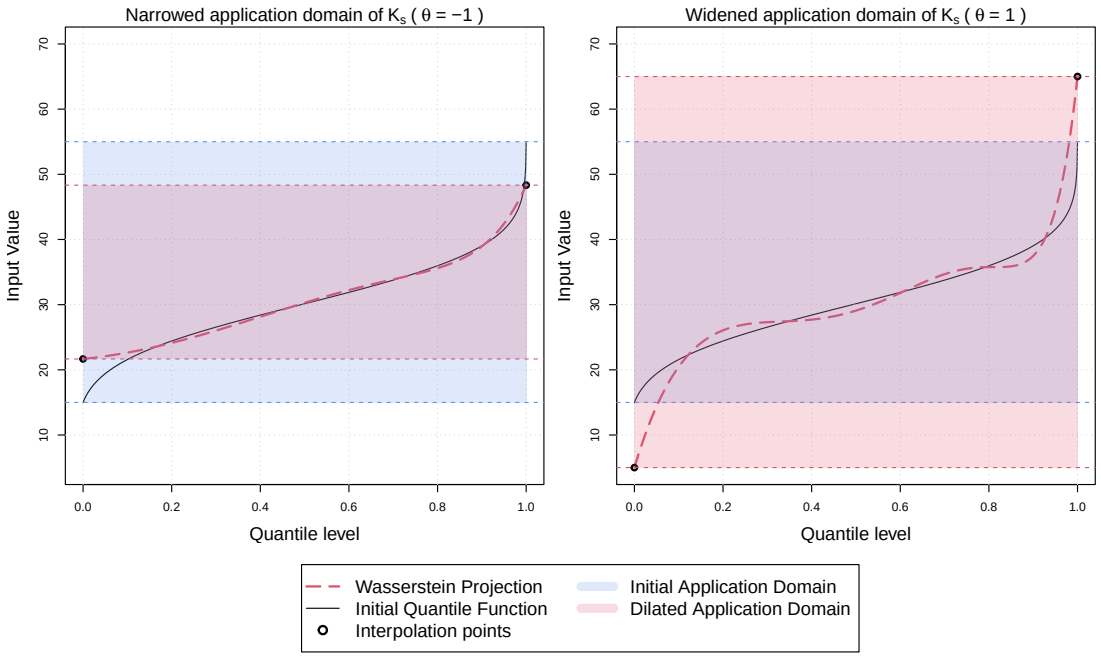

Figure 5.9: Initial quantile function, application domain and corresponding optimally perturbed quantile functions for K_s, for θ being equal to -1 (left) and 1 (right), for a scaling parameter $\eta = 2$.

Additionally, the perturbations' smoothness is enforced using piece-wise continuous isotonic polynomials of degree up to 12, chosen arbitrarily.

Robustness of the sensitivity analysis From a global standpoint, one can be interested in the impact of the distributional perturbations on key statistics of the random output of the river water level model. Figure 5.10 presents estimated values for the mean, standard deviation, 2.5% and 97.5%-quantiles (shown by the 95% coverage), and minimum and maximum values of the random output, computed on Monte Carlo samples of size 4×10^5, w.r.t. the dilatation intensity θ. These values are compared to the reference ones according to the initial distribution of the inputs, estimated on a simulated sample of size 10^6.

Notice that the expectation, standard deviation, 95% coverage quantiles, and minimum value of the model output remain stable under the distributional perturbations on the application domain of the Strickler coefficient. However, the estimated upper bound of the output support increases exponentially for positive values of θ. Widening the uncertainty on the riverbed type allows for relatively rare events of high river water levels since the 97.5%-quantile does not seem dramatically affected by the distributional perturbations.

Figure 5.11 presents the Shapley effects (see, Chapter 3) w.r.t. the perturbation intensity θ. These indices have been computed using a Monte Carlo scheme as depicted in [206, 115] and recalled in Section C.1.1, with fixed simulated sample sizes, for each perturbed distribution Q driven by a value of θ, $N_v = 10^4$ for estimating $\mathrm{Var}_Q(Y)$, as well as $N_o = 10^3$ and $N_i = 100$ to estimate the conditional elements for every subset of inputs $X_A, A \subseteq D$. Additionally, the displayed reference Shapley effects have been computed as in Section 3.4.2.

Note that the distributional perturbations have an impact on the importance measures. More precisely, increasing the range of the uncertainty of the riverbed roughness increases its importance for positive values of θ. Conversely, the importance of both Q and Z_v decreases accordingly. However, the variable importance hierarchy induced by the Shapley effects seems to be generally preserved. It is also essential to notice that the gap in importance between Q and Z_v and K_s decreases as θ gets large. Hence, this SA does not seem robust to distributional perturbations and, more precisely, to a widening of the support of the Strickler coefficient in combination with the quantile perturbations put on Q, L, and Z_m. In other words, the distribution of K_s impacts the subsequent importance quantification, and its definition

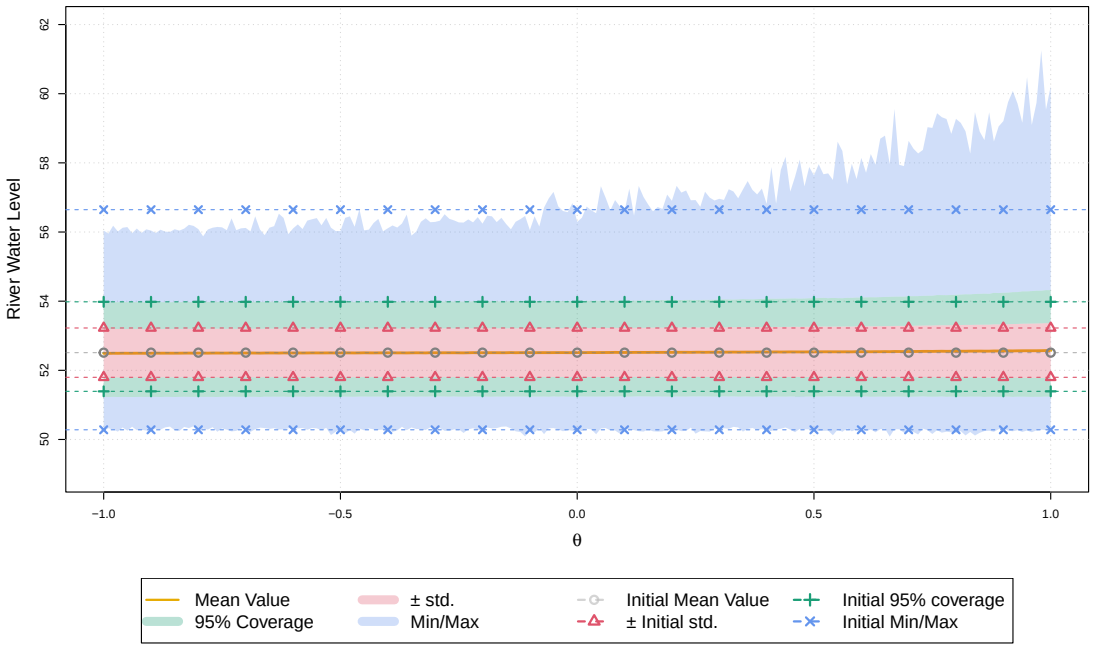

Figure 5.10: Expectation, standard deviation, 95% coverage, minimum and maximum estimators of the river water level, w.r.t. the application domain dilatation intensity θ.

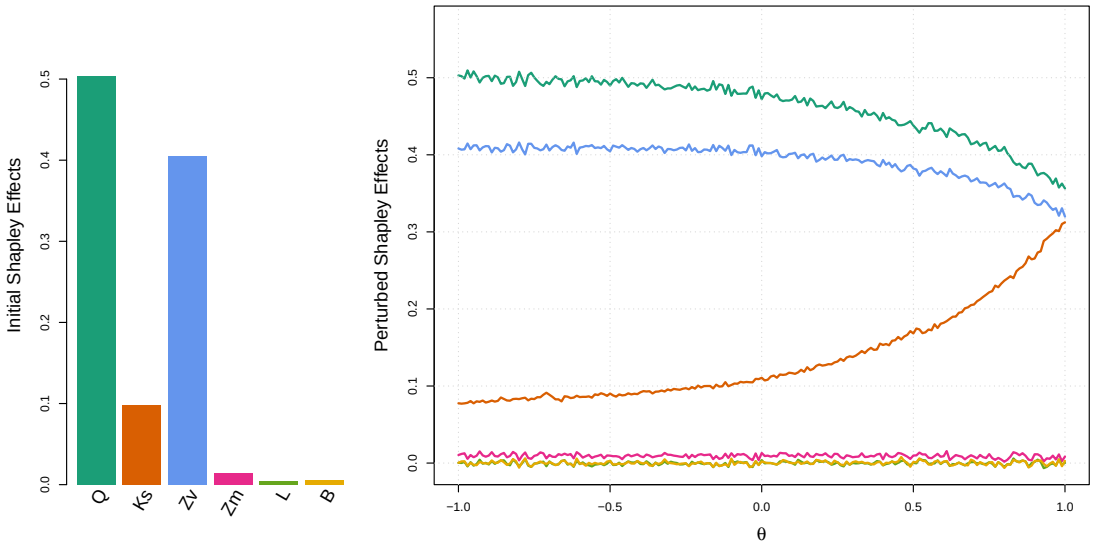

Figure 5.11: Reference Shapley effects (left) and Shapley effects of the river water level model under optimally dilated application domain w.r.t. θ (right), using the same color panel.

requires particular care.

Surrogate model validation A surrogate model is trained on a simulated input-output sample of size 10^6 of the initial probabilistic structure and validated on a validation dataset of size 10^5. The surrogate model is a neural network comprised of 3 hidden layers, 64 neurons each, and ReLu as an activation function. The model's R^2 is 98.18% on both the training and validation data. Despite the model's good results on the validation data, it does not behave the same way as the initial model when perturbed similarly. Echoing Figure 5.10, Figure 5.12 illustrates the model's behavior when subject to the previously introduced perturbations.

One can notice that the surrogate model does not display the same behavior as the numerical model w.r.t. the epistemic uncertainty of the riverbed roughness. Even though the surrogate model generalizes well on validation data, its behavior on the perturbed data differs from the initial numerical model.

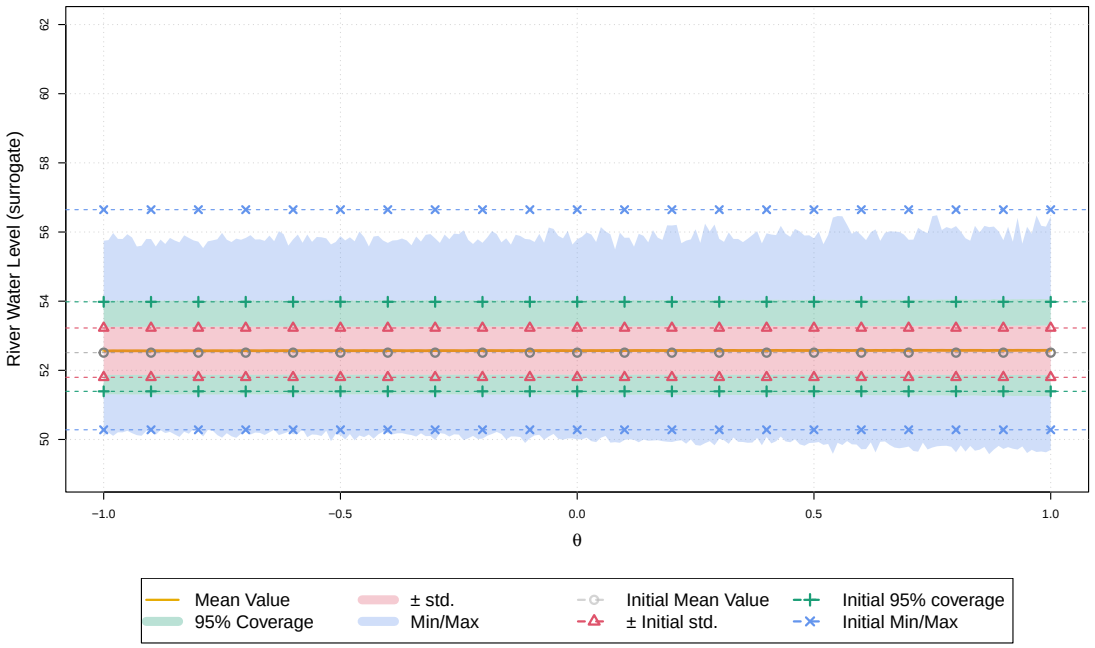

Figure 5.12: Expectation, standard deviation, 95% coverage, minimum and maximum estimators of the surrogate model, w.r.t. the application domain dilatation intensity θ.

More precisely, the maximal value of the river water level does not seem to be impacted by the epistemic uncertainty of the riverbed roughness. However, the other statistics (mean, variance, and 95% coverage) align with the numerical model. Hence, despite its good fit, using this surrogate model would not be advised if the goal of the sensitivity analysis is to study rare events.

5.5.3 Conclusions

These two use cases illustrate the different insights the perturbation methodology can offer in UQ and ML studies. On the ML side, for classification tasks, it allows for assessing the global behavior of black-box models under input perturbations. This assessment is quantified either through studying the prediction shifts due to the perturbation or through the behavior of feature importance metrics. Locally, it allows the detection of low-stability regions of interest (regions where small perturbations induce a classification change). In addition to classical accuracy metrics, our method can be used to assess confidence in a predictive model. On the UQ side, it allows for studying the impact of distributional perturbations (whose intensity can be tuned to represent epistemic uncertainties) on the model output, even in situations where inputs are correlated. Furthermore, in an SA context, the behavior of classical sensitivity indices under those perturbations can also be studied, and their robustness (for instance, the preservation of the input importance hierarchy) w.r.t. the probabilistic modeling of the inputs can be assessed. In both cases, meaningful perturbations allow for a more complete picture, beyond classical validation metrics, of a black-box model's behavior outside of the initial distribution.

5.6 Discussion

Obtaining robustness diagnoses on the influence of input variables and the behavior of a model considered a black box is essential for its acceptance and use. This chapter provides a tool to answer this problem by modifying the distributions of the inputs in a controlled manner. Four desirability criteria are introduced and discussed to ensure the interpretability method's meaningfulness. The developed method relies on the choice of the 2-Wasserstein distance with perturbations on univariate quantiles, which allows for preserving the input's dependence structure (i.e., copula). Regularity conditions can be enforced, and the case of piece-wise interpolating isotonic polynomials is studied. The robustness analyses conducted on real case studies illustrate its potential flexibility and adequate computational cost, which are essential for high-dimensional cases. These studies highlighted validation insights be-

yond classical tools, allowing for a more complete understanding of the black-box model's behavior. However, it is essential to note that while the tool presented in this chapter allows exploring *some aspects* of the robustness of black-box models, many other aspects are also worth exploring. Formally defining the notion of robustness is a complicated task, and the presented work only tackles input perturbations. The interested reader is referred to [78] for a more complete picture of robustness in the ML field.

Several technical avenues of improvement can be considered. First, concerning the piece-wise interpolating isotonic polynomials as a smoothing vehicle. Throughout the chapter, the degrees of the polynomials have been chosen arbitrarily or through obscure heuristics. To guide the choice of this degree, one could use prior information on the order of differentiability of the sought-after perturbed gqf. In an ML framework, nonparametric approaches to isotonic regression of the marginal gqfs can provide answers through statistical testing [64, 49] or criteria enforcing a trade-off between approximation error and sparsity (e.g., inspired from AIC or BIC). Moreover, while the proposed methodology allows for continuous results, differentiability is not guaranteed. However, the literature on isotonic splines [97, 196, 77, 221] can be leveraged to offer a better range of smoothness constraints. Finally, other spaces of functions can also be used for smoothing purposes. Following the work of [14], abstract reproducing kernel Hilbert space of nonnegative functions can be reached through particular kernels. Hence, it would allow access to different sets of nonnegative functions whose regularities can be assessed through a thorough study of these kernels.

The proposed methodology solely focuses on marginal perturbation, preserving the dependence structure of the inputs. However, one may wish to perturb the dependence structure as well. However, it is argued that copula perturbation should be done independently of marginal perturbations for the sake of the final interpretation of the robustness analyses. It allows separating the effects in the marginal perturbation of the effects of the stochastic dependence perturbation. Association and concordance measures appear as the most interpretable tools for copula manipulation (and are frequently used to incorporate expert opinion) [42, 225, 22]. An alternative approach to perturb the stochastic dependence structure and the marginal would be to consider multivariate quantile functions. However, defining multivariate quantile functions is not trivial and not as natural as in the univariate case. Among the many approaches to defining such a notion, the most theoretically accomplished today is the one resulting from the concept of *center-outward distribution function* [40, 94, 24]. Perturbing these quantile contours can be leveraged to go beyond marginal consideration.

One of the primary motivations for using the 2-Wasserstein distance as a projection metric is that it metricizes weak convergence on a broad set of probability measures. This property allows being generic on the initial probability measures P and does not restrict the perturbed probability measure to be in a particular class (e.g., with a density). Other distances between probability measures are endowed with similar properties, such as the Prokhorov-Levy distance. Leveraging the different relationships between such distances (see [88]) could be beneficial for generalizing the proposed approach.

Chapter 6

Conclusion and perspectives

Contents

6.1 Conclusion **100**
6.2 Perspectives **101**
6.2.1 Allocations as attribution methods 101
6.2.2 Orthocanonical decompositions 101
6.2.3 Robustness assessment 102

6.1 Conclusion

The general purpose of this manuscript is to explore the interpretation of black-box models and propose a first take on a suitable mathematical framework. This framework allows justifying the use and drives the development of interpretability methods to enhance the trustworthiness of black-box models of critical systems and strive towards their acceptance by regulatory instances.

Chapter 1 contextualizes this problem and lays down the overall direction taken in the manuscript. It aims to propose a unified view of post-hoc interpretability and sensitivity analysis. To that extent, the proposed framework relies on a measure-theoretic definition of the inputs and the random output, seen as random elements. The notion of quantity of interest is introduced and illustrated. In order to answer specific conundrums (i.e., interpretation questions), two scientific questions are explored:

i) **Influence quantification:** being able to quantify and rank the subsets of inputs by their influence on a model or some QoIs ;

ii) **Robustness assessment of black-box models:** being able to study the behavior of a model or some QoIs under the perturbation of its inputs.

Three use-cases are also presented, which are studied using the proposed methods throughout the manuscript.

In Chapter 2, the algebraic roots of the fundamental question of influence measuring are highlighted, driven by the mere assumption that subsets of inputs "can be ranked by influence". It naturally opens the way to coalitional decompositions of QoIs in order to produce influence measures. These measures must express the influence order between subsets of inputs. These measures can be linked to the field of combinatorics and, in particular, to Rota's generalization of the Möbius inversion formula. This connexion highlights two approaches to define influence measures: the input-centric approach, which requires a value measure allowing the total influence of subsets of inputs to be quantified, and the model-centric approach, which requires an intrinsic decomposition of the random output. These two approaches are illustrated for the problem of importance quantification, i.e., the decomposition of the random output's variance.

Chapter 3 is a deep-dive into the input-centric approach. This question has been extensively studied under the paradigm of cooperative game theory, by analogy between players and inputs of a black-box model. It allows defining allocations built upon an input-centric influence measure known as the Harsanyi dividends. Using this framework, importance attributions can be defined (i.e., decomposing the variance between the inputs themselves instead of between every subset of inputs), such as the egalitarian redistribution of dividends proposed by the Shapley effects. The latter presents a drawback: inputs that are not in the model but are correlated with the inputs in the model can be granted some importance. This problem is solved with the PMEs, which rely on a proportional redistribution of dividends. These two methods offer different ways to quantify importance, which are compared and illustrated in use-cases. Finally, a fundamental issue with the input-centric approach is highlighted: the choice of the value function. This problem motivates exploring the model-centric approach.

The model-centric approach is tackled in Chapter 4, starting with the random output decomposition. This problem does have a solution for mutually independent inputs, known as Hoeffding's decomposition. However, while many developments have been proposed in the literature, a definite answer has not yet been found concerning its potential generalization for dependent inputs under reasonable assumptions. By approaching this generalization as a direct-sum decomposition of Lebesgue spaces (i.e., Hilbert spaces), it has been shown that a generalization exists under two fairly reasonable assumptions: non-perfect functional dependence, and non-degenerate stochastic dependence. These developments stem from studying the intrinsic subspaces of Lebesgue spaces generated by multivariate random elements and, in particular, their relationships using Dixmier's and Friedrichs' angles. It leads to a rather intuitive and geometric result, relying on oblique projections. Finally, this approach enabled the definition of various influence measures, which can be easily theoretically interpreted and whose theoretical properties are discussed. This decomposition has then been illustrated using a simple toy-case.

Finally, Chapter 5, has been dedicated to studying the problem of the assessment of the robustness of black-box models. In particular, the study of the behavior of a model whenever the distribution of its inputs is perturbed. A general formalized view of this problem is proposed. It relies on an optimization problem in the space of probability measures. The definition of suitable perturbations is discussed, and

four desirability criteria are proposed. Based on these criteria, the choice of the 2-Wasserstein distance as a mean for comparing probability measures is motivated. Moreover, under this choice of discrepancy, perturbations based on quantiles coupled with the preservation of the initial dependence structure are explored. The problem can be analytically solved, but this solution is not suitable for practical studies. Smoothness constraints are introduced to solve this issue, in addition to quantile perturbations and dependence preservation. The use of isotonic interpolating polynomials is studied, leading to a well-posed optimization problem with a unique solution. The presented method is then illustrated and discussed on use-cases, opening the way to validation methods of ML models going beyond the classical metrics.

6.2 Perspectives

Some perspectives and areas of improvement related to the developments of this manuscript are presented in this section.

6.2.1 Allocations as attribution methods

The following perspectives are related to the developments presented in Chapter 3.

Estimation The computational burden associated with estimating any allocation of the Sobol' cooperative games remains a drawback. They require calculating an exponential number $(2^d - 1)$ of evaluation of the value function. An avenue to alleviate some of the computations would be to use surrogate models to estimate the conditional elements. For instance, random forests [20] or Gaussian process-based meta-models [121, 23] can be leveraged for that task, potentially reducing the need for costly numerical model evaluations. Additionally, the bias induced by using the nearest neighbor estimation method (which is the only one usable in costly application cases) does not guarantee the detection of exogenous inputs by PMEs. New given-data algorithms are required.

Other allocations As seen in Chapter 3, the Shapley effects and the PMEs are designed to extract different insights, the interest of which depends on the task at hand. While the PMEs are a reasonable option for factor fixing and factor prioritization, the Shapley effects provide a tool for model exploration that allows for a good overview of all the inputs that might impact the output, even though it is only due to correlation with other inputs. Other allocations, such as weighted Shapley values [125] or proportional Shapley values [17], may be defined with different specific UQ tasks in mind, allowing for domain-specific tools for more accurate and relevant indices. In the XAI literature, recent developments, such as [218], introduce the notion of correlation distortion due to using the Shapley values. It expresses their inability to detect exogenous inputs in a regression modeling context and to solve this issue. The authors proposed to focus on a different set of axioms to define better-suited allocations, with exogeneity detection in mind.

6.2.2 Orthocanonical decompositions

The following perspectives are related to the developments presented in Chapter 4, which are highlighted in Section 4.6.

Oblique projection estimation The first main challenge towards adopting the measures introduced in Chapter 4 is statistical estimation, the main bottleneck being to produce estimators of oblique projections. The literature seems relatively scarce when it comes to the estimation of these particular operators. A first approach would be to find estimators based on a variational problem, around the same idea that the problem stipulated the Hilbert projection theorem allows seeing orthogonal projections (and hence conditional expectations) as a distance-minimizing problem. A second idea would be to take advantage of the particular expression of oblique projections (see, e.g., [2, 46]). However, this approach involves estimating inverses of operators, which is a challenging feat. A final idea would be to find suitable bases for each $(V_A)_{A\in\mathcal{P}_D}$ to project $G(X)$ onto. However, it remains relatively complicated since these

subspaces can be infinite-dimensional (i.e., these bases are most likely Schauder). Non-orthogonal polynomial bases would be a great start to study this problem whenever X is endowed with a multivariate Gaussian probabilistic structure.

Dependence seen as angles The second main challenge is understanding the extent of such an approach. As one can notice, the point of view taken in order to prove Theorem 4.6 leverages the (somewhat surprising) linear nature of possibly highly non-linear problems (due to the function G and/or to the stochastic dependence on X) by decomposing the model into 2^d summands. There remains some work to do to explore the extent to which this approach can be linked to the more "traditional" probabilistic treatment of uncertainties. The development in this chapter uncovers an exciting path towards a more complete overview of non-linear multivariate statistics. However, many aspects remain to be mastered, implications to be discovered, and links with existing literature to unveil.

Different algebraic structures Another surprising connection leveraged in this chapter is with the field of abstract algebra, particularly with algebraic structures. The Boolean lattice naturally comes up because the sought-after decompositions rely on a power-set. However, since Rota's result is very general (see, Theorem B.1) and does not only apply to Boolean lattices, studying other algebraic structures can pave the way for more complex analysis. In particular, the study of *graphical models* [136], which relies on graph structures to define dependence between random elements, seems an excellent place to start.

6.2.3 Robustness assessment

The following perspectives are related to the developments presented in Chapter 5, which are highlighted in Section 5.6.

Beyond interpolating isotonic polynomials The developments presented in this chapter relied on interpolating isotonic polynomials in order to smooth the resulting gqfs. However, the degree choice for the fitted polynomials has yet to be studied and remains an area of improvement. Moreover, while continuity is a first step towards smoothness, one may be inclined to explore smoother results. For instance, guaranteeing differentiability (up to a particular order) could lead to more control over the perturbation scheme. The literature on isotonic splines [97, 196, 77, 221] can be leveraged to improve the proposed approach. Other spaces of functions can be worth exploring, too. For instance, the work of [14] on abstract reproducing kernel Hilbert space of nonnegative functions can be an exciting path towards different smooth solutions.

Dependence perturbation Preserving the dependence structure between the inputs is central to the proposed developments. However, one may wish to perturb the dependence structure. Following the desirability criteria, it can be argued that the dependence structure should be perturbed *independently* of the marginal perturbation for the sake of the final "causal" interpretation. Association and concordance measures are prime candidates for defining suitable copula manipulations [42, 225, 22]. The notion of multivariate gqfs can also be worth exploring for perturbation purposes, especially with the choice of the 2-Wasserstein distance. The concept of *center-outward distribution function* [40, 94, 24] seems one of the most promising to go beyond marginal perturbations.

Other discrepancies Other discrepancies can also be considered. Many distances, dissimilarities, and discrepancies between probability measures have been studied in the literature (see [88]). The choice of the 2-Wasserstein distance in the presented developments is mainly motivated by the fact that it does not restrict the initial or perturbed probability measures. The Prokhorov-Levy distance is another example of non-restrictive distance, which can be leveraged for input perturbation purposes.

References

[1] Y. A. Abramovich and C. D. Aliprantis. *An invitation to operator theory*. Graduate studies in mathematics. American Mathematical Society, 2002. ISBN: 978-0-8218-2146-6.

[2] S. N. Afriat. Orthogonal and oblique projectors and the characteristics of pairs of vector spaces. *Mathematical Proceedings of the Cambridge Philosophical Society*, 53(4):800–816, 1957. ISSN: 0305-0041, 1469-8064. DOI: 10.1017/S0305004100032916.

[3] A. Ajenjo. *Info-gap robustness assessment of reliability evaluations for the safety of critical industrial systems*. Ph.D Thesis, Université Bourgogne Franche-Comté, 2022.

[4] A. Alfonsi and B. Jourdain. A remark on the optimal transport between two probability measures sharing the same copula. *Statistics & Probability Letters*, 84:131–134, 2014. ISSN: 0167-7152. DOI: 10.1016/j.spl.2013.09.035.

[5] D.L. Allaire and K. E. Willcox. Distributional sensitivity analysis. *Procedia - Social and Behavioral Sciences*, 2:7595–7596, 2010.

[6] J. An and A.B. Owen. Quasi-regression. *Journal of Complexity*, 17(588–607), 2001.

[7] J. D. Angrist and J. S. Pischke. *Mostly harmless econometrics: an empiricist's companion*. Princeton University Press, 2009. ISBN: 978-0-691-12034-8. OCLC: ocn231586808.

[8] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 214–223, 2017.

[9] N. Asher, L. De Lara, S. Paul, and C. Russell. Counterfactual Models for Fair and Adequate Explanations. *Machine Learning and Knowledge Extraction*, 4(2):316–349, 2022. ISSN: 2504-4990. DOI: 10.3390/make4020014. Number: 2 Publisher: Multidisciplinary Digital Publishing Institute.

[10] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok. Synthesizing robust adversarial examples. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning (ICML), 10-15, 2018*, volume 80, pages 284–293, 2018.

[11] European Banking Authority. *2021 EU-Wide Stress Test*. European Banking Authority, 2020.

[12] S. Axler. *Linear Algebra Done Right*. Undergraduate Texts in Mathematics. Springer International Publishing, 2015. ISBN: 978-3-319-11079-0. DOI: 10.1007/978-3-319-11080-6.

[13] F. Bachoc, F. Gamboa, M. Halford, J. M. Loubes, and L. Risser. Explaining machine learning models using entropic variable projection. *Information and Inference: A Journal of the IMA*, 12(3), 2023. ISSN: 2049-8772. DOI: 10.1093/imaiai/iaad010.

[14] J. A. Bagnell and A-M Farahmand. Learning positive functions in a Hilbert space. *Preprint*, 2015.

[15] A. Balbás, J. Garrido, and S. Mayoral. Properties of Distortion Risk Measures. *Methodology and Computing in Applied Probability*, 11(3):385–399, 2009. ISSN: 1573-7713. DOI: 10.1007/s11009-008-9089-z.

[16] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115, 2020. ISSN: 1566-2535. DOI: 10.1016/j.inffus.2019.12.012.

[17] S. Béal, S. Ferrières, E. Rémila, and P. Solal. The proportional Shapley value and applications. *Games and Economic Behavior*. Special Issue in Honor of Lloyd Shapley: Seven Topics in Game Theory, 108:93–112, 2018. ISSN: 0899-8256. DOI: 10.1016/j.geb.2017.08.010.

[18] C. Bénard, G. Biau, S. Da Veiga, and E. Scornet. Interpretable Random Forests via Rule Extraction. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, pages 937–945. PMLR, 2021. ISSN: 2640-3498.

[19] C. Bénard, G. Biau, S. Da Veiga, and E. Scornet. SHAFF: Fast and consistent SHApley eFfect estimates via random Forests. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151, pages 5563–5582, 2022.

[20] C. Bénard, S. Da Veiga, and E. Scornet. Mean decrease accuracy for random forests: inconsistency, and a practical solution via the Sobol-MDA. *Biometrika*, 109(4):881–900, 2022. ISSN: 1464-3510. DOI: 10.1093/biomet/asac017.

[21] C. Bénesse, F. Gamboa, J. M. Loubes, and T. Boissin. Fairness seen as global sensitivity analysis. *Machine Learning*, 2022. ISSN: 0885-6125, 1573-0565. DOI: 10.1007/s10994-022-06202-y.

[22] N. Benoumechiara, N. Bousquet, B. Michel, and P. Saint-Pierre. Detecting and modeling critical dependence structures between random inputs of computer models. *Dependence Modeling*, 8(1):263–297, 2020. DOI: doi:10.1515/demo-2020-0016.

[23] N. Benoumechiara and K. Elie-Dit-Cosaque. Shapley effects for sensitivity analysis with dependent inputs: bootstrap and kriging-based algorithms. *ESAIM: Proceedings and Surveys*, 65:266–293, 2019. ISSN: 2267-3059. DOI: 10.1051/proc/201965266. (Visited on 12/13/2023). Publisher: EDP Sciences.

[24] B. Bercu, J. Bigot, and G. Thurin. Monge-Kantorovich superquantiles and expected shortfalls with applications to multivariate risk measurements, 2023.

[25] D. P. Bertsekas. *Nonlinear programming*. Athena scientific, 3rd ed edition, 2016. ISBN: 978-1-886529-05-2.

[26] M. Besner. Value dividends, the Harsanyi set and extensions, and the proportional Harsanyi solution. *International Journal of Game Theory*, 49(3):851–873, 2020. ISSN: 0020-7276, 1432-1270. DOI: 10.1007/s00182-019-00701-4.

[27] N. Bloom. The impact of uncertainty shocks. *Econometrica*, 77(3):623–685, 2009. DOI: https://doi.org/10.3982/ECTA6248.

[28] E. Borgonovo, A. Figalli, E. Plischke, and G. Savare. Probabilistic Sensitivity with Optimal Transport. *Preprint*, 2022.

[29] E. Borgonovo, V. Ghidini, R. Hahn, and E. Plischke. Explaining classifiers with measures of statistical association. *Computational Statistics & Data Analysis*, 182:107701, 2023. ISSN: 0167-9473. DOI: 10.1016/j.csda.2023.107701.

[30] E. Borgonovo and E. Plischke. Sensitivity analysis: A review of recent advances. *European Journal of Operational Research*, 248(3):869–887, 2016. ISSN: 0377-2217. DOI: 10.1016/j.ejor.2015.06.032.

[31] A. Brandenburger. Cooperative Game Theory: Characteristic Functions, ALlocations, Marginal Contribution, 2007.

[32] B. Broto. *Sensitivity analysis with dependent random variables: Estimation of the Shapley effects for unknown input distribution and linear Gaussian models*. Ph.D thesis, Université Paris-Saclay, 2020.

[33] B. Broto, F. Bachoc, and M. Depecker. Variance Reduction for Estimation of Shapley Effects and Adaptation to Unknown Input Distribution. *SIAM/ASA Journal on Uncertainty Quantification*, 8(2):693–716, 2020. ISSN: 2166-2525. DOI: 10.1137/18M1234631.

[34] L. Bruzzone and M. Marconcini. Domain Adaptation Problems: A DASVM Classification Technique and a Circular Validation Strategy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(5):770–787, 2010.

[35] W. Bryc. Conditional expectation with respect to dependent sigma-fields. In *Proceedings of VII conference on Probability Theory*, pages 409–411, 1984.

[36] W. Bryc. Conditional Moment Representations for Dependent Random Variables. *Electronic Journal of Probability*, 1:1–14, 1996. ISSN: 1083-6489, 1083-6489. DOI: 10.1214/EJP.v1-7. Publisher: Institute of Mathematical Statistics and Bernoulli Society.

[37] V. Chabridon. *Reliability-oriented sensitivity analysis under probabilistic model uncertainty – Application to aerospace systems*. PhD thesis, Université Clermont Auvergne, 2018.

[38] G. Chastaing, F. Gamboa, and C. Prieur. Generalized Hoeffding-Sobol decomposition for dependent variables - Application to sensitivity analysis. *Electronic Journal of Statistics*, 6:2420–2448, 2012.

[39] A Chateauneuf and J-Y. Jaffray. Some characterizations of lower probabilities and other monotone capacities through the use of Möbius inversion. *Mathematical Social Sciences*, 17(3):263–283, 1989. ISSN: 0165-4896. DOI: 10.1016/0165-4896(89)90056-5.

[40] V. Chernozhukov, A. Galichon, M. Hallin, and M. Henry. Monge-Kantorovich depth, quantiles, ranks and signs. *The Annals of Statistics*, 45(1):223–256, 2017. DOI: 10.1214/16-AOS1450.

[41] Y. Chung, W. Neiswanger, I. Char, and J. Schneider. Beyond pinball loss: quantile methods for calibrated uncertainty quantification. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 10971–10984, 2021.

[42] R. T. Clemen and T. Reilly. Correlations and copulas for decision and risk analysis. *Management Science*, 45(2):208–224, 1999. DOI: 10.1287/mnsc.45.2.208.

[43] T. J. Cleophas and A. H. Zwinderman. *Regression Analysis in Medical Research: for Starters and 2nd Levelers*. Springer International Publishing, 2021. ISBN: 978-3-030-61393-8. DOI: 10.1007/978-3-030-61394-5.

[44] L. Clouvel, B. Iooss, V. Chabridon, M. Il Idrissi, and F. Robin. A review on variance-based importance measures in the linear regression context. Preprint, 2023.

[45] J.B. Conway. *A Course in Functional Analysis*, volume 96 of *Graduate Texts in Mathematics*. Springer, 2007. ISBN: 978-1-4419-3092-7. DOI: 10.1007/978-1-4757-4383-8.

[46] G. Corach, A. Maestripieri, and D. Stojanoff. A classification of projectors. In *Topological Algebras, their Applications, and Related Topics*, pages 145–160. Institute of Mathematics Polish Academy of Sciences, 2005. DOI: 10.4064/bc67-0-12.

[47] I. Csiszár. I-Divergence Geometry of Probability Distributions and Minization problems. *The Annals of Probability*, 3(1):146–158, 1975. DOI: 10.1214/aop/1176996454.

[48] S. Da Veiga, F. Gamboa, B. Iooss, and C. Prieur. *Basics and Trends in Sensitivity Analysis. Theory and Practice in R*. SIAM. Computational Science and Engineering, 2021.

[49] S. Da Veiga and A. Marrel. Gaussian process modeling with inequality constraints. *Annales de la Faculté des Sciences de Toulouse*, 3:529–555, 2012.

[50] J. Dauxois and G. M. Nkiet. Canonical analysis of two Euclidean subspaces and its applications. *Linear Algebra and its Applications*. Sixth Special Issue on Linear Algebra and Statistics, 264:355–388, 1997. ISSN: 0024-3795. DOI: 10.1016/S0024-3795(96)00244-3.

[51] J. Dauxois, G. M. Nkiet, and Y Romain. Canonical analysis relative to a closed subspace. *Linear Algebra and its Applications*. Tenth Special Issue (Part 1) on Linear Algebra and Statistics, 388:119–145, 2004. ISSN: 0024-3795. DOI: 10.1016/j.laa.2004.02.036.

[52] B. A. Davey and H. A. Priestley. *Introduction to Lattices and Order*. Cambridge University Press, 2nd edition, 2002. ISBN: 978-0-521-78451-1. DOI: 10.1017/CBO9780511809088.

[53] L. De Lara, A. González-Sanz, N. Asher, and J. M. Loubes. Transport-based counterfactual models. *arXiv preprint arXiv:2108.13025*, 2021.

[54] L. De Lara, A. González-Sanz, and J. M. Loubes. Diffeomorphic registration using Sinkhorn divergences. *SIAM Journal on Imaging Sciences*, 16(1):250–279, 2023.

[55] A. Dembo, A. Kagan, and L. A. Shepp. Remarks on the Maximum Correlation Coefficient. *Bernoulli*, 7(2):343–350, 2001. ISSN: 1350-7265. DOI: 10.2307/3318742.

[56] J. Derks, H. Haller, and H. Peters. The selectope for cooperative games. *International Journal of Game Theory*, 29(1):23–38, 2000. ISSN: 1432-1270. DOI: 10.1007/s001820050003.

[57] E. de Rocquigny, N. Devictor, and S. Tarantola, editors. *Uncertainty in Industrial Practice*. John Wiley and Sons, Ltd, 2008. ISBN: 978-0-470-99447-4. DOI: 10.1002/9780470770733.

[58] H. Dette and W. J. Studden. *The theory of canonical moments with applications in statistics, probability, and analysis*. Wiley series in probability and statistics. Wiley, 1997. ISBN: 978-0-471-10991-4.

[59] F. Deutsch. The Angle Between Subspaces of a Hilbert Space. In S. P. Singh, editor, *Approximation Theory, Wavelets and Applications*, NATO Science Series, pages 107–130. Springer Netherlands, 1995. ISBN: 978-94-015-8577-4. DOI: 10.1007/978-94-015-8577-4_7.

[60] J. Dixmier. Étude sur les variétés et les opérateurs de Julia, avec quelques applications. *Bulletin de la Société Mathématique de France*, 77:11–101, 1949.

[61] P. Doukhan. *Mixing*, volume 85 of *Lecture Notes in Statistics*. Springer, 1994. ISBN: 978-0-387-94214-8. DOI: 10.1007/978-1-4612-2642-0.

[62] J. C. Duchi, P. W. Glynn, and H. Namkoong. Statistics of Robust Optimization: A Generalized Empirical Likelihood Approach. *Mathematics of Operations Research*, 43:835–1234, 3, 2021. DOI: 10.1287/moor.2020.1085.

[63] J.-M. Dufour. Distribution and quantile functions. *McGill University Report*, 1995.

[64] C. Durot and A.-S. Tocquet. Goodness of fit test for isotonic regression. *ESAIM:P&S*, 5:119–140, 2001.

[65] G. Ecoto, A. Bibault, and A. Chambaz. One-step ahead Super Learning from short time series of many slightly dependent data, and anticipating the cost of natural disasters. *arXiv:2107:13291*, 2021.

[66] U. Faigle and J. Voss. A system-theoretic model for cooperation, interaction and allocation. *Discrete Applied Mathematics*. 8th Cologne/Twente Workshop on Graphs and Combinatorial Optimization (CTW 2009), 159(16):1736–1750, 2011. ISSN: 0166-218X. DOI: 10.1016/j.dam.2010.07.007.

[67] E. Fekhari, B. Iooss, J. Muré, L. Pronzato, and J. Rendas. Model predictivity assessment: incremental test-set selection and accuracy evaluation. In N. Salvati, C. Perna, S. Marchetti, and R. Chambers, editors, *Studies in Theoretical and Applied Statistics, SIS 2021, Pisa, Italy, June 21-25*, pages 315–347. Springer, 2023.

[68] B. E Feldman. A dual model of cooperative value. *Available at SSRN 317284*, 2002. DOI: 10.2139/ssrn.317284.

[69] B. E. Feldman. A Theory of Attribution. *SSRN Electronic Journal*, 2007. ISSN: 1556-5068. DOI: 10.2139/ssrn.988860.

[70] B. E. Feldman. Relative Importance and Value. *SSRN Electronic Journal*, 2005. ISSN: 1556-5068. DOI: 10.2139/ssrn.2255827.

[71] B. E. Feldman. The proportional value of a cooperative game. *Manuscript. Chicago: Scudder Kemper Investments*, 1999.

[72] I. Feshchenko. When is the sum of closed subspaces of a Hilbert space closed?, 2020. DOI: 10.48550/arXiv.2012.08688.

[73] I. S. Feshchenko. On closeness of the sum of n subspaces of a Hilbert space. *Ukrainian Mathematical Journal*, 63(10):1566–1622, 2012. ISSN: 1573-9376. DOI: 10.1007/s11253-012-0601-9.

[74] I. Florescu and C. A. Tudor. *Handbook of Probability*. Wiley Handbooks in Applied Statistics. John Wiley & Sons, Ltd, 2013. ISBN: 978-0-470-64727-1.

[75] J. C. Fort, T. Klein, and A. Lagnoux. Global Sensitivity Analysis and Wasserstein Spaces. *SIAM/ASA Journal on Uncertainty Quantification*, 9(2):880–921, 2021. DOI: 10.1137/20M1354957.

[76] A. Foucault, M. Il Idrissi, B. Iooss, and S. Ancelet. Shapley effects and proportional marginal effects for global sensitivity analysis: application to computed tomography scan organ dose estimation. Preprint, 2023.

[77] S. Fredenhagen, H. J. Oberle, and G. Opfer. On the Construction of Optimal Monotone Cubic Spline Interpolations. *Journal of Approximation Theory*, 96(2):182–201, 1999. ISSN: 00219045. DOI: 10.1006/jath.1998.3247.

[78] T. Freiesleben and T. Grote. Beyond generalization: a theory of robustness in machine learning. *Synthese*, 202(4):109, 2023. ISSN: 1573-0964. DOI: 10.1007/s11229-023-04334-9. (Visited on 11/08/2023).

[79] K. Friedrichs. On Certain Inequalities and Characteristic Value Problems for Analytic Functions and For Functions of Two Variables. *Transactions of the American Mathematical Society*, 41(3):321–364, 1937. ISSN: 0002-9947. DOI: 10.2307/1989786.

[80] C. Frogner, C. Zhang, H. Mobahi, M. Araya, and T.A. Poggio. Learning with a Wasserstein loss. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

[81] A. Fu, B. Narasimhan, and S. Boyd. CVXR: an R package for disciplined convex optimization. *Journal of Statistical Software*, 94(14):1–34, 2020. DOI: 10.18637/jss.v094.i14.

[82] S. Fu, M. Couplet, and N. Bousquet. An adaptive kriging method for solving nonlinear inverse statistical problems. *Environmetrics*, 28(4):e2439, 2017. ISSN: 1099-095X. DOI: 10.1002/env.2439.

[83] Y. Funaki. Dual axiomatizations of solutions of cooperative games, 1996.

[84] A. Galántai. *Projectors and Projection Methods*. Springer US, 2004. ISBN: 978-1-4613-4825-2. DOI: 10.1007/978-1-4419-9180-5.

[85] F. Gamboa, A. Janon, T. Klein, and A. Lagnoux. Sensitivity indices for multivariate outputs. *Comptes Rendus Mathematique*, 351(7):307–310, 2013. ISSN: 1631-073X. DOI: 10.1016/j.crma.2013.04.016.

[86] C. Gauchy, J. Stenger, R. Sueur, and B. Iooss. An information geometry approach to robustness analysis for the uncertainty quantification of computer codes. *Technometrics*, 64:80–91, 2022. DOI: 10.1080/00401706.2021.1905072.

[87] H. Gebelein. Das statistische Problem der Korrelation als Variations- und Eigenwertproblem und sein Zusammenhang mit der Ausgleichsrechnung. *ZAMM - Zeitschrift für Angewandte Mathematik und Mechanik*, 21(6):364–379, 1941. ISSN: 00442267, 15214001. DOI: 10.1002/zamm.19410210604.

[88] A.L. Gibbs and F. E. Su. On choosing and bounding probability metrics. *International Statistical Review / Revue Internationale de Statistique*, 70(3):419–435, 2002. ISSN: 03067734, 17515823.

[89] T. Goda. A simple algorithm for global sensitivity analysis with Shapley effects. *Reliability Engineering & System Safety*, 213:107702, 2021. ISSN: 09518320. DOI: 10.1016/j.ress.2021.107702.

[90] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[91] U. Grömping. Estimators of relative importance in linear regression based on variance decomposition. *The American Statistician*, 61(2), 2007.

[92] U. Grömping. Variable importance in regression models. *Wiley Interdisciplinary Reviews: Computational Statistics*, 7:137–152, 2015.

[93] Shimodaira; H. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000. ISSN: 0378-3758. DOI: https://doi.org/10.1016/S0378-3758(00)00115-4.

[94] M. Hallin, E. del Barrio, J. Cuesta-Albertos, and C. Matrán. Distribution and quantile functions, ranks and signs in dimension d: A measure transportation approach. *The Annals of Statistics*, 49(2):1139–1165, 2021. ISSN: 0090-5364, 2168-8966. DOI: 10.1214/20-AOS1996.

[95] J. C. Harsanyi. A simplified bargaining model for the n-person Cooperative Game. *International Economic Review*, 4(2):194–220, 1963. ISSN: 0020-6598. DOI: 10.2307/2525487. Publisher: [Economics Department of the University of Pennsylvania, Wiley, Institute of Social and Economic Research, Osaka University].

[96] J. Hart and P. A. Gremaud. An approximation theoretic perspective of Sobol' indices with dependent variables. *International Journal for Uncertainty Quantification*, 8(6), 2018.

[97] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer: New York, 2009. ISBN: 978-0-387-84857-0. DOI: 10.1007/978-0-387-84858-7.

[98] D. Hendrycks, S. Basart, N. Mu, S. Kadavath, F. Wang, E. Dorundo, R. Desai, T. Zhu, S. Parajuli, M. Guo, D. Song, J. Steinhardt, and J. Gilmer. The many faces of robustness: a critical analysis of out-of-distribution generalization. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8320–8329, 2021. DOI: 10.1109/ICCV48922.2021.00823.

[99] M. Herin. Proportional values: an alternative to Shapley values in sensitivity analysis. Msc Internship Dissertation, EDF R&D, 2021.

[100] M. Herin, M. Il Idrissi, V. Chabridon, and B. Iooss. Proportional marginal effects for global sensitivity analysis. *SIAM/ASA Journal on Uncertainty Quantification*, 2024. In press.

[101] S. Hilal. *Thermo-mechanical modelling of the Wire Arc Additive Manufacturing process (WAAM)*. Ph.D Thesis, Université Paris sciences et lettres, 2022.

[102] W. Hoeffding. A Class of Statistics with Asymptotically Normal Distribution. *The Annals of Mathematical Statistics*, 19(3):293–325, 1948. ISSN: 0003-4851, 2168-8990. DOI: 10.1214/aoms/1177730196.

[103] G. Hooker. Diagnosing extrapolation: tree-based density estimation. In W. Kim and R. Kohavi, editors, *Proceedings of the tenth ACM SIGKDD International conference on Knowledge Discovery and Data Mining (KDD '04)*, pages 569–574, 2004.

[104] G. Hooker. *Diagnostic and extrapolation in machine learning*. PhD thesis, Stanford University, 2004.

[105] G. Hooker. Generalized Functional ANOVA Diagnostics for High-Dimensional Functions of Dependent Variables. *Journal of Computational and Graphical Statistics*, 16(3):709–732, 2007.

[106] G. Hooker and S. Rosset. Prediction-based regularization using data augmented regression. *Statistics and Computing*, 22:237–249, 2012.

[107] M. Il Idrissi, N. Bousquet, F. Gamboa, B. Iooss, and J. M. Loubes. Coalitional decomposition of quantities of interest. In *2023 Annual Meeting of MASCOT-NUM Research Group*, 2023.

[108] M. Il Idrissi, N. Bousquet, F. Gamboa, B. Iooss, and J. M. Loubes. Cooperative game theory and importance quantification. In *23rd European Young Statisticians Meeting of the Bernoulli Society*, 2023.

[109] M. Il Idrissi, N. Bousquet, F. Gamboa, B. Iooss, and J. M. Loubes. Hoeffding decomposition of black-box models with dependent inputs. Preprint, 2023.

[110] M. Il Idrissi, N. Bousquet, F. Gamboa, B. Iooss, and J. M. Loubes. Hoeffding decomposition, revisited. In *SIAM Conference on Uncertainty Quantification 2024*, 2024.

[111] M. Il Idrissi, N. Bousquet, F. Gamboa, B. Iooss, and J. M. Loubes. On the coalitional decomposition of parameters of interest. *Comptes Rendus. Mathématique*, 361:1653–1662, 2023. DOI: 10.5802/crmath.521.

[112] M. Il Idrissi, N. Bousquet, F. Gamboa, B. Iooss, and J. M. Loubes. Projection de mesures de probabilité sous contraintes de quantile par distance de Wasserstein et approximation monotone polynomiale. In *53èmes Journées de Statistique de la Société Française de Statistique (SFdS)*, 2022.

[113] M. Il Idrissi, N. Bousquet, F. Gamboa, B. Iooss, and J. M. Loubes. Quantile-constrained Wasserstein projections for robust interpretability of numerical and machine learning models. Preprint, 2023.

[114] M. Il Idrissi, N. Bousquet, F. Gamboa, B. Iooss, and J. M. Loubes. Robustness assessment of black-box models using quantile-constrained wasserstein projections. In *2022 Annual Meeting of MASCOT-NUM Research Group*, 2022. (Poster).

[115] M. Il Idrissi, V. Chabridon, and B. Iooss. Developments and applications of Shapley effects to reliability-oriented sensitivity analysis with correlated inputs. *Environmental Modelling and Software*, 143:105115, 2021. ISSN: 1364-8152. DOI: 10.1016/j.envsoft.2021.105115.

[116] M. Il Idrissi, V. Chabridon, and B. Iooss. Shapley effects for reliability-oriented sensitivity analysis with correlated inputs. In *Proceedings of the 10th International Conference on Sensitivity Analysis of Model Output*, 2022.

[117] M. Il Idrissi, M. Hérin, and V. Chabridon. Cooperative game theory and global sensitivity analysis. École Thématique sur les Incertitudes en Calcul Scientifique (ETICS), Erdeven, France, 2021.

[118] M. Il Idrissi, B. Iooss, and V. Chabridon. Mesures d'importance relative par décomposition de la performance de modèles de régression. In *52èmes Journées de Statistique de la Société Française de Statistique (SFdS)*, 2021.

[119] B. Iooss, R. Kennet, and P. Secchi. Different views of interpretability. In A. Lepore, B. Palumbo, and J. M. Poggi, editors, *Interpretability for Industry 4.0: Statistical and Machine Learning Approaches*. Springer, 2022. ISBN: 978-3-031-12401-3. DOI: https://doi.org/10.1007/978-3-031-12402-0.

[120] B. Iooss and P. Lemaître. A review on global sensitivity analysis methods. In C. Meloni and G. Dellino, editors, *Uncertainty management in Simulation-Optimization of Complex Systems: Algorithms and Applications*. Springer, 2015.

[121] B. Iooss and C. Prieur. Shapley effects for Sensitivity Analysis with correlated inputs : Comparisons with Sobol' Indices, Numerical Estimation and Applications. *International Journal for Uncertainty Quantification*, 9(5):493–514, 2019.

[122] J. Jacod and P. Protter. *Probability Essentials*. Universitext. Springer, 2004. ISBN: 978-3-540-43871-7. DOI: 10.1007/978-3-642-55682-1.

[123] H. Joe. *Multivariate Models and Multivariate Dependence Concepts*. Chapman and Hall/CRC, 1997. ISBN: 978-0-367-80389-6. DOI: 10.1201/9780367803896.

[124] M. I. Jordan and T. M. Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, 2015. DOI: 10.1126/science.aaa8415. Publisher: American Association for the Advancement of Science.

[125] E. Kalai and D. Samet. On weighted Shapley values. *International Journal of Game Theory*, 16(3):205–222, 1987. ISSN: 1432-1270.

[126] O. Kallenberg. *Foundations of Modern Probability*. Probability theory and stochastic modelling. Springer, 2021. ISBN: 978-3-030-61871-1. DOI: 10.1007/978-3-030-61871-1.

[127] B. Ketema, F. Costantino, F. Gamboa, R. Sueur, N. Bousquet, and B. Iooss. Robustness analysis for uncertainty quantification by optimization on riemannian manifolds. In *2023 Annual Meeting of MASCOT-NUM Research Group*, 2023. Poster.

[128] M. Koklu and Y. S. Taspinar. Determining the Extinguishing Status of Fuel Flames With Sound Wave by Machine Learning Methods. *IEEE Access*, 9:86207–86216, 2021. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2021.3088612.

[129] R. A. Koyak. On Measuring Internal Dependence in a Set of Random Variables. *The Annals of Statistics*, 15(3):1215–1228, 1987. ISSN: 0090-5364, 2168-8966. DOI: 10.1214/aos/1176350501.

[130] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[131] D. Krueger, E. Caballero, J.-H. Jacobsen, A. Zhang, J. Binas, D. Zhang, R. Le Priol, and A. Courville. Out-of-distribution generalization via risk extrapolation (rex). In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 5815–5826, 2021.

[132] M. Kuhn and K. Johnson. *Applied Predictive Modeling*. Springer, 2013. ISBN: 978-1-4614-6848-6. DOI: 10.1007/978-1-4614-6849-3.

[133] J. P. S. Kung, G. C. Rota, and C. Hung Yan. *Combinatorics: the Rota way*. Cambridge University Press, 2012. ISBN: 978-0-511-80389-5. OCLC: 1226672593.

[134] F. Y. Kuo, I. H. Sloan, G. W. Wasilkowski, and H. Woźniakowski. On decompositions of multivariate functions. *Mathematics of Computation*, 79(270):953–966, 2009. ISSN: 0025-5718. DOI: 10.1090/S0025-5718-09-02319-9.

[135] J-B. Lasserre. *An Introduction to Polynomial and Semi-Algebraic Optimization*. Cambridge Texts in Applied Mathematics. Cambridge University Press, 2015. ISBN: 978-1-107-06057-9. DOI: 10.1017/CBO9781107447226.

[136] S.L. Lauritzen. *Graphical Models*. Oxford Statistical Science Series. Oxford University Press, 1996. ISBN: 978-0-19-852219-5.

[137] R. Lebrun and A. Dutfoy. A generalization of the Nataf transformation to distributions with elliptical copula. *Probabilistic Engineering Mechanics*, 24(2):172–178, 2009. ISSN: 0266-8920. DOI: 10.1016/j.probengmech.2008.05.001.

[138] R. Lebrun and A. Dutfoy. Do Rosenblatt and Nataf isoprobabilistic transformations really differ? *Probabilistic Engineering Mechanics*, 24(4):577–584, 2009. ISSN: 0266-8920. DOI: 10.1016/j.probengmech.2009.04.006.

[139] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. ISSN: 1558-2256. DOI: 10.1109/5.726791. Conference Name: Proceedings of the IEEE.

[140] P. Lemaître. *Analyse de sensibilité en fiabilité des structures*. PhD thesis, Université de Bordeaux, 2014.

[141] P. Lemaître, E. Sergienko, A. Arnaud, N. Bousquet, F. Gamboa, and B. Iooss. Density modification-based reliability sensitivity analysis. *Journal of Statistical Computation and Simulation*, 85(6):1200–1223, 2015. DOI: 10.1080/00949655.2013.873039.

[142] A. Lepore, B. Palumbo, and J. M. Poggi, editors. *Interpretability for Industry 4.0 : Statistical and Machine Learning Approaches*. Springer International Publishing, 2022. ISBN: 978-3-031-12401-3. DOI: 10.1007/978-3-031-12402-0.

[143] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis. Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy*, 23(1):18, 2020. ISSN: 1099-4300. DOI: 10.3390/e23010018.

[144] R. H. Lindeman, P. F. Merenda, and R. Z. Gold. *Introduction to Bivariate and Multivariate Analysis*. Scott, Foresman, 1980. ISBN: 978-0-673-15099-8.

[145] K. Liu, H. Kargupta, and J. Ryan. Random projection-based multiplicative data perturbation for privacy preserving distributed data mining. *IEEE Transactions on Knowledge and Data Engineering*, 18:92–106, 2006.

[146] R. Liu and A. B. Owen. Estimating Mean Dimensionality of Analysis of Variance Decompositions. *Journal of the American Statistical Association*, 101(474):712–721, 2006. ISSN: 0162-1459. Publisher: [American Statistical Association, Taylor & Francis, Ltd.]

[147] P-L Loh and M. J. Wainwright. Structure estimation for discrete graphical models: Generalized covariance matrices and their inverses. *The Annals of Statistics*, 41(6), 2013. ISSN: 0090-5364. DOI: 10.1214/13-AOS1162.

[148] X. Lu and E. Borgonovo. Is time to intervention in the COVID-19 outbreak really important? A global sensitivity analysis approach. *Preprint*, 2020. arXiv:2005.01833.

[149] S. M. Lundberg and S-I. Lee. A Unified Approach to Interpreting Model Predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[150] P. Malliavin. *Integration and Probability*, volume 157 of *Graduate Texts in Mathematics*. Springer, 1995. ISBN: 978-1-4612-8694-3. DOI: 10.1007/978-1-4612-4202-4.

[151] T. A. Mara, S. Tarantola, and P. Annoni. Non-parametric methods for global sensitivity analysis of model output with dependent inputs. *Environmental Modelling & Software*, 72:173–183, 2015. ISSN: 1364-8152. DOI: 10.1016/j.envsoft.2015.07.010.

[152] G. Mazo and L. Tournier. An inference method for global sensitivity analysis. Preprint, 2023.

[153] S. A. Mead, S. J. Gezan, A. Clark, and S. J. Welham. *Statistical Methods in Biology: Design and Analysis of Experiments and Regression*. Chapman and Hall/CRC, 2014. ISBN: 978-0-429-11298-0. DOI: 10.1201/b17336.

[154] A. Meurer, C.P. Smith, M. Paprocki, O. Čertík, S.B. Kirpichev, M. Rocklin, A. Kumar, S. Ivanov, J.K. Moore, S. Singh, T. Rathnayake, S. Vig, B.E. Granger, R.P. Muller, F. Bonazzi, H. Gupta, S. Vats, F. Johansson, F. Pedregosa, M.J. Curry, A.R. Terrel, Š. Roučka, A. Saboo, I. Fernando, S. Kulal, R. Cimrman, and A. Scopatz. Sympy: symbolic computing in python. *PeerJ Computer Science*, 3:e103, 2017. ISSN: 2376-5992. DOI: 10.7717/peerj-cs.103.

[155] A.F. Möbius. Über eine Besondere Art von Umkehrung der Reihen. *Journal für die reine und angewandte Mathematik*, 9:105–123, 1832.

[156] C. Molnar. *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*. leanpub.com, 1st edition, 2021.

[157] S-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1765–1773, 2017.

[158] J. Morio. Global and local sensitivity analysis methods for a physical system. *European Journal of Physics*, 32(6):1577–1583, 2011. ISSN: 0143-0807, 1361-6404. DOI: 10.1088/0143-0807/32/6/011.

[159] R. K. Mothilal, A. Sharma, and C. Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, pages 607–617. Association for Computing Machinery, 2020. ISBN: 978-1-4503-6936-7. DOI: 10.1145/3351095.3372850.

[160] H. Moulin. *Fair division and collective welfare*. MIT Press, 1. mit press paperback ed edition, 2004. ISBN: 978-0-262-63311-6.

[161] K. Murray, S. Müller, and B. A. Turlach. Fast and flexible methods for monotone polynomial fitting. *Journal of Statistical Computation and Simulation*, 86(15):2946–2966, 2016. ISSN: 0094-9655. DOI: 10.1080/00949655.2016.1139582.

[162] A. Narayan and D. Xiu. Distributional sensitivity for uncertainty quantification. *Communications in Computational Physics*, 10(1):140–160, 2011. ISSN: 2152-7385, 2152-7393. DOI: 10.4236/am.2012.312A289.

[163] P. Natekar and M. Sharma. Representation Based Complexity Measures for Predicting Generalization in Deep Learning, 2020. DOI: 10.48550/arXiv.2012.02775. arXiv:2012.02775 [cs].

[164] R. B. Nelsen. *An introduction to copulas*. Springer series in statistics (2nd edition). Springer, 2006. ISBN: 978-0-387-28659-4.

[165] B. Neyshabur, S. Bhojanapalli, D. McAllester, and N. Srebro. Exploring generalization in deep learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 5949–5958, 2017. ISBN: 9781510860964.

[166] B. O'Donoghue, E. Chu, N. Parikh, and S. Boyd. Conic Optimization via Operator Splitting and Homogeneous Self-Dual Embedding. *Journal of Optimization Theory and Applications*, 169(3):1042–1068, 2016.

[167] K. M. Ortmann. The proportional value for positive cooperative games. *Mathematical Methods of Operations Research (ZOR)*, 51(2):235–248, 2000.

[168] M.J. Osborne and A. Rubinstein. *A Course in Game Theory*. MIT Press, 1994, pages 291–294.

[169] A. B. Owen. Sobol' Indices and Shapley Value. *SIAM/ASA Journal on Uncertainty Quantification*, 2(1):245–251, 2014. ISSN: 2166-2525. DOI: 10.1137/130936233.

[170] A. B. Owen and C. Prieur. On Shapley value for measuring importance of dependent inputs. *SIAM/ASA Journal on Uncertainty Quantification*, 5(1):986–1002, 2017.

[171] P. A. Parrilo. Algebraic Optimization and Semidefinite Optimization. *MIT Lectures Notes (EIDMA Minicourse)*, 2010.

[172] P. A. Parrilo. Polynomial optimization, sums of squares, and applications. In *Semidefinite Optimization and Convex Algebraic Geometry*, pages 47–157. SIAM, 2012. DOI: 10.1137/1.9781611972290.ch3.

[173] M.K. Paul, M.R. Islam, and Sattar S. An efficient perturbation approach for multivariate data in sensitive and reliable data mining. *Journal of Information Security and Applications*, 62:102954, 2021. ISSN: 2214-2126. DOI: https://doi.org/10.1016/j.jisa.2021.102954.

[174] J. Pelamatti and V. Chabridon. Sensitivity Analysis in the Presence of Hierarchical Variables. In *Programme and abstracts of the 23th Annual Conference of the European Network for Business and Industrial Statistics (ENBIS)*, volume 1, page 84. Department of Applied Statistics, Operational Research, and Quality, Universitat Politecnica de Valencia, 2023. ISBN: 978-84-12-54449-7.

[175] G. Perrin and G. Defaux. Efficient Evaluation of Reliability-Oriented Sensitivity Indices. *Journal of Scientific Computing*, 79:1433–1455, 2019.

[176] S.M. Pesenti. Reverse Sensitivity Analysis for Risk Modelling. *Risks*, 10:141, 2022.

[177] E. Plischke and E. Borgonovo. Copula theory and probabilistic sensitivity analysis: Is there a connection? *European Journal of Operational Research*, 277(3):1046–1059, 2019. ISSN: 0377-2217. DOI: https://doi.org/10.1016/j.ejor.2019.03.034.

[178] E. Plischke, G. Rabitti, and E. Borgonovo. Computing Shapley Effects for Sensitivity Analysis. *SIAM/ASA Journal on Uncertainty Quantification*, 9(4):1411–1437, 2021. DOI: 10.1137/19M1304738. Publisher: Society for Industrial and Applied Mathematics.

[179] H. Rabitz and Ö. Aliş. General foundations of high-dimensional model representations. *Journal of Mathematical Chemistry*, 25(2):197–233, 1999. ISSN: 1572-8897. DOI: 10.1023/A:1019188517934.

[180] H. Raguet and A. Marrel. Target and conditional sensitivity analysis with emphasis on dependence measures. *Working paper*, 2018.

[181] D. S Rakic and D. S Djordjevic. A note on topological direct sum of subspaces. *Funct. Anal. Approx. Comput*, 10(1):9–20, 2018.

[182] S. Razavi, A. Jakeman, A. Saltelli, C. Prieur, B. Iooss, E. Borgonovo, E. Plischke, S. Lo Piano, T. Iwanaga, W. Becker, S. Tarantola, J.H.A. Guillaume, J. Jakeman, H. Gupta, N. Melillo, G. Rabitti, V. Chabridon, Q. Duan, X. Sun, S. Smith, R. Sheikholeslami, N. Hosseini, M. Asadzadeh, A. Puy, S. Kucherenko, and H.R. Maier. The Future of Sensitivity Analysis: An essential discipline for systems modeling and policy support. *Environmental Modelling and Software*, 137:104954, 2021. ISSN: 1364-8152. DOI: https://doi.org/10.1016/j.envsoft.2020.104954.

[183] A. Rényi. On measures of dependence. *Acta Mathematica Academiae Scientiarum Hungarica*, 10(3):441–451, 1959. ISSN: 1588-2632. DOI: 10.1007/BF02024507.

[184] S. I. Resnick. *A Probability Path*. Birkhäuser Boston, 2014. ISBN: 978-0-8176-8408-2. DOI: 10.1007/978-0-8176-8409-9.

[185] S. I. Resnick. Preliminaries. In S. I. Resnick, editor, *Extreme Values, Regular Variation and Point Processes*, Springer Series in Operations Research and Financial Engineering, pages 1–37. Springer, 1987. ISBN: 978-0-387-75953-1. DOI: 10.1007/978-0-387-75953-1_1.

[186] M. T. Ribeiro, S. Singh, and C. Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 1135–1144. Association for Computing Machinery, 2016. ISBN: 978-1-4503-4232-2. DOI: 10.1145/2939672.2939778.

[187] G. C. Rota. On the foundations of combinatorial theory I. Theory of Möbius Functions. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 2(4):340–368, 1964. ISSN: 1432-2064. DOI: 10.1007/BF00531932.

[188] O. Roustant, D. Ginsbourger, and Y. Deville. DiceKriging, DiceOptim: Two R packages for the analysis of computer experiments by kriging-based metamodeling and optimization. *Journal of Statistical Software*, 21:1–55, 2012.

[189] C.J. Roy and W.L. Oberkampf. A comprehensive framework for verification, validation, and uncertainty quantification in scientific computing. *Computer Methods in Applied Mechanics and Engineering*, 200(25):2131–2144, 2011. ISSN: 0045-7825. DOI: https://doi.org/10.1016/j.cma.2011.03.016.

[190] W. Rudin. *Functional analysis*. International series in pure and applied mathematics. McGraw-Hill, 2. ed. Edition, 1996. ISBN: 978-0-07-100944-7.

[191] F. Rupin, G. Blatman, S. Lacaze, T. Fouquet, and B. Chassignole. Probabilistic approaches to compute uncertainty intervals and sensitivity factors of ultrasonic simulations of a weld inspection. *Ultrasonics*, 54:1037–1046, 2013.

[192] J. Sacks, W.J. Welch, T.J. Mitchell, and H.P. Wynn. Design and analysis of computer experiments. *Statistical Science*, 4:409–435, 1989.

[193] A. Saltelli, G. Bammer, I. Bruno, E. Charters, M. Di Fiore, et al. Five ways to ensure that models serve society: a manifesto (short comments). *Nature*, 582:482–484, 2020.

[194] F. Santambrogio. *Optimal Transport for Applied Mathematicians*, volume 87 of *Progress in Nonlinear Differential Equations and Their Applications*. Springer International Publishing, 2015. ISBN: 978-3-319-20827-5. DOI: 10.1007/978-3-319-20828-2.

[195] A. Sasane. *A Friendly Approach to Functional Analysis*. WORLD SCIENTIFIC (EUROPE), 2017. ISBN: 978-1-78634-333-8. DOI: 10.1142/q0096.

[196] J. W. Schmidt and W. Heß. Positivity of cubic polynomials on intervals and positive spline interpolation. *BIT Numerical Mathematics*, 28(2):340–352, 1988. ISSN: 0006-3835, 1572-9125. DOI: 10.1007/BF01934097.

[197] B. S. W. Schröder. *Ordered Sets*. Birkhäuser, 2003. ISBN: 978-1-4612-6591-7. DOI: 10.1007/978-1-4612-0053-6.

[198] B. B. Seiler. *Applications of cooperative game theory to interpretable machine learning*. Ph.D thesis, Stanford University, 2023.

[199] G. Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, 1976. ISBN: 978-0-691-10042-5. DOI: 10.2307/j.ctv10vm1qb.

[200] L. S. Shapley. *A value for n-person games*. In *Contributions to the Theory of Games (AM-28), Volume II*. Harold William Kuhn and Albert William Tucker, editors. Princeton University Press, 2016, pages 307–318. DOI: doi:10.1515/9781400881970-018.

[201] L. S. Shapley. Notes on the n-Person Game – II: The Value of an n-Person Game. Research Memorandum ATI 210720, RAND Corporation, 1951.

[202] Z. Shen, J. Liu, Y. He, X. Zhang, R. Xu, H. Yu, and P. Cui. Towards Out-Of-Distribution Generalization: A Survey. *arXiv:2108.13624*, 2021. DOI: 10.48550/ARXIV.2108.13624.

[203] Z. Sidák. On Relations Between Strict-Sense and Wide-Sense Conditional Expectations. *Theory of Probability & Its Applications*, 2(2):267–272, 1957. ISSN: 0040-585X. DOI: 10.1137/1102020.

[204] I.M Sobol. Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Mathematics and Computers in Simulation*, 55(1):271–280, 2001. ISSN: 03784754.

[205] O. Sobrie, N. Gillis, V. Mousseau, and M. Pirlot. UTA-poly and UTA-splines: Additive value functions with polynomial marginals. *European Journal of Operational Research*, 264(2):405–418, 2018. ISSN: 0377-2217. DOI: 10.1016/j.ejor.2017.03.021.

[206] E. Song, B. L. Nelson, and J. Staum. Shapley Effects for Global Sensitivity Analysis: Theory and Computation. *SIAM/ASA Journal on Uncertainty Quantification*, 4(1):1060–1083, 2016. ISSN: 2166-2525. DOI: 10.1137/15M1048070.

[207] T. Speith. A review of taxonomies of explainable artificial intelligence (xai) methods. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 2239–2250, 2022.

[208] E. Spiegel and C. J. O'Donnell. *Incidence algebras*, number 206 in Monographs and textbooks in pure and applied mathematics. M. Dekker, 1997. ISBN: 978-0-8247-0036-2.

[209] C. J. Stone. The Use of Polynomial Splines and Their Tensor Products in Multivariate Function Estimation. *The Annals of Statistics*, 22(1):118–171, 1994. ISSN: 0090-5364, 2168-8966. DOI: 10.1214/aos/1176325361. Publisher: Institute of Mathematical Statistics.

[210] J. Stufken. Letters to the Editor: on hierarchical partitioning. *The American Statistician*, 46(1):70–77, 1992. ISSN: 0003-1305, 1537-2731. DOI: 10.1080/00031305.1992.10475852.

[211] Y. S. Taspinar, M. Koklu, and M. Altin. Acoustic-Driven Airflow Flame Extinguishing System Design and Analysis of Capabilities of Low Frequency in Different Fuels. *Fire Technology*, 58(3):1579–1597, 2022. ISSN: 1572-8099. DOI: 10.1007/s10694-021-01208-9.

[212] Y. S. Taspinar, M. Koklu, and M. Altin. Classification of flame extinction based on acoustic oscillations using artificial intelligence methods. *Case Studies in Thermal Engineering*, 28:101561, 2021. ISSN: 2214-157X. DOI: 10.1016/j.csite.2021.101561.

[213] N. Tripuraneni, B. Adlam, and J. Pennington. Overparameterization improves robustness to covariate shift in high dimensions. In *35th Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

[214] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 2000. ISBN: 978-1-4757-3264-1. DOI: 10.1007/978-1-4757-3264-1.

[215] V. Vasil'ev and G. van der Laan. The Harsanyi Set for Cooperative TU-Games. Working Paper 01-004/1, Tinbergen Institute Discussion Paper, 2001.

[216] O. Vasseur, A. Azarian, V. Jolivet, and P. Bourdon. Capability of high intrinsic quality Space Filling Design for global sensitivity analysis and metamodelling of interference optical systems. *Chemometrics and Intelligent Laboratory Systems*, 113:10–18, 2012.

[217] O. Vasseur, M. Claeys-Bruno, M. Cathelinaud, and M. Sergent. High-dimensional sensitivity analysis of complex optronic systems by experimental design: applications to the case of the design and the robustness of optical coatings. *Chinese Optics Letters*, 8(s1):21–24, 2010.

[218] I. Verdinelli and L. Wasserman. Feature Importance: A Closer Look at Shapley Values and LOCO, 2023. DOI: 10.48550/arXiv.2303.05981. arXiv:2303.05981 [stat].

[219] C. Villani. *Topics in Optimal Transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, 2003. ISBN: 978-0-8218-3312-4. DOI: 10.1090/gsm/058.

[220] J. Wang, C. Lan, C. Liu, Y. Ouyang, and T. Qin. Generalizing to unseen domains: a survey on domain generalization. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4627–4635, 2021. DOI: 10.24963/ijcai.2021/628.

[221] X. Wang and F. Li. Isotonic Smoothing Spline Regression. *Journal of Computational and Graphical Statistics*, 17(1):21–37, 2008. ISSN: 1061-8600. DOI: 10.1198/106186008X285627.

[222] L. Wasserman. *All of Statistics: A Concise Course in Statistical Inference*. Springer Texts in Statistics. Springer, 2004. ISBN: 978-1-4419-2322-6. DOI: 10.1007/978-0-387-21736-9.

[223] R. J. Weber. Probabilistic values for games. In A. E. Roth, editor, *The Shapley value: essays in honor of Lloyd S. Shapley*, chapter 7, pages 101–120. Cambridge University Press, 1988.

[224] R. Yousefzadeh. Deep Learning Generalization and the Convex Hull of Training Sets. *arXiv:2101.09849*, 2021.

[225] M. Zondervan-Zwijnenburg, W. van de Schoot-Hubeek, K. Lek, H. Hoijtink, and R. van de Schoot. Application and Evaluation of an Expert Judgment Elicitation Procedure for Correlations. *Frontiers in Psychology*, 8:90, 2017.

APPENDIX A

MEASURE AND PROBABILITY THEORY PRELIMINARIES

The interested reader is referred to, e.g., [150, 184, 74, 126] for a thorough primer on probability and measure theory. The definitions and results introduced below have been extracted from these references for the developments presented in the manuscript.

A few definitions

Definition A.1 (Standard Borel measurable space). A measurable space $(E, \mathcal{E})$ is said to be standard Borel if E is a Polish space (i.e., a separable, completely metrizable topological space) and $\mathcal{E}$ is the subsequent Borel σ-algebra on E, w.r.t. to its metric.

Definition A.2 (σ-algebra generated by a mapping). Let $(E_1, \mathcal{E}_1)$ and $(E_2, \mathcal{E}_2)$ be two measurable spaces. For a mapping $T : E_1 \to E_2$, the σ-algebra generated by T is the set:

$$\sigma_T := \{T^{-1}(A), \forall A \in \mathcal{E}_2\},$$

where T^{-1} denotes the *inverse image* of T.

Definition A.3 (Measurable mapping). Let $(E_1, \mathcal{E}_1)$ and $(E_2, \mathcal{E}_2)$ be two measurable spaces. A mapping $T : E_1 \to E_2$ is said to be *measurable* if

$$\sigma_T \subset \mathcal{E}_1.$$

Additionally, for any sub-σ-algebra $\mathcal{G} \subseteq \mathcal{E}_1$, T is said to be *$\mathcal{G}$-measurable* if

$$\sigma_T \subset \mathcal{G}.$$

Definition A.4 (Random element, random variable). Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $(E, \mathcal{E})$ be a measurable space. A mapping $T : \Omega \to E$ is called *random element* if it is measurable. In particular, if $E = \mathbb{R}$, T is called *random variable*.

Additionally, if E is the Cartesian product of at least two Polish spaces, T is called *a vector of random elements*, and in particular, if $E = \mathbb{R}^d$ for a positive integer d, T is called a *random vector*.

Definition A.5 (Induced probability measure). Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $(E, \mathcal{E})$ be a measurable space. The *probability measure induced by a measurable mapping* $T : \Omega \to E$, denoted $P_T : \mathcal{E} \to [0, 1]$ is defined, $\forall A \in \mathcal{E}$ as:

$$P_T(A) := \mathbb{P}\left(T^{-1}(A)\right).$$

Definition A.6 ($\mathbb{P}$-trivial σ-algebra). Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. The *$\mathbb{P}$-trivial σ-algebra*, denoted $\sigma_\emptyset$, is defined as the smallest (coarsest) sub-σ-algebra of $\mathcal{F}$ containing the null-set w.r.t. the measure $\mathbb{P}$, i.e., the set of elements of $\mathcal{F}$ with measure 0.

Some useful results

Lemma A.1 ($\sigma_\emptyset$-measurable implies almost-sure constance). *Let X be a random variable. If X is $\sigma_\emptyset$-measurable, then there exists a $c \in \mathbb{R}$ such that*

$$X = c \quad \textit{almost surely.}$$

Proof: see, [184] Lemma 4.5.1.

Lemma A.2 (Functional representation and measurability (Doob-Dynkin)). *Let Y be an A-valued random element, and let X be an E-valued random element. If Y is σ_X-measurable, then there exists a measurable function $f : E \to A$ such that:*

$$Y = f(X) \text{ a.s.}$$

Proof: see, [126], Lemma 1.14

Appendix B

Supplementary material for Chapter 2

B.1 Preliminaries on sets and orders

B.1.1 Some group theory

Definition B.1 (Group). Let N be a set, and $\circ$ be an operation on N. $(N, \circ)$ is said to be a *group* if it satisfies the four following conditions:

- $\forall a, b \in N$, $a \circ b \in N$ (Closure);
- $\forall a, b, c \in N$, $a \circ (b \circ c) = (a \circ b) \circ c$ (Associativity);
- $\exists e \in N$ such that $\forall a \in N$, $e \circ a = a = a \circ e$ (Identity);
- $\forall a \in N$, $\exists b \in N$ such that $a \circ b = e = b \circ a$ (Inverse).

Definition B.2 (Abelian group). Let $(\mathbb{A}, +)$ be a group. If, in addition, the operation $+$ respects the condition

- $\forall a, b \in \mathbb{A}$, $a + b = b + a$ (Commutativity);

then $(\mathbb{A}, +)$ is said to be an *abelian group*.

Definition B.3 (Commutative ring with identity). A commutative ring with identity is a triplet $(\mathbb{A}, +, \times)$, where $+$ is usually called *addition* and $\times$ *multiplication*. $(\mathbb{A}, +)$ is an abelian group, and $\times$ respects closure, associativity, commutativity, identity, and is distributive w.r.t. $+$ on $\mathbb{A}$. However, one does not require $\times$ to admit a *multiplicative inverse* (i.e., an inverse w.r.t. to $\times$).

B.1.2 Some order theory

Definition B.4 (Partially ordered set). Let N be a set, and $\preceq$ be a binary relation. If $\preceq$ is:

- Reflexive, i.e., for any $a \in N$, $a \preceq a$;
- Transitive, i.e., for any $a, b, c \in N$, if $a \preceq b$ and $b \preceq c$, then $a \preceq c$;
- Antisymetric, i.e., for any $a, b \in N$, if $a \preceq b$ and $b \preceq a$, then $a = b$ (where "$=$" represent equivalence).

then $(N, \preceq)$ is said to be a *partially ordered set* (or poset).

Definition B.5 (Totally ordered set). Let $(N, \preceq)$ be a partially ordered set. If, in addition, for any $a, b \in N$, either $a \preceq b$ or $b \preceq a$ (i.e., any two elements can be compared), then $(N, \preceq)$ is said to be a *totally ordered set*.

Definition B.6 (Order isomorphism). Let $(N, \preceq)$ and $(M, \leq)$ be two partially ordered sets. A bijective mapping $\phi : N \to M$ is said to be an *order isomorphism* if

$$\forall a, b \in N, \quad a \preceq b \iff \phi(a) \leq \phi(b).$$

If there exists an order isomorphism between two partially ordered sets, then both sets are said to be *order isomorphic* (i.e., the order structures are the same).

Order isomorphisms rely on a bijection between the two sets. However, the requirement of the mapping to be bijective can be relaxed. Essentially, order embeddings can be understood as mapping an initial set into a subset of another set, while preserving the order. For instance, a total order on a finite set can be embedded into a total order on an infinite set (i.e., mapped to a finite subset of the infinite set), but it can never be isomorphic to it.

Definition B.7 (Order embedding). Let $(N, \preceq)$ and $(M, \leq)$ be two partially ordered sets. A mapping $\phi : N \to M$ is said to be an *order embedding* if

$$\forall a, b \in N, \quad a \preceq b \iff \phi(a) \leq \phi(b).$$

If there exists an order embedding between N and M, then N is said to be embedded into M (i.e., the order structure of N is preserved on a subset of M). In other words, N is order isomorphic to the image of ϕ (which is not necessarily equal to M).

Definition B.8 (Locally finite partially ordered set). Let $(N, \preceq)$ be a partially ordered set. For every $a, b \in N$, the sets

$$[c \in N : a \preceq c \preceq b],$$

are cal called *segments of* N.

If every segment of N is finite, then N is said to be *locally finite*.

B.1.3 Rota's generalization of the Möbius inversion formula

Definition B.9 (Incidence algebra). Let $(N, \preceq)$ be a locally finite partially ordered set, and let $(\mathbb{A}, +, \times)$ be a commutative ring with identity. Define the set of functions from the cartesian product of N by itself, valued :

$$I_{\mathbb{A}}(N) := \{f : N \times N \to \mathbb{A} \ : \ f(a, b) = 0 \text{ if } a \not\preceq b\}.$$

The triplet $(I_{\mathbb{A}}(N), +_I, \star)$ where, for any $f, g \in I_{\mathbb{A}}$ $a, b \in N$ and $\alpha \in \mathbb{A}$:

$$\begin{aligned}(f +_I g)(a, b) &= f(a, b) + g(a, b)\\ (f \star g)(a, b) &= \sum_{a \preceq z \preceq b} f(a, z) \times g(z, b)\\ (\alpha \star f)(x, y) &= \alpha \times f(a, b)\end{aligned}$$

is called the *incidence algebra of* $(N, \preceq)$ *over* $(\mathbb{A}, +, \times)$ (in short, the incidence algebra of N over $\mathbb{A}$). $+_I$ is called the addition of $I_{\mathbb{A}}(N)$, and $\star$ is called the convolution of $I_{\mathbb{A}}(N)$.

Definition B.10 (Zeta and Möbius functions). Let $(N, \preceq)$ be a partially ordered set, and let $(I_{\mathbb{A}}(N), +_I, \star)$ be the incidence algebra of N over a commutative ring with identity $\mathbb{A}$. The *zeta* function $\zeta \in I_{\mathbb{A}}(N)$, defined, for any $a, b \in N$ as

$$\zeta(a, b) = \begin{cases} 1_{\mathbb{A}} & \text{if } x = y, \\ 0_{\mathbb{A}} & \text{otherwise.} \end{cases},$$

where $1_{\mathbb{A}}$ is the multiplicative identity of $\mathbb{A}$ and $0_{\mathbb{A}}$ is its zero element. The zeta function is the *convolutional identity of* $I_{\mathbb{A}}(N)$, i.e., for any $f \in I_{\mathbb{A}}(N)$,

$$f \star \zeta = f.$$

The Möbius function $\mu \in I_{\mathbb{A}}(N)$ is defined as the *convolutional inverse of the zeta function on the incidence algebra of* N *over* $\mathbb{A}$. It is defined recursively, for any $a, b \in N$ with $a \preceq b$ as:

$$\mu(a, b) = \begin{cases} 1_{\mathbb{A}} & \text{if } x = y \\ -\sum_{a \preceq z \preceq b} \mu(a, z) & \text{otherwise.} \end{cases}$$

For any $f \in I_{\mathbb{A}}(N)$, $f \star \mu$ is called the *Möbius transform of* f.

Theorem B.1 (Rota's inversion formula). *Let N be any non-empty set and $(N, \preceq)$ be a locally finite partially ordered set, and let $(I_{\mathbb{A}}(N), +_I, \star)$ be its incidence algebra over a commutative ring with identity*

$(\mathbb{A}, +, \times)$. *Let* $v, \phi \in I_{\mathbb{A}}(N)$. *The following equivalence hold:*

$$v(a) = \sum_{z:z\leq a} \phi(z),\ \forall a \in N \quad \iff \quad \phi(a) = \sum_{z:z\leq a} v(z) \times \mu(z,a), \quad \forall a \in N,$$

where μ *is the Möbius function on* $I_{\mathbb{A}}(N)$.

Proof: See [133], p.108 or [208] Theorems 2.1.2 and 2.1.3.

B.2 Proofs

Proof of Proposition 2.2.
First, assume that $v(D) = \text{QoI}(G(X))$. By construction of ϕ, and by Corollary 2.1, one has that in particular

$$\text{QoI}(G(X)) = v(D) = \sum_{A\in\mathcal{P}_D} \phi(A),$$

and thus ϕ is a coalitional decomposition of $\text{QoI}(G(X))$.
Now assume that ϕ is a coalitional decomposition. One then has that,

$$\begin{aligned}\text{QoI}(G(X)) &= \sum_{A\in\mathcal{P}_D} \phi(A) \\ &= \sum_{A\in\mathcal{P}_D} \sum_{B\in\mathcal{P}_A} (-1)^{|A|-|B|} v(B) = v(D),\end{aligned}$$

using Corollary 2.1.

Appendix C

Supplementary material for Chapter 3

Contents

B.1 Preliminaries on sets and orders **118**
B.1.1 Some group theory 118
B.1.2 Some order theory 118
B.1.3 Rota's generalization of the Möbius inversion formula 119
B.2 Proofs **120**

C.1 Estimators for the conditional elements

C.1.1 Monte-Carlo estimator

This procedure, introduced in [206] for the estimation of Shapley effects, relies on a Monte Carlo estimation of the conditional elements. It requires the ability to sample from the marginal distributions of the inputs (i.e., P_{X_A} for all $A \subseteq \{1,\dots,d\} \setminus \emptyset$), as well as from all the conditional distributions (i.e., $P_{X_{\overline{A}}|X_A}$, for all possible subsets of inputs A). Additionally, one also needs to be able to evaluate the model G, which is usually the case in the context of uncertainty quantification of numerical computer models (ignoring the potential difficulties related to the cost of a single evaluation of $G(\cdot)$) [57].

In order to estimate a conditional element S^{clos}, one needs to randomly draw several i.i.d. samples:

- an i.i.d. sample of size N drawn from P_X and denoted by $(X^{(1)},\dots,X^{(N)})$;
- another i.i.d. sample of size N_v drawn from P_{X_A} and denoted by $(X_A^{(1)},\dots,X_A^{(N_v)})$;
- for each element $X_A^{(i)}, i = 1,\dots,N_v$, a corresponding sample of size N_p drawn from $P_{X_{\overline{A}}|X_A}$ given that $X_A = X_A^{(i)}$ and denoted by $(\widetilde{X}_i^{(1)},\dots,\widetilde{X}_i^{(N_p)})$.

Then, the Monte Carlo estimator of S^{clos} can be defined as:

$$\widehat{S^{\text{clos}}} = \frac{1}{N_v - 1}\sum_{i=1}^{N_v}\left(\frac{1}{N_p}\sum_{j=1}^{N_p} G(\widetilde{X}_i^{(j)}, X_A^{(i)}) - \overline{G(X)}\right)^2 \tag{C.1}$$

where

$$\overline{G(X)} = \frac{1}{N}\sum_{i=1}^{N} G(X^{(i)}). \tag{C.2}$$

Concerning the complexity in number of model evaluations, the interested reader is referred to the recent developments presented in [178].

C.1.2 Nearest-neighbor estimator

A given-data estimation method has been introduced by [33] to estimate the conditional elements. This method can be seen as an extension of the Monte Carlo estimator when only a single i.i.d. input-output sample is available. This method is appropriate when the input distributions are not known or when the numerical model G is not available. The main idea behind this method is to replace the exact samples from the conditional distributions $P_{X_{\overline{A}}|X_A}$ by approximated ones based on a non-parametric nearest-neighbor procedure.

Let $\left(X^{(1)},\dots,X^{(N)}\right)$ be an i.i.d. sample of the inputs X and $A \in \mathcal{P}_d \setminus \{\emptyset, \{1:d\}\}$. Let $k_N^A(l,n)$ be the index such that $X_A^{\left(k_N^A(l,n)\right)}$ is the n-th closest element to $X_A^{(l)}$ in $\left(X_A^{(1)},\dots,X_A^{(N)}\right)$. Note that, if two or more observations are at an equal distance from $X_A^{(l)}$, then one of them is uniformly randomly selected. Then, one can express an estimator for S^T:

$$\widehat{S^T}_{A,\text{KNN}} = \frac{1}{N}\sum_{l=1}^{N}\left(\frac{1}{N_s - 1}\sum_{i=1}^{N_s}\left[G\left(X^{\left(k_N^{\overline{A}}(l,i)\right)}\right) - \frac{1}{N_s}\sum_{h=1}^{N_s} G\left(X^{\left(k_N^{\overline{A}}(l,h)\right)}\right)\right]^2\right). \tag{C.3}$$

Under some mild assumptions, [33] showed that this estimator does asymptotically converge towards S^T.

This method is less computationally expensive (in terms of model evaluations) compared to the Monte Carlo sampling-based method, since no additional model evaluation, other than the ones in the i.i.d. sample, is required in order to produce estimates of the target Shapley effects. Since the samples of the conditional and marginal distributions are approximated by a non-parametric procedure, this method also reduces the possible input modeling error (e.g., in the context of ill-defined input distributions), at the cost of less accurate estimates. Another constraint is due to the fact that the input-output sample has to

be i.i.d. which prevents it from being used, for instance, in advanced orthogonal designs of computer experiments.

C.2 Proofs

Proof of Theorem 3.2.
Let (D, v) be a nonnegative, monotonic cooperative game. For any non-empty coalition $A \subseteq D$, and denote $|A|$ its cardinality.

Denote $\mathcal{S}_A$ the set of permutations of players in A. Let $\pi \in \mathcal{S}_A$, and for the sake of clarity, denote $|\pi| = |A|$, i.e., the number of elements in the permutation. Moreover, by convention, $|\emptyset| = 0$ and $v\left(C_0(\pi)\right) = v(\emptyset) = 0$ for any $\pi \in \mathcal{S}_D$.

By monotonicity, $\forall j \in \{0, \ldots, |\pi| - 1\}$, notice that,

$$0 \leq v\left(C_j(\pi)\right) \leq v\left(C_{j+1}(\pi)\right).$$

Moreover, for any permutation $\pi \in \mathcal{S}_A$, define:

$$k_\pi(v) = \max\ \{j \in \{0, \ldots, |\pi|\} : v(C_j(\pi)) = 0\}.$$

For the sake of conciseness and readability, the argument v is omitted and the notation $k_\pi := k_\pi(v)$ is adopted.

Now, let $(\epsilon_p)_{p\in\mathbb{N}}$ be a sequence such that:

$$\forall p \in \mathbb{N}, \epsilon_p > 0, \quad \text{and} \quad \lim_{p\to\infty} \epsilon_p = 0.$$

Let $(v_p)_{p\in\mathbb{N}}$ be a sequence of positive, monotonic value functions acting on the set of players D and defined, for any $p \in \mathbb{N}$ and for any $A \subseteq D$, as:

$$v_p(A) = \begin{cases} \epsilon_p & \text{if } v(A) = 0, \\ v(A) & \text{otherwise.} \end{cases}$$

Alternatively, one can notice that, for any $A \subseteq D$, for every permutation $\pi \in \mathcal{S}_A$, and for every $j \in \{0, \ldots, |\pi|\}$, one has that

$$v_p\big(C_j(\pi)\big) = \begin{cases} \epsilon_p & \text{if } j \leq k_\pi, \\ v(C_j(\pi)) & \text{otherwise.} \end{cases} \tag{C.4}$$

Let $p \in \mathbb{N}$, and from the recursive definition of PV (see, Eq. (3.9)), for the positive games (D, v_p), one has, for every $i \in D$:

$$\mathrm{PV}_i = \frac{\sum_{\pi\in\mathcal{S}_{D_{-i}}} \prod_{m=1}^{d-1} v_p\left(C_m(\pi)\right)^{-1}}{\sum_{\sigma\in\mathcal{S}_D} \prod_{m=1}^{d} v_p\left(C_m(\sigma)\right)^{-1}}.$$

For the sake of conciseness and clarity, and for every permutation $\pi \in \mathcal{S}_A$ such that $A \subseteq D$, and for $l, k \in 1, \ldots, d_1$, denote:

$$\Upsilon_k^l(\pi, v) = \begin{cases} \prod_{j=k}^{l} v\left(C_j(\pi)\right)^{-1} & \text{if } k \leq l, \\ 1 & \text{otherwise.} \end{cases}$$

One then has that, for any $i \in D$:

$$\begin{aligned} \mathrm{PV}_i &= \frac{\sum_{\pi\in\mathcal{S}_{D_{-i}}} \Upsilon_1^{d-1}(\pi, v_p)}{\sum_{\sigma\in\mathcal{S}_D} \Upsilon_1^{d}(\sigma, v_p)} = \frac{\sum_{\pi\in\mathcal{S}_{D_{-i}}} \Upsilon_1^{k_\pi}(\pi, v_p)\Upsilon_{k_\pi+1}^{d-1}(\pi, v_p)}{\sum_{\sigma\in\mathcal{S}_D} \Upsilon_1^{k_\sigma}(\sigma, v_p)\Upsilon_{k_\sigma+1}^{d}(\sigma, v_p)} \\ &= \frac{\sum_{\pi\in\mathcal{S}_{D_{-i}}} \epsilon_p^{-k_\pi}\Upsilon_{k_\pi+1}^{d-1}(\pi, v_p)}{\sum_{\sigma\in\mathcal{S}_D} \epsilon_p^{-k_\sigma}\Upsilon_{k_\sigma+1}^{d}(\sigma, v_p)}. \end{aligned}$$

From Eq. (C.4), for any permutation $\pi \in \mathcal{S}_A$ such that $A \subseteq D$, one has that:

$$\Upsilon_1^{k_\pi}(\pi, v_p) = \prod_{j=1}^{k_\pi} v_p\big(C_j(\pi)\big)^{-1} = \epsilon_p^{-k_\pi},$$

Denote, for any $j \in D$, $k_{\max}^{-j}$ the size of the largest zero coalition in D_{-j}, i.e.,

$$k_{\max}^{-j} = \max_{A \in \mathcal{P}(D_{-j})} \{|A| : v(A) = 0\},$$

and let $k_{\max}$ be the size of the largest zero coalition in D, and notice that necessarily,

$$\forall j \in D, \forall \pi \in \mathcal{S}_{D_{-j}}, \quad k_{\max}^{-j} \leq k_{\max}. \tag{C.5}$$

Moreover, denote, for any $j \in D \cup \{\emptyset\}$, the two following sets of permutations:

$$\mathcal{R}_{\max}^{-j} = \{\pi \in \mathcal{S}_{D_{-j}} : k_\pi = k_{\max}^{-j}\}, \quad \text{and} \quad \mathcal{R}^{-j} = \{\pi \in \mathcal{S}_{D_{-j}} : k_\pi < k_{\max}^{-j}\}$$

and notice that $\mathcal{R}_{\max}^{-j} \cup \mathcal{R}^{-j} = \mathcal{S}_{D_{-j}}$. Moreover, notice that

$$\forall j \in D \cup \emptyset, \forall \pi \in \mathcal{R}^{-j}, \quad k_\pi < k_{\max}. \tag{C.6}$$

Again, denote $\mathcal{R}_{\max} = \mathcal{R}_{\max}^{-\emptyset}$ and $\mathcal{R} = \mathcal{R}^{-\emptyset}$. Hence, for any $j \in D$, one has that:

$$\begin{aligned}\sum_{\pi \in \mathcal{S}_{D^{-j}}} \epsilon_p^{-k_\pi} \Upsilon_{k_\pi+1}^{d-1}(\pi, v) &= \sum_{\pi \in \mathcal{R}_{\max}^{-j}} \epsilon_p^{-k_\pi} \Upsilon_{k_\pi+1}^{d-1}(\pi, v) + \sum_{\pi \in \mathcal{R}^{-j}} \epsilon_p^{-k_\pi} \Upsilon_{k_\pi+1}^{d-1}(\pi, v) \\ &= \sum_{\pi \in \mathcal{R}_{\max}^{-j}} \epsilon_p^{-k_{\max}^{-j}} \Upsilon_{k_\pi+1}^{d-1}(\pi, v) + \sum_{\pi \in \mathcal{R}^{-j}} \epsilon_p^{-k_\pi} \Upsilon_{k_\pi+1}^{d-1}(\pi, v)\end{aligned}$$

and in the particular case of $j = \emptyset$, one has that:

$$\begin{aligned}\sum_{\pi \in \mathcal{S}_D} \epsilon_p^{-k_\pi} \Upsilon_{k_\pi+1}^{d}(\pi, v) &= \sum_{\pi \in \mathcal{R}_{\max}} \epsilon_p^{-k_{\max}} \Upsilon_{k_\pi+1}^{d}(\pi, v) + \sum_{\pi \in \mathcal{R}} \epsilon_p^{-k_\pi} \Upsilon_{k_\pi+1}^{d}(\pi, v) \\ &= \epsilon_p^{-k_{\max}} \left(\sum_{\pi \in \mathcal{R}_{\max}} \Upsilon_{k_\pi+1}^{d}(\pi, v) + \sum_{\pi \in \mathcal{R}} \epsilon_p^{k_{\max}-k_\pi} \Upsilon_{k_\pi+1}^{d}(\pi, v) \right).\end{aligned}$$

It entails that:

$$\mathrm{PV}_i = \frac{\sum_{\pi \in \mathcal{R}_{\max}^{-i}} \epsilon_p^{k_{\max}-k_{\max}^{-i}} \Upsilon_{k_\pi+1}^{d-1}(\pi, v) + \sum_{\pi \in \mathcal{R}^{-i}} \epsilon_p^{k_{\max}-k_\pi} \Upsilon_{k_\pi+1}^{d-1}(\pi, v)}{\sum_{\sigma \in \mathcal{R}_{\max}} \Upsilon_{k_\sigma+1}^{d}(\sigma, v) + \sum_{\sigma \in \mathcal{R}} \epsilon_p^{k_{\max}-k_\sigma} \Upsilon_{k_\sigma+1}^{d}(\sigma, v)}.$$

From Eq. (C.6), one can notice that, for any $j \in D \cup \{\emptyset\}$:

$$\lim_{p \to \infty} \sum_{\pi \in \mathcal{R}^{-j}} \epsilon_p^{k_{\max}-k_\pi} \Upsilon_{k_\pi+1}^{d-|j|}(\pi, v) = 0$$

and additionally, from Eq. (C.5), notice that for any $j \in D$:

$$\lim_{p \to \infty} \sum_{\pi \in \mathcal{R}_{\max}^{-j}} \epsilon_p^{k_{\max}-k_{\max}^{-j}} \Upsilon_{k_\pi+1}^{d-1}(\pi, v) = \begin{cases} \sum_{\pi \in \mathcal{R}_{\max}^{-j}} \Upsilon_{k_\pi+1}^{d-1}(\pi, v) & \text{if } k_{\max} = k_{\max}^{-j} \\ 0 & \text{otherwise.} \end{cases}$$

Denote:

$$\mathrm{PV}^*\big((D, v)\big) = \lim_{p \to \infty} \mathrm{PV}\big(D, v_p\big),$$

and notice that, for any $i \in D$:

$$\mathrm{PV}_i^* = \begin{cases} \dfrac{\sum_{\pi \in \mathcal{R}_{\max}^{-i}} \Upsilon_{k_\pi+1}^{d-1}(\pi, v)}{\sum_{\sigma \in \mathcal{R}_{\max}} \Upsilon_{k_\sigma+1}^{d-1}(\sigma, v)} & \text{if } k_{\max} = k_{\max}^{-i}, \\ 0 & \text{otherwise.} \end{cases}$$

Notice that, for any $j \in D$, the condition $k_{\max} = k_{\max}^{-j}$ is equivalent to having a coalition $A \subseteq D_{-j}$ such that $|A| = k_{\max}$ and $v(A) = 0$. The complement of this condition is that j must be in every coalition $A \subseteq D$ such that $|A| = k_{\max}$ and $v(A) = 0$, leading to the condition for which $\mathrm{PV}_j^* = 0$.

For any $j \in D \cup \{\emptyset\}$, and assuming that $k_{\max}^{-j} = k_{\max}$, one can notice that $\mathcal{R}_{\max}^{-j}$ only contains the permutations $\pi \in \mathcal{S}_{D_{-j}}$ such that $v\left(C_{k_{\max}}(\pi)\right) = 0$, and by monotonicity, this implies that for any $\pi \in \mathcal{R}_{\max}^{-j}$:

$$v\left(C_1(\pi)\right) = v\left(C_2(\pi)\right) = \cdots = v\left(C_{k_{\max}-1}(\pi)\right) = v\left(C_{k_{max}}(\pi)\right) = 0,$$

and that for $k_{\max} < k \leq |\pi|$,

$$v\left(C_k(\pi)\right) > 0.$$

For any $j \in D \cup \{\emptyset\}$, denote $\mathcal{K}_{-j} = \{A \subseteq D_{-j} : v(A) = 0 \text{ and } |A| = k_{\max}\}$, and notice that $\mathcal{R}_{\max}^{-j}$ is necessarily composed of permutations having permutations of elements in $\mathcal{K}_{-j}$ as their first $k_{\max}$ elements. In other words, for every $\pi \in \mathcal{R}_{\max}^{-j}$,

$$C_{k_{\max}}(\pi) \in \mathcal{K}_{-j}.$$

Thus, for any $j \in D \cup \emptyset$, one has that:

$$\begin{aligned}
\sum_{\pi \in \mathcal{R}_{\max}^{-j}} \Upsilon_{k_\pi+1}^{d-1}(\pi, v) &= \sum_{A \in \mathcal{K}_{-j}} k_{\max}! \sum_{\pi \in \mathcal{S}_{D_{-j} \setminus A}} \Upsilon_1^{|\pi|}(\pi, v_A) \\
&= k_{\max}! \sum_{A \in \mathcal{K}_{-j}} \sum_{\pi \in \mathcal{S}_{D_{-j} \setminus A}} \prod_{k=1}^{|\pi|} v(A \cup C_k(\pi))^{-1} \\
&= k_{\max}! \sum_{A \in \mathcal{K}_{-j}} R(D_{-j} \setminus A, v_A)^{-1}
\end{aligned}$$

where for any $B \subseteq D \setminus A$, $v_A(B) = v(A \cup B)$, and using results from [69] on the ratio potential. This leads to the following rewriting of PV^*, for any $i \in D$:

$$\mathrm{PV}_i^* = \begin{cases} 0 & \text{if } \forall A \in \mathcal{K}, i \in A, \\ \dfrac{\sum_{A \in \mathcal{K}_{-i}} R(D_{-i} \setminus A, v_A)^{-1}}{\sum_{A \in \mathcal{K}} R(D \setminus A, v_A)^{-1}} & \text{otherwise.} \end{cases}$$

Finally, notice that for any positive game (D, v), i.e., , where v is positively valued, then necessarily, for any permutation and sub-permutations π of players $k_\pi = k_{\max} = 0$. Moreover, for any $j \in D \cup \{\emptyset\}$, $\mathcal{K}_{-j} = \emptyset$, and for any $i \in D$,

$$\mathrm{PV}_i^* = \frac{R(D, v)}{R(D_{-i}, v)} = \mathrm{PV}_i,$$

and hence the allocation PV^* is a continuous extension of PV to cooperative games with nonnegative value function.

Proof of Lemma 3.1.
Let $A \subseteq D$. First, assume that $S_A^T = 0$, then necessarily,

$$\mathbb{E}_{-A}\left[\left(G(X) - \mathbb{E}_{-A}\left[G(X)\right]\right)^2\right] = 0 \text{ a.s.}$$

which can only be attained, by non-negativity of the squared distance, if

$$G(X) = \mathbb{E}\left[G(X) \mid X_{\overline{A}}\right] \text{ a.s.}$$

Now assume that $G(X) = \mathbb{E}_{-A}\left[G(X)\right]$ a.s.. Then necessarily,

$$S_A^T = \mathbb{E}_{-A}\left[\left(G(X) - \mathbb{E}_{-A}\left[G(X)\right]\right)^2\right] = 0 \text{ a.s..}$$

Proof of Proposition 3.2.

First, the equivalence

$$S_E^T = 0 \iff X_E \text{ is an exogenous,}$$

is proven.

Let $E \subseteq D$ and suppose that X_E is an exogenous vector. Then it entails that there exists some $f(X_{-E}) \in \mathbb{L}^2(\sigma_{-E})$ such that:

$$G(X) = f(X_{-E}) \text{ a.s.}$$

and hence, $G(X) \in \mathbb{L}^2(\sigma_{-E})$. Thus, since the conditional expectation $\mathbb{E}_{-E}[\cdot]$ is the orthogonal projection onto $\mathbb{L}^2(\sigma_{-E})$, one has that:

$$G(X) = \mathbb{E}_{-E}[G(X)],$$

and by Lemma 3.1, it is equivalent to

$$S_E^T = 0.$$

Now, suppose that $S_E^T = 0$, and thus, by Lemma 3.1, $G(X) = \mathbb{E}_{-E}[G(X)]$ a.s. and since $\mathbb{E}_{-E}[G(X)] \in \mathbb{L}^2(\sigma_{-E})$ then X_E is necessarily an exogenous vector. Thus, the between X_E being an exogenous vector and $S_E^T = 0$ is proved.

Suppose that for every $i \in E$, X_i is an exogenous input, and for every $j \in -E$, X_j is not exogenous. Suppose that Assumption 1 hold. Then X_E is an exogenous vector containing every exogenous input, and then, from the previous equivalence, $S_E^T = 0$.

Now, suppose that there exists another subset $A \subseteq D$, such that $E \subset A$ and $S_A^T = 0$. Necessarily one has that $A \setminus E \neq \emptyset$. Moreover, for any $i \in A \setminus E$, one has $0 \leq S_i^T \leq S_A^T = 0$ since S^T is monotonic w.r.t. set inclusion. Then $S_i^T = 0$, which from the previous equivalence, entails that X_j is exogenous. However, since $j \notin E$, this is impossible since X_E contains every exogenous input. Hence, there is no coalition A, bigger than E such that $S_A^T = 0$.

Now, from the value function S^T, this entails that

$$\mathcal{K} = \left\{ A \in \mathcal{P}_D \ : \ |A| = \max_{B \in \mathcal{P}_D} \{|B|\} \text{ s.t. } S_B^T = 0 \right\} = \{E\},$$

and thus, coming from Theorem 3.2,

$$\text{PME}_i = 0 \iff i \in \bigcap_{A \in \mathcal{K}} A = E,$$

and thus, an input is granted a PME equal to zero if and only if it is exogenous.

Appendix D

Supplementary material for Chapter 4

Contents

C.1 Estimators for the conditional elements 122
C.1.1 Monte-Carlo estimator 122
C.1.2 Nearest-neighbor estimator 122
C.2 Proofs 123

D.1 Some technical preliminaries

D.1.1 Hilbert space external direct sums

Let $H_1, \dots, H_n$ be a collection of Hilbert spaces with respective inner products $\langle .,. \rangle_1, \dots, \langle .,. \rangle_n$ and induced norms $\|.\|_1, \dots, \|.\|_n$. The Hilbert space (external) direct-sum is the space denoted and defined as (see, [45], Definition 6.4)

$$\bigoplus_{i=1}^{n} H_i = \left\{ x = (x_1, \dots, x_n) \in \bigtimes_{i=1}^{n} H_i : \sum_{i=1}^{n} \|x_i\|_i^2 < \infty \right\}.$$

A Hilbert space direct-sum is itself a Hilbert space w.r.t. the inner product $\langle .,. \rangle$ (see, [45], Proposition 6.2)

$$\forall x, y \in \bigoplus_{i=1}^{n} H_i, \quad \langle x, y \rangle = \sum_{i=1}^{n} \langle x_i, y_i \rangle_i .$$

D.1.2 Closed range operator

Theorem D.1 (Closed range operator). *Let $(\mathcal{M}_1, \|.\|_1)$ and $(\mathcal{M}_2, \|.\|_2)$ be two Banach spaces, and let $T : \mathcal{M}_1 \to \mathcal{M}_2$ be a continuous operator between the two spaces. T is bounded from below, i.e., there exists some $\gamma > 0$ such that, $\forall x \in \mathcal{M}_1$*

$$\|T(x)\|_2 \geq \gamma \|x\|_1 ,$$

if and only if T is one-to-one and $\mathrm{Ran}(T)$ is closed in $\mathcal{M}_2$.

Proof: See, [1], Theorem 2.5.

D.2 Proof of Theorem 4.6

D.2.1 Intermediary results

In order to prove Theorem 4.6, two preliminary results are required.

Proposition D.1. *Let $A \in \mathcal{P}_D$, and let $B, C \in \mathcal{P}_{-A}$ be non-empty proper subsets of A such that $B \neq C$. Let V_B, V_C be a closed subspace of $\mathbb{L}^2(\sigma_B)$ and $\mathbb{L}^2(\sigma_C)$ respectively. If one has that*

$$V_B \subseteq \left[\mathbb{L}^2(\sigma_{B \cap C})\right]^{\perp}, \quad \text{and } V_C \subseteq \left[\mathbb{L}^2(\sigma_{B \cap C})\right]^{\perp},$$

then, assuming that Assumption 2 hold, then

$$c_0(V_B, V_C) \leq c\left(\mathbb{L}^2(\sigma_B), \mathbb{L}^2(\sigma_C)\right).$$

Proof of Proposition D.1.
First, recall that, if Assumption 2 holds and thanks to Theorem 4.1

$$\mathbb{L}^2(\sigma_B) \cap \mathbb{L}^2(\sigma_C) = \mathbb{L}^2(\sigma_B \cap \sigma_C) = \mathbb{L}^2(\sigma_{B \cap C}).$$

Then, notice that since

$$V_B \subseteq \mathbb{L}^2(\sigma_B) \cap \left[\mathbb{L}^2(\sigma_{B \cap C})\right]^{\perp}, \quad \text{and } V_C \subseteq \mathbb{L}^2(\sigma_C) \cap \left[\mathbb{L}^2(\sigma_{B \cap C})\right]^{\perp},$$

one has that

$$\begin{aligned} c_0(V_B, V_C) &= c_0\left(\mathbb{L}^2(\sigma_B) \cap V_B, \mathbb{L}^2(\sigma_C) \cap V_C\right) \\ &\leq c_0\left(\mathbb{L}^2(\sigma_B) \cap \left[\mathbb{L}^2(\sigma_{B \cap C})\right]^{\perp}, \mathbb{L}^2(\sigma_C) \cap \left[\mathbb{L}^2(\sigma_{B \cap C})\right]^{\perp}\right). \end{aligned}$$

Hence, if Assumption 2 is assumed

$$\begin{aligned} c_0\left(V_B, V_C\right) &\leq c_0\left(\mathbb{L}^2\left(\sigma_B\right) \cap \left[\mathbb{L}^2\left(\sigma_B\right) \cap \mathbb{L}^2\left(\sigma_C\right)\right]^{\perp}, \mathbb{L}^2\left(\sigma_C\right) \cap \left[\mathbb{L}^2\left(\sigma_B\right) \cap \mathbb{L}^2\left(\sigma_C\right)\right]^{\perp}\right) \\ &= c\left(\mathbb{L}^2\left(\sigma_B\right), \mathbb{L}^2\left(\sigma_C\right)\right) \end{aligned}$$

where the last equality is achieved using Lemma 4.2.

Proposition D.2. *Let $A \in \mathcal{P}_D$, and let $(V_B)_{B \in \mathcal{P}_A, B \neq A}$ be a collection of closed subspaces of $\mathbb{L}^2(\sigma_A)$ such that, $\forall B, C \in \mathcal{P}_{-A}, B \neq C$,*

$$c_0\left(V_B, V_C\right) \leq c\left(\mathbb{L}^2\left(\sigma_B\right), \mathbb{L}^2\left(\sigma_C\right)\right)$$

then, under Assumption 3, there exist a $\rho > 0$ such that, for any $\sum_{A \in \mathcal{P}_{-A}} Y_A \in +_{B \in \mathcal{P}_{-A}} V_B$

$$\sqrt{\mathbb{E}\left[\left(\sum_{B \in \mathcal{P}_{-A}} Y_A\right)^2\right]} \geq \rho \sum_{B \in \mathcal{P}_{-A}} \sqrt{\mathbb{E}\left[Y_A^2\right]},$$

and additionally,

$$\dotplus_{B \in \mathcal{P}_{-A}} V_B$$

is closed in $\mathbb{L}^2(\sigma_A)$.

Proof of Proposition D.2.
Let $H_A = \bigoplus_{B \in \mathcal{P}_A : B \neq A}^n V_A$ be the Hilbert space external direct-sum of the collection of closed (and thus Hilbert) subspaces $(V_B)_{B \in \mathcal{P}_A, B \neq A}$. Let T_A be the operator defined as

$$\begin{aligned} T_A : H_A &\to \mathbb{L}^2\left(\sigma_A\right) \\ Y = \left(Y_B\right)_{B \in \mathcal{P}_{-A}} &\mapsto \sum_{B \in \mathcal{P}_{-A}} Y_B \end{aligned}$$

and notice that

$$\operatorname{Ran}\left(T_A\right) = \dotplus_{B \in \mathcal{P}_{-A}} V_B \subseteq \mathbb{L}^2\left(\sigma_A\right).$$

One then has that

$$\begin{aligned} \mathbb{E}\left[\left(\sum_{B \in \mathcal{P}_{-A}} Y_B\right)^2\right] &= \sum_{B \in \mathcal{P}_{-A}} \mathbb{E}\left[Y_B^2\right] + \sum_{B, C \in \mathcal{P}_{-A} : B \neq C} \mathbb{E}\left[Y_A Y_B\right] \\ &\geq \sum_{B \in \mathcal{P}_{-A}} \mathbb{E}\left[Y_B^2\right] - \sum_{B, C \in \mathcal{P}_{-A} : B \neq C} c_0\left(V_A, V_B\right) \sqrt{\mathbb{E}\left[Y_A^2\right]} \sqrt{\mathbb{E}\left[Y_B^2\right]} \\ &\geq \sum_{B \in \mathcal{P}_{-A}} \mathbb{E}\left[Y_B^2\right] - \sum_{B, C \in \mathcal{P}_{-A} : B \neq C} c\left(\mathbb{L}^2\left(\sigma_A\right), \mathbb{L}^2\left(\sigma_B\right)\right) \sqrt{\mathbb{E}\left[Y_A^2\right]} \sqrt{\mathbb{E}\left[Y_B^2\right]} \end{aligned}$$

where the first inequality is achieved thanks to Theorem 4.3. Denote $E_A = \left(\sqrt{\mathbb{E}\left[Y_B^2\right]}\right)_{B \in \mathcal{P}_{-A}}$ and notice that

$$\sum_{B \in \mathcal{P}_{-A}} \mathbb{E}\left[Y_B^2\right] - \sum_{B, C \in \mathcal{P}_{-A} : B \neq C} c\left(\mathbb{L}^2\left(\sigma_A\right), \mathbb{L}^2\left(\sigma_B\right)\right) \sqrt{\mathbb{E}\left[Y_A^2\right]} \sqrt{\mathbb{E}\left[Y_B^2\right]} = E_A^{\top} \Delta|_A E_A$$

Denote λ_A the smallest eigenvalue of $\Delta|_A$, and notice that if Assumption 3 holds, $\Delta|_A$ is definite positive and $\lambda_A > 0$. Thus, one has that

$$\begin{aligned} E_A^{\top} \Delta|_A E_A &\geq \lambda_A E_A^{\top} E_A \\ &= \lambda_A \sum_{B \in \mathcal{P}_{-A}} \mathbb{E}\left[Y_A^2\right]. \end{aligned}$$

Hence, one has that

$$\sqrt{\mathbb{E}\left[\left(\sum_{B\in\mathcal{P}_{-A}} Y_B\right)^2\right]} \geq \sqrt{\lambda_A \sum_{B\in\mathcal{P}_{-A}} \mathbb{E}\left[Y_A^2\right]}$$
$$\geq \sqrt{\frac{\lambda_A}{2^d-1}} \sum_{B\in\mathcal{P}_{-A}} \sqrt{\mathbb{E}\left[Y_A^2\right]}$$

where the last inequality is achieved using Jensen's inequality. Hence, one has that, for any $Y \in H_A$

$$\sqrt{\mathbb{E}\left[T_A(Y)^2\right]} \geq \sqrt{\frac{\lambda_A}{2^d-1}} \sum_{B\in\mathcal{P}_{-A}} \sqrt{\mathbb{E}\left[Y_A^2\right]}$$

where $\sqrt{\frac{\lambda_A}{2^d-1}} > 0$, and $\sum_{B\in\mathcal{P}_A} \sqrt{\mathbb{E}\left[Y_A^2\right]}$ is the norm of Y product on H_A. Hence, by Theorem D.1,

$$\mathrm{Ran}\left(T_A\right) = \dotplus_{B\in\mathcal{P}_{-A}} V_B \text{ is closed in } \mathbb{L}^2\left(\sigma_A\right).$$

D.2.2 Proof of the main result

Proof of Theorem 4.6.
The proof is done in two steps. First, we prove by induction that, $\forall A \in \mathcal{P}_D$

$$\mathbb{L}^2\left(\sigma_A\right) = \dotplus_{C\in\mathcal{P}_A} V_C,$$

and then we show that the sum is indeed direct.

Statement Let $n = 1, \ldots, d-1$. We will show that if for every non-empty $B \in \mathcal{P}_D$, B such that $|B| = n$, one has that

- $\mathbb{L}^2\left(\sigma_C\right) = +_{Z\in\mathcal{P}_C} V_Z$ where $V_C = \left[+_{Z\in\mathcal{P}_{-C}} V_Z\right]^{\perp_C}$;

Then, for every $A \in \mathcal{P}_D$ such that $|A| = n+1$, it holds that

$$\mathbb{L}^2\left(\sigma_A\right) = \dotplus_{C\in\mathcal{P}_A} V_C \text{ where } V_A = \left[\dotplus_{Z\in\mathcal{P}_{-A}} V_Z\right]^{\perp_A}$$

Base case We start for $n = 1$. For any $i \in D$, denote $V_i = [V_\emptyset]^{\perp_i}$, and notice that since $V_\emptyset$ is closed in $\mathbb{L}^2\left(\sigma_i\right)$

$$\mathbb{L}^2\left(\sigma_i\right) = V_\emptyset \oplus V_i.$$

and notice that $\forall i \in D$,

$$V_i = \left[\mathbb{L}^2\left(\sigma_\emptyset\right)\right]^{\perp_i} \subseteq \left[\mathbb{L}^2\left(\sigma_\emptyset\right)\right]^{\perp},$$

by Lemma 4.3.

Next, consider the case where $n = 2$. Notice from the previous step that for any $i, j \in D$ such that $i \neq j$, notice that $\mathbb{L}^2\left(\sigma_{i\cap j}\right) = \mathbb{L}^2\left(\sigma_\emptyset\right)$, and thus one has that

$$V_i \subset \left[\mathbb{L}^2\left(\sigma_\emptyset\right)\right]^{\perp} \text{ and } V_j \subset \left[\mathbb{L}^2\left(\sigma_\emptyset\right)\right]^{\perp}.$$

Hence, assuming that Assumption 2 hold, from Proposition D.1, one can conclude that, for any $i, j \in D$ such that $i \neq j$,

$$c_0\left(V_i, V_j\right) \leq c\left(\mathbb{L}^2\left(\sigma_i\right), \mathbb{L}^2\left(\sigma_j\right)\right).$$

Now, let $A \in \mathcal{P}_D$ such that $|A| = 2$, and denote $A = \{i, j\}$, and notice that, assuming that Assumption 3 hold, by Proposition D.2, one has that

$$V_\emptyset + V_i + V_j \text{ is closed in } \mathbb{L}^2(\sigma_A).$$

Hence, let

$$V_A = [V_\emptyset + V_i + V_j]^{\perp_A},$$

and notice that

$$\mathbb{L}^2(\sigma_A) = [V_\emptyset + V_i + V_j] \oplus V_A.$$

Since A has been chosen arbitrarily, this holds for any $A \in \mathcal{P}_D$ such that $|A| = 2$.

Induction Suppose that, for every $B \in \mathcal{P}_D$ such that $|B| = n$, one has that

$$\mathbb{L}^2(\sigma_B) = \dotplus_{Z \in \mathcal{P}_B} V_Z, \text{ where } V_B = \left[\dotplus_{Z \in \mathcal{P}_{-B}} V_Z\right]^{\perp_B}.$$

Let $A \in \mathcal{P}_D$ such that $|A| = n + 1$. Notice then that, for any non-empty $B, C \in \mathcal{P}_{-A}$, since $B \cap C \in \mathcal{P}_{-B} \cap \mathcal{P}_{-C}$, that

$$\mathbb{L}^2(\sigma_{B\cap C}) = \dotplus_{Z \in \mathcal{P}_{B\cap C}} V_Z,$$

is necessarily contained of $\dotplus_{Z\in\mathcal{P}_{-B}} V_Z$ and of $\dotplus_{Z\in\mathcal{P}_{-C}} V_Z$. Thus, one has that

$$V_B = \left[\dotplus_{Z \in \mathcal{P}_{-B}} V_Z\right]^{\perp_B} \subset \left[\dotplus_{Z \in \mathcal{P}_{-B}} V_Z\right]^{\perp} \subset \left[\mathbb{L}^2(\sigma_{B\cap C})\right]^{\perp}.$$

and analogously

$$V_C \subset \left[\mathbb{L}^2(\sigma_{B\cap C})\right]^{\perp}.$$

Hence, assuming that Assumption 2 hold, from Proposition D.1, one can conclude that, for every non-empty $B, C \in \mathcal{P}_{-A}$ such that $B \neq C$,

$$c_0(V_B, V_C) \leq c\left(\mathbb{L}^2(\sigma_B), \mathbb{L}^2(\sigma_C)\right),$$

which, under Assumption 3 and thanks to Proposition D.2, in turn implies that

$$\dotplus_{Z \in \mathcal{P}_{-A}} V_Z \text{ is closed in } \mathbb{L}^2(\sigma_A).$$

Denote $V_A = \left[\dotplus_{Z\in\mathcal{P}_{-A}} V_Z\right]^{\perp_A}$, and notice that

$$\mathbb{L}^2(\sigma_A) = \left[\dotplus_{Z \in \mathcal{P}_{-A}} V_Z\right] \oplus V_A = \dotplus_{Z \in \mathcal{P}_A} V_Z.$$

Since A has been taken arbitrarily, this holds for any $A \in \mathcal{P}_D$ such that $|A| = n$.

Now, we show that these additive decompositions are direct. Let $A \in \mathcal{P}_D$, and notice that for any non-empty $\forall B \in \mathcal{P}_A$, $V_B \perp \mathbb{L}^2(\sigma_\emptyset)$, meaning that any $f(X_B) \in V_B$ is centered. Next, notice that the principal $\left(2^{|A|} \times 2^{|A|}\right)$ submatrix of Δ, indexed by the elements of $\mathcal{P}_A$ and denoted Δ_A, is also definite-positive, and hence its smallest eigenvalue λ_A is positive. Next, notice that for any $Y \in \mathbb{L}^2(\sigma_A)$, by definition, one has that:

$$Y = \sum_{B \in \mathcal{P}_A} Y_B, \quad \text{where } Y_B \in V_B.$$

Now, suppose that $Y = 0$ a.s., which is equivalent to $\mathbb{E}[Y] = 0$ and $\mathbb{E}\left[Y^2\right] = 0$. However, under Assumptions 2 and 3, notice that

$$\begin{aligned}
\mathbb{E}\left[Y^2\right] &= \mathbb{E}\left[\left(\sum_{B\in\mathcal{P}_A} Y_B\right)^2\right] \\
&= \sum_{B\in\mathcal{P}_A} \mathbb{E}\left[Y_B^2\right] + \sum_{B,C\in\mathcal{P}_A:B\neq C} \mathbb{E}\left[Y_B Y_C\right] \\
&\geq \sum_{B\in\mathcal{P}_A} \mathbb{E}\left[Y_B^2\right] - \sum_{B,C\in\mathcal{P}_A:B\neq C} c_0\left(V_B, V_C\right)\sqrt{\mathbb{E}\left[Y_B^2\right]}\sqrt{\mathbb{E}\left[Y_C^2\right]} \\
&\geq \sum_{B\in\mathcal{P}_A} \mathbb{E}\left[Y_B^2\right] - \sum_{B,C\in\mathcal{P}_A:B\neq C} c\left(\mathbb{L}^2\left(\sigma_B\right), \mathbb{L}^2\left(\sigma_C\right)\right)\sqrt{\mathbb{E}\left[Y_B^2\right]}\sqrt{\mathbb{E}\left[Y_C^2\right]}
\end{aligned}$$

Let $E_A = \left(\sqrt{\mathbb{E}\left[Y_B^2\right]}\right)^\top_{B\in\mathcal{P}_A}$ and notice that

$$\begin{aligned}
\mathbb{E}\left[Y^2\right] &\geq E_A^\top \Delta_A E_A \\
&\geq \lambda_A E_A^\top E_A \\
&= \lambda_A \sum_{B\in\mathcal{P}_A} \mathbb{E}\left[Y_B^2\right]
\end{aligned}$$

since Δ_A is definite positive, and $\lambda_A > 0$ is its smallest eigenvalue. Thus, one has that if $\mathbb{E}\left[Y^2\right] = 0$, then necessarily

$$\sum_{B\in\mathcal{P}_A} \mathbb{E}\left[Y_B^2\right] = 0,$$

and since this is a sum of positive elements, $\forall B \in \mathcal{P}_A$, $\mathbb{E}\left[Y_B^2\right] = 0$, which, in addition to the fact that each Y_B is centered, is equivalent to $Y_B = 0$ a.s. Hence,

$$Y = 0 \text{ a.s.} \implies \forall B \in \mathcal{P}_D, \quad Y_B = 0 \text{ a.s.}$$

which ultimately proves that

$$\mathbb{L}^2\left(\sigma_A\right) = \bigoplus_{B\in\mathcal{P}_A} V_B.$$

D.3 Proofs

Proof of Proposition 4.11.
First, recall that, for any $A \in \mathcal{P}_D$,

$$G_A(X_A) = \sum_{B\in\mathcal{P}_A} (-1)^{|A|-|B|}\mathbb{M}_B(G(X)),$$

and hence,

$$\begin{aligned}
\sum_{B\in\mathcal{P}_A} (-1)^{|A|-|B|}\mathrm{Cov}\left(\mathbb{M}_B(G(X)), [I-\mathbb{M}_A](G(X))\right) &= \mathrm{Cov}\left(G_A(X_A), [I-\mathbb{M}_A](G(X))\right) \\
&= \sum_{B\in\mathcal{P}_D:B\notin\mathcal{P}_A} \mathrm{Cov}\left(G_A(X_A), G_B(X_B)\right).
\end{aligned}$$

However, notice that $\mathcal{U}_A \subset \mathcal{P}_D \setminus \mathcal{P}_A$, and that, for any $B \in \mathcal{P}_D \setminus \mathcal{P}_A$ with $B \notin \mathcal{U}_A$,

$$\mathrm{Cov}\left(G_A(X_A), G_B(X_B)\right) = 0,$$

and hence,

$$\begin{aligned}\sum_{B\in\mathcal{P}_A}(-1)^{|A|-|B|}\mathrm{Cov}\left(\mathbb{M}_B(G(X)),[I-\mathbb{M}_A]\left(G(X)\right)\right) &= \sum_{B\in\mathcal{P}_D:B\notin\mathcal{P}_A}\mathrm{Cov}\left(G_A(X_A),G_B(X_B)\right)\\ &= \sum_{B\in\mathcal{U}_A}\mathrm{Cov}\left(G_A(X_A),G_B(X_B)\right)\\ &= S_A^C.\end{aligned}$$

Proof of Proposition 4.12.
First, recall that

$$\mathbb{M}_A(G(X)) = \sum_{B\in\mathcal{P}_A} G_B(X_B).$$

Thus,

$$\begin{aligned}\mathbb{V}\left(\mathbb{M}_A(G(X))\right) &= \mathbb{V}\left(\sum_{B\in\mathcal{P}_A} G_B(X_B)\right)\\ &= \sum_{B\in\mathcal{P}_A}\mathbb{V}\left(G_B(X_B)\right) + \sum_{C\in\mathcal{U}_A}\mathrm{Cov}\left(G_B(X_B),G_C(X_C)\right)\\ &= \sum_{B\in\mathcal{P}_A} S_B^U + S_B^C\end{aligned}$$

which is equivalent to

$$\mathbb{V}\left(\mathbb{M}_A(G(X))\right) - \sum_{B\in\mathcal{P}_A} S_B^C = \sum_{B\in\mathcal{P}_A} S_B^U.$$

However, notice that, $\forall A\in\mathcal{P}_D$

$$\sum_{B\in\mathcal{P}_A} S_B^C = \mathrm{Cov}\left(\mathbb{M}_A(G(X)),[I-\mathbb{M}_A]\left(G(X)\right)\right),$$

and thus, $\forall A\in\mathcal{P}_D$

$$\mathbb{V}\left(\mathbb{M}_A(G(X))\right) - \mathrm{Cov}\left(\mathbb{M}_A(G(X)),[I-\mathbb{M}_A]\left(G(X)\right)\right) = \sum_{B\in\mathcal{P}_D} S_B^U.$$

Using Rota's generalization of the Möbius inversion formula applied to the power-set, it yields that $\forall A\in\mathcal{P}_D$

$$S_A^U = \sum_{B\in\mathcal{P}_A}(-1)^{|A|-|B|}\left[\mathbb{V}\left(\mathbb{M}_B(G(X))\right) - \mathrm{Cov}\left(\mathbb{M}_B(G(X)),[I-\mathbb{M}_A]\left(G(X)\right)\right)\right].$$

Proof of Lemma 4.5.
First, suppose that X is mutually independent. By Proposition 4.9, one has that

$$\forall A,B\in\mathcal{P}_D, B\neq A \quad V_A\perp V_B,$$

which entails that

$$V_A\perp W_A = \bigoplus_{B\in\mathcal{P}_D:B\neq A} V_B.$$

However, notice that W_A still complements V_A in $\mathbb{L}^2\left(\sigma_X\right)$. Furthermore, by unicity of the orthogonal complement, one has that $W_A = V_A^\perp$. Thus,

$$\mathrm{Ran}\left(Q_A\right) = V_A, \quad \mathrm{Ker}\left(Q_A\right) = V_A^\perp,$$

and thus $Q_A = P_A$, leading to

$$Q_A\left(G(X)\right) = P_A\left(G(X)\right) \text{ a.s.}$$

Now, suppose that for any $A \in \mathcal{P}_D$, $Q_A\left(G(X)\right) = P_A\left(G(X)\right)$ a.s. Hence, it implies that

$$\forall A \in \mathcal{P}_D, G_A(X_A) = P_A\left(G(X)\right).$$

which implies that $P_A = Q_A$, since the above equation defines the operator Q_A. Thus, P_A and Q_A must share the same ranges and nullspaces. In particular,

$$\mathrm{Ker}\left(Q_A\right) = \mathrm{Ker}\left(P_A\right) = V_A^{\perp},$$

implying that $W_A = V_A^{\perp}$, which leads to

$$V_A \perp W_A, \forall A \in \mathcal{P}_D \quad \Longleftrightarrow \quad V_A \perp V_B, \forall A, B \in \mathcal{P}_D, B \neq A.$$

Finally, thanks to Proposition 4.9, notice that this is equivalent to X being mutually independent.

D.4 Analytical results

This section is dedicated to presenting the analytical results for the toy-case in Section 4.5.

```
import sympy as sym
import numpy as np
sym.init_printing()
```

1 Framework

We have that $X = (X_1, X_2)$ with

$$X_1 \sim \mathcal{B}(q_1), \quad X_2 \sim \mathcal{B}(q_2).$$

Additionnally, let

$$\rho = \mathbb{P}(\{X_1 = 1\} \cap \{X_2 = 1\}).$$

```
## Variables
q1, q2= sym.symbols('q1,q2', positive=True)
rho=sym.Symbol("rho")
```

The joint law of X is given by

$$\begin{aligned}
p_{00} &= \mathbb{P}(\{X_1 = 0\} \cap \{X_2 = 0\}) = 1 - q_1 - q_2 + \rho \\
p_{01} &= \mathbb{P}(\{X_1 = 0\} \cap \{X_2 = 1\}) = q_2 - \rho \\
p_{10} &= \mathbb{P}(\{X_1 = 1\} \cap \{X_2 = 0\}) = q_1 - \rho \\
p_{11} &= \mathbb{P}(\{X_1 = 1\} \cap \{X_2 = 1\}) = \rho
\end{aligned}$$

```
p00=1-q1-q2+rho
p01=q2-rho
p10=q1-rho
p11=rho

probs=[p00, p01, p10, p11]
```

```
## Functions
def prosca(pr,a,b): #E[ab] = a^T P b
    tmp=[x*y*z for x,y,z in zip(pr, a, b)]
    return(sum(tmp))

def norm_sq(pr, x): #E[x^2] = x^T P x
    return(prosca(pr, x,x))

def orthProj(probs, vec, subspace):
    weight=prosca(probs, vec, subspace)/norm_sq(probs, subspace)
    return([x*weight for x in subspace])
```

2 Defining the subspaces and the canonical decomposition

First, notice that the functions of X can only take four different values. Hence, any $G : \{0,1\}^2 \to \mathbb{R}$ can be represented as a vector in $\mathbb{R}^4$.

2.1 $V_\emptyset$

Notice that $V_\emptyset$ is comprised of constant functions of X. In other words, for every $v_\emptyset \in V_\emptyset$, it can be represented as

$$v_\emptyset = \begin{pmatrix} c \\ c \\ c \\ c \end{pmatrix},$$

where $c \in \mathbb{R}$.

```
c=sym.Symbol("c")
ve=[c,c,c,c]
param_ve=[c]
```

We want to find a vector $v_\emptyset \in V_\emptyset$ such that its norm (w.r.t. the joint law of X) is equal to 1.

```
eq_ve=[norm_sq(probs, ve)-1] #Norm(ve) = 1
eq_ve
```

$$\left[c^2\rho + c^2\left(q_1 - \rho\right) + c^2\left(q_2 - \rho\right) + c^2\left(-q_1 - q_2 + \rho + 1\right) - 1\right]$$

2.2 V_1 and V_2

By definition, $V_1 \subset \mathbb{L}^2(\sigma_1)$, and thus, elements of V_1 are functions of X_1. It implies that the values they take does not change w.r.t. the values taken by X_2. Hence, a function of X in V_1 can be represented as

$$v_1 = \begin{pmatrix} g_0 \\ g_0 \\ g_1 \\ g_1 \end{pmatrix},$$

where $g_0, g_1 \in \mathbb{R}$.

```
g0,g1 = sym.symbols('g0,g1')
v1=[g0,g0,g1,g1]
param_v1=[g0,g1]
```

For G_1 to be in V_1, it must be orthogonal (w.r.t. the joint law to X) to the functions in $V_\emptyset$. Additionnally, we want to find a vector $v_1 \in V_1$ with unit norm.

```
eq1_v1=prosca(probs, v1, ve) #v1 \perp ve
eq2_v1=norm_sq(probs, v1)-1 #Norm(v1)= 1
eqs_v1= [eq1_v1, eq2_v1]
```

Analogously, since $V_2 \subset \mathbb{L}^2(\sigma_2)$, any element of $v_2 \in V_2$ admits the form:

$$v_2 = \begin{pmatrix} h_0 \\ h_1 \\ h_0 \\ h_1 \end{pmatrix},$$

where $h_0, h_1 \in \mathbb{R}$.

```
h0,h1 = sym.symbols('h0,h1')
v2=[h0,h1,h0,h1]
param_v2=[h0,h1]

eq1_v2=prosca(probs, v2, ve) #v2 \perp ve
eq2_v2=norm_sq(probs, v2)-1 #Norm(v2)= 1
```

```
eqs_v2= [eq1_v2, eq2_v2]
```

2.3 V_{12}

Since $V_{12} \subset \mathbb{L}^2(\sigma_{12})$, any element of $v_{12} \in V_{12}$ admits the form:

$$v_{12} = \begin{pmatrix} k_{00} \\ k_{01} \\ k_{10} \\ k_{11} \end{pmatrix},$$

where $k_{00}, k_{01}, k_{10}, k_{11} \in \mathbb{R}$.

```
k00, k01, k10, k11 = sym.symbols('k00, k01, k10, k11')
v12=[k00, k01, k10, k11]
param_v12=[k00, k01, k10, k11]
```

By definition of V_{12} it is the orthogonal complement in $\mathbb{L}^2(\sigma_{12}$ of

$$V_\emptyset + V_1 + V_2,$$

which is equivalent to V_{12} to be orthogonal to each of the summands.

```
eq1_v12=prosca(probs, v12, ve) ## v12 \perp ve
eq2_v12=prosca(probs, v12, v1) ## v12 \perp v1
eq3_v12=prosca(probs, v12, v2) ## v12 \perp v2
eq4_v12=norm_sq(probs, v12)-1 ## Norm(v12)=1

eqs_v12=[eq1_v12, eq2_v12, eq3_v12, eq4_v12]
```

2.4 Canonical decomposition

Any function $G(X)$ in $\mathbb{L}^2(\sigma_{12})$ can be represented as

$$G = \begin{pmatrix} G_{00} \\ G_{01} \\ G_{10} \\ G_{11} \end{pmatrix}$$

```
G00,G01,G10,G11 = sym.symbols('G00,G01,G10,G11')
G=[G00,G01,G10,G11]
```

And the canonical decomposition entails that

$$\begin{aligned} G &= e \times v_\emptyset + \alpha \times v_1 + \beta \times v_2 + \delta \times v_{12} \\ &= G_\emptyset + G_1 + G_2 + G_{12}. \end{aligned}$$

```
alpha,beta,delta,e = sym.symbols('alpha, beta, delta, e')
decomp_params = [alpha,beta,delta,e]

eqs_params=[e*x + alpha*y + beta*z + delta*t - g for x,y,z,t, g in zip(ve, v1, v2, v12, G)]
```

2.5 Solution to the problem

We have 13 parameters related to 13 different equations, which can be solved.

```
parameters= decomp_params + param_v12 + param_v2 + param_v1 + param_ve
equations=eq_ve + eqs_v1 + eqs_v2 + eqs_v12 + eqs_params

print("Number of parameters:")
print(len(parameters))

print("Number of equations:")
print(len(equations))
```

```
Number of parameters:
13
Number of equations:
13
```

```
%%time
res_params=sym.solve(equations,
                     parameters,
                     dict=True,
                    check=False)
```

```
Wall time: 44.9 s
```

```
e_=res_params[2][e]
ve=[res_params[2][c], res_params[2][c], res_params[2][c], res_params[2][c]]

alpha_=res_params[2][alpha]
v1 = [res_params[2][g0], res_params[2][g0], res_params[2][g1], res_params[2][g1]]

beta_=res_params[2][beta]
v2 = [res_params[2][h0], res_params[2][h1], res_params[2][h0], res_params[2][h1]]

delta_=res_params[2][delta]
v12 = [res_params[2][k00], res_params[2][k01], res_params[2][k10], res_params[2][k11]]
```

```
Ge=[sym.simplify(e_*x).factor(G00,G01,G10,G11) for x in ve]
G1=[sym.simplify(alpha_*x).factor(G00,G01,G10,G11) for x in v1]
G2=[sym.simplify(beta_*x).factor(G00,G01,G10,G11) for x in v2]
G12=[sym.simplify(delta_*x).factor(G00,G01,G10,G11) for x in v12]
```

2.6 Observations and verification

2.6.1 Evaluation decomposition

$$G(X) = eV_{\emptyset} + \alpha V_1 + \beta V_2 + \delta V_{12}$$

```
[sym.simplify(x + y + z + t) for x,y,z,t in zip(Ge, G1, G2, G12)]
```

$[G_{00}, G_{01}, G_{10}, G_{11}]$

2.6.2 e is equal to the expectation of $G(X)$:

$$e = \mathbb{E}[G(X)]$$

```
sym.simplify(prosca(probs, G, [1,1,1,1]) - e_)==0
```

```
True
```

2.6.3 Variance decomposition

$$\mathbb{E}[G^2(X)] = e^2 + \alpha^2 + \beta^2 + \delta^2 + 2\alpha\beta\mathbb{E}[V_1(X_1), V_2(X_2)]$$

```
normG= e_**2 + alpha_**2 + beta_**2 + delta_**2 + 2*alpha_*beta_*prosca(probs, v1,v2)
sym.simplify(norm_sq(probs, G) - normG) == 0
```

```
True
```

2.6.4 Annihilating property

V_{12}

$$P_1(Q_{12}(G(X))) = 0$$

```
[sym.simplify(x) for x in orthProj(probs, G12, v1)]
```

[0, 0, 0, 0]

$$P_2(Q_{12}(G(X))) = 0$$

```
[sym.simplify(x) for x in orthProj(probs, G12, v2)]
```

[0, 0, 0, 0]

$$P_\emptyset(Q_{12}(G(X))) = 0$$

```
[sym.simplify(x) for x in orthProj(probs, G12, ve)]
```

[0, 0, 0, 0]

V_1

$$P_\emptyset(Q_1(G(X))) = 0$$

```
[sym.simplify(x) for x in orthProj(probs, G1, ve)]
```

[0, 0, 0, 0]

V_2

$$P_\emptyset(Q_2(G(X))) = 0$$

```
[sym.simplify(x) for x in orthProj(probs, G2, ve)]
```

[0, 0, 0, 0]

2.6.5 Correlation of X and $c_0(V_1, V_2)$

$$c_0(V_1, V_2) = \mathbb{E}[v_1(X_1)v_2(X_2)] = \mathrm{Corr}(X_1, X_2)$$

```
corrX=(rho-q1*q2)/(sym.sqrt(q1*(1-q1))*sym.sqrt(q2*(1-q2)))
sym.simplify(prosca(probs, v1,v2) - corrX)
```

0

2.6.6 Bounds for definite-positiveness of Δ

Δ is definite-positive if

$$|\mathbb{E}[v_1(X_1)v_2(X_2)]| < 1$$

which restricts rho to takes the values:

```
bd_rho=sym.solve(sym.sqrt(prosca(probs, v1,v2)**2)-1, rho)
bd_rho
```

$$\left[-\sqrt{q_1}\sqrt{q_2}\sqrt{(q_1-1)\,(q_2-1)}+q_1q_2,\ \sqrt{q_1}\sqrt{q_2}\sqrt{(q_1-1)\,(q_2-1)}+q_1q_2\right]$$

3 Evaluation Decomposition

3.1 Canonical Evaluation Decomposition

$G(X) = eV_\emptyset + \alpha V_1 + \beta V_2 + \delta V_{12}$

```
[sym.simplify(x + y + z + t) for x,y,z,t in zip(Ge, G1, G2, G12)]
```

$[G_{00},\ G_{01},\ G_{10},\ G_{11}]$

3.2 Canonical Shapley

```
CSh1 = [sym.simplify(x+sym.Rational("1/2")*y) for x,y in zip(G1, G12)]
CSh2 = [sym.simplify(x+sym.Rational("1/2")*y) for x,y in zip(G2, G12)]
```

```
[sym.simplify(x+y+z) for x,y,z in zip(Ge, CSh1, CSh2)]
```

$[G_{00},\ G_{01},\ G_{10},\ G_{11}]$

4 Variance Decomposition

4.1 Organic indices

4.1.1 Pure interaction

```
%%time
Se=sym.simplify(norm_sq(probs,Ge).subs(rho, q1*q2))
S1=sym.simplify(norm_sq(probs,G1).subs(rho, q1*q2))
S2=sym.simplify(norm_sq(probs,G2).subs(rho, q1*q2))
S12=sym.simplify(norm_sq(probs,G12).subs(rho, q1*q2))
S=[Se, S1, S2, S12]
```

```
Wall time: 2.37 s
```

4.1.2 Dependence indices

```
%%time
Pe=orthProj(probs, G, ve)
P1=orthProj(probs, G, v1)
P2= orthProj(probs, G, v2)
P12= orthProj(probs, G, v12)

De=sym.simplify(norm_sq(probs, [sym.simplify(x-y) for x,y in zip(Ge, Pe)]))
D1=sym.simplify(norm_sq(probs, [sym.simplify(x-y) for x,y in zip(G1, P1)]))
D2=sym.simplify(norm_sq(probs, [sym.simplify(x-y) for x,y in zip(G2, P2)]))
```

```
D12=sym.simplify(norm_sq(probs, [sym.simplify(x-y) for x,y in zip(G12, P12)]))
D=[De, D1, D2, D12]
```

```
Wall time: 49.6 s
```

4.2 Canonical indices

4.2.1 Structural indices

```
%%time
Sue=sym.simplify(norm_sq(probs, Ge))
Su1=sym.simplify(norm_sq(probs, G1))
Su2=sym.simplify(norm_sq(probs, G2))
Su12=sym.simplify(norm_sq(probs, G12))
Su=[Sue, Su1, Su2, Su12]
```

```
Wall time: 11 s
```

4.2.2 Correlative indices

```
%%time
Sce=sym.zeros(1)[0]
Sc1=sym.simplify(prosca(probs, G1,G2))
Sc2=sym.simplify(prosca(probs, G2,G1))
Sc12=sym.zeros(1)[0]
Sc=[Sce, Sc1, Sc2, Sc12]
```

```
Wall time: 11.5 s
```

5 Case of $q_1 = q_2 = 0.5$

```
rplc=[(q1, sym.Rational("1/2")), (q2,sym.Rational("1/2"))]
```

5.1 Evaluation decomposition

$$e \times v_\emptyset$$

```
[sym.simplify((e_*x).subs(rplc)).factor(rho) for x in ve]
```

$$\left[\frac{G_{01}+G_{10}+\rho\left(2G_{00}-2G_{01}-2G_{10}+2G_{11}\right)}{2},\right.$$
$$\frac{G_{01}+G_{10}+\rho\left(2G_{00}-2G_{01}-2G_{10}+2G_{11}\right)}{2},$$
$$\frac{G_{01}+G_{10}+\rho\left(2G_{00}-2G_{01}-2G_{10}+2G_{11}\right)}{2},$$
$$\left.\frac{G_{01}+G_{10}+\rho\left(2G_{00}-2G_{01}-2G_{10}+2G_{11}\right)}{2}\right]$$

$$\alpha \times v_1$$

```
[sym.simplify((alpha_*x).subs(rplc)).factor() for x in v1]
```

$$\left[\frac{G_{00}+G_{01}-G_{10}-G_{11}}{4}, \frac{G_{00}+G_{01}-G_{10}-G_{11}}{4}, -\frac{G_{00}+G_{01}-G_{10}-G_{11}}{4}, -\frac{G_{00}+G_{01}-G_{10}-G_{11}}{4}\right]$$

$$\beta \times v_2$$

```
[sym.simplify((beta_*x).subs(rplc)) for x in v2]
```

$$\left[\frac{G_{00}}{4} - \frac{G_{01}}{4} + \frac{G_{10}}{4} - \frac{G_{11}}{4}, -\frac{G_{00}}{4} + \frac{G_{01}}{4} - \frac{G_{10}}{4} + \frac{G_{11}}{4}, \frac{G_{00}}{4} - \frac{G_{01}}{4} + \frac{G_{10}}{4} - \frac{G_{11}}{4}, -\frac{G_{00}}{4} + \frac{G_{01}}{4} - \frac{G_{10}}{4} + \frac{G_{11}}{4}\right]$$

$$\delta \times v_2$$

```
[sym.simplify((delta_*x).subs(rplc)) for x in v12]
```

$$\left[\frac{\left(-4\rho^2 + 4\rho - 1\right)\left(G_{00} - G_{01} - G_{10} + G_{11}\right)}{2 \cdot (2\rho - 1)}, \rho\left(-G_{00} + G_{01} + G_{10} - G_{11}\right), \rho\left(-G_{00} + G_{01} + G_{10} - G_{11}\right), \frac{\rho\sqrt{\frac{1-2\rho}{\rho}}\sqrt{\frac{-4\rho^2+4\rho-1}{\rho(2\rho-1)}}\left(G_{00} - G_{01} - G_{10} + G_{11}\right)}{2}\right]$$

5.2 Variance Decomposition

5.2.1 Organic indices

Pure interaction

```
[sym.simplify(x.subs(rplc)) for x in S]
```

$$\left[\frac{\left(G_{00} + G_{01} + G_{10} + G_{11}\right)^2}{16}, \frac{\left(G_{00} + G_{01} - G_{10} - G_{11}\right)^2}{16}, \frac{\left(G_{00} - G_{01} + G_{10} - G_{11}\right)^2}{16}, \frac{\left(G_{00} - G_{01} - G_{10} + G_{11}\right)^2}{16}\right]$$

Dependence indices

```
[sym.simplify(x.subs(rplc)).factor(rho) for x in D]
```

$$\left[0, \frac{(4\rho - 1)^2\left(-G_{00} + G_{01} - G_{10} + G_{11}\right)^2}{16}, \frac{(4\rho - 1)^2\left(-G_{00} - G_{01} + G_{10} + G_{11}\right)^2}{16}, 0\right]$$

5.2.2 Canonical indices

Structural indices

```
[sym.simplify(x.subs(rplc)).factor(rho) for x in Su]
```

$$\left[\frac{\left(G_{01}+G_{10}+\rho\left(2G_{00}-2G_{01}-2G_{10}+2G_{11}\right)\right)^2}{4},\right.$$
$$\frac{\left(G_{00}+G_{01}-G_{10}-G_{11}\right)^2}{16},$$
$$\frac{\left(G_{00}-G_{01}+G_{10}-G_{11}\right)^2}{16},$$
$$\left.-\frac{\rho\left(2\rho-1\right)\left(G_{00}-G_{01}-G_{10}+G_{11}\right)^2}{2}\right]$$

Correlative indices

```
[sym.simplify(x.subs(rplc)).factor(rho) for x in Sc]
```

$$\left[0,\right.$$
$$\left(4\rho-1\right)\left(\frac{G_{00}}{16}+\frac{G_{01}}{16}-\frac{G_{10}}{16}-\frac{G_{11}}{16}\right)\left(G_{00}-G_{01}+G_{10}-G_{11}\right),$$
$$\left(4\rho-1\right)\left(\frac{G_{00}}{16}+\frac{G_{01}}{16}-\frac{G_{10}}{16}-\frac{G_{11}}{16}\right)\left(G_{00}-G_{01}+G_{10}-G_{11}\right),$$
$$\left.0\right]$$

6 Further numerical testing

```
# Some Boolean function valued in R
def model(x1,x2):
    return(x1+x2-2*np.exp(x1*x2) + 5*np.log(np.pi)**(x1*x2))

#Marginal probabilities
p1,p2=np.random.random_sample(size=2)

# Lower and upper bounds on the correlation between X1 and X2
L = max(0, p1+p2-1)
U = min(p1,p2)

#Random sample of the correlation
p12=(L-U)*np.random.random_sample(1)[0] + U



#Sample the values of G
g00 = model(0,0)
g01 = model(0,1)
g10 = model(1,0)
g11 = model(1,1)

Gb = [g00, g01, g10, g11]


#Vector of replacements for the sympy formulas
rplc=[(q1, p1), (q2, p2), (rho, p12), (G00, g00), (G01, g01), (G10, g10), (G11, g11)]
```

```
#Variance of G under dependence
VG=(norm_sq(probs, Gb)).subs(rplc)

#Variance of G under mutual independence
VG_=(norm_sq(probs, Gb)).subs(rho, q1*q2).subs(rplc)



print('Var G')
print(VG)
print("\n")

print('Var G indep')
print(VG_)
print("\n")

print('Probas : q1, q2, rho')
print([p1, p2, p12])
print("\n")

print('Correlation X1 and X2')
print((p12-p1*p2)/(np.sqrt(p1*(1-p1))*np.sqrt(p2*(1-p2))))
print("\n")

print('c(L²(1),L²(2))')
print(prosca(probs, v1,v2).subs(rplc))
print("\n")

print('rho LU bounds')
print([L,U])
print("\n")

print("rho DP bounds")
dpb=[max(0,bd_rho[0].subs([(q1,p1),(q2,p2)])), min(1, bd_rho[1].subs([(q1,p1),(q2,p2)]))]
print(dpb)
print("\n")

print('Values : G00, G01, G10, G11')
print([g00, g01, g10, g11])
print("\n")

print('##########################################################################')
print('Evaluation decomposition')
print('##########################################################################')
print("Ge")
print([sym.simplify(x.subs(rplc)) for x in Ge])
print("G1")
print([sym.simplify(x.subs(rplc)) for x in G1])
print("G2")
print([sym.simplify(x.subs(rplc)) for x in G2])
print("G12")
print([sym.simplify(x.subs(rplc)) for x in G12])
print("Ge + G1 + G2 + G12")
print([sym.simplify((x+y+z+t).subs(rplc)) for x,y,z,t in zip(Ge, G1, G2, G12)])
```

```
print("\n")

print('##############################################################################')
print('Organic decomposition')
print('##############################################################################')

print('Pure interaction')
print([sym.simplify(x.subs(rplc)/VG_) for x in S])
print(sum([sym.simplify(x.subs(rplc)/VG_) for x in S]))
print("\n")

print('Dependence indices')
print([sym.simplify(x.subs(rplc)/VG) for x in D])
print(sum([sym.simplify(x.subs(rplc)/VG) for x in D]))
print("\n")


print('##############################################################################')
print('Canonical Decomposition')
print('##############################################################################')
print('Structural indices')
print([sym.simplify(x.subs(rplc)/VG) for x in Su])
print(sum([sym.simplify(x.subs(rplc)/VG) for x in Su]))
print("\n")


print('Correlative indices')
print([sym.simplify(x.subs(rplc)/VG) for x in Sc])
print(sum([sym.simplify(x.subs(rplc)/VG) for x in Sc]))
print("\n")
```

```
Var G
13.6806203605639


Var G indep
13.4088889317681


Probas : q1, q2, rho
[0.028850246229265686, 0.6484828481519346, 0.003416651733481213]


Correlation X1 and X2
-0.19135097530836204


c(L²(1),L²(2))
-0.191350975308362


rho LU bounds
[0, 0.028850246229265686]
```

```
rho DP bounds
[0, 0.0986261104729329]


Values : G00, G01, G10, G11
[3.0, 4.0, 4.0, 2.2870857723289104]


##########################################################################
Evaluation decomposition
##########################################################################

Ge
[3.66806401128244, 3.66806401128244, 3.66806401128244, 3.66806401128244]
G1
[-0.0189952475655124, -0.0189952475655124, 0.639413260097192, 0.639413260097192]
G2
[-0.624353328200793, 0.338437484200944, -0.624353328200793, 0.338437484200944]
G12
[-0.0247154355161360, 0.0124937520821266, 0.316876056821160, -2.35882898325167]
Ge + G1 + G2 + G12
[3.00000000000000, 4.00000000000000, 4.00000000000000, 2.28708577232891]


##########################################################################
Organic decomposition
##########################################################################

Pure interaction
[0.980846682537967, 0.00120460583555726, 0.0144431300645849,
0.00350558156189049]
1.00000000000000


Dependence indices
[0, 0.000565541472312945, 3.25073947774616e-5, 0]
0.000598048867090407


##########################################################################
Canonical Decomposition
##########################################################################

Structural indices
[0.983485634149333, 0.000887811579599808, 0.0154455400471367,
0.00159818643345630]
1.00141717220953


Correlative indices
[0, -0.000708586104763110, -0.000708586104763110, 0]
-0.00141717220952622
```

Appendix E

Supplementary material for Chapter 5

Contents

D.1 Some technical preliminaries **128**
D.1.1 Hilbert space external direct sums 128
D.1.2 Closed range operator 128
D.2 Proof of Theorem 4.6 . . . . . **128**
D.2.1 Intermediary results 128
D.2.2 Proof of the main result 130
D.3 Proofs **132**
D.4 Analytical results **135**

E.1 Some preliminaries

E.1.1 Application domain

The application domain of inputs X is a subset of $\Omega_X \subseteq \mathbb{R}^d$, representing the *region of the inputs where the black-box model is intended to be used* [189]. Figure E.1 illustrates a typical situation for a univariate marginal of X.

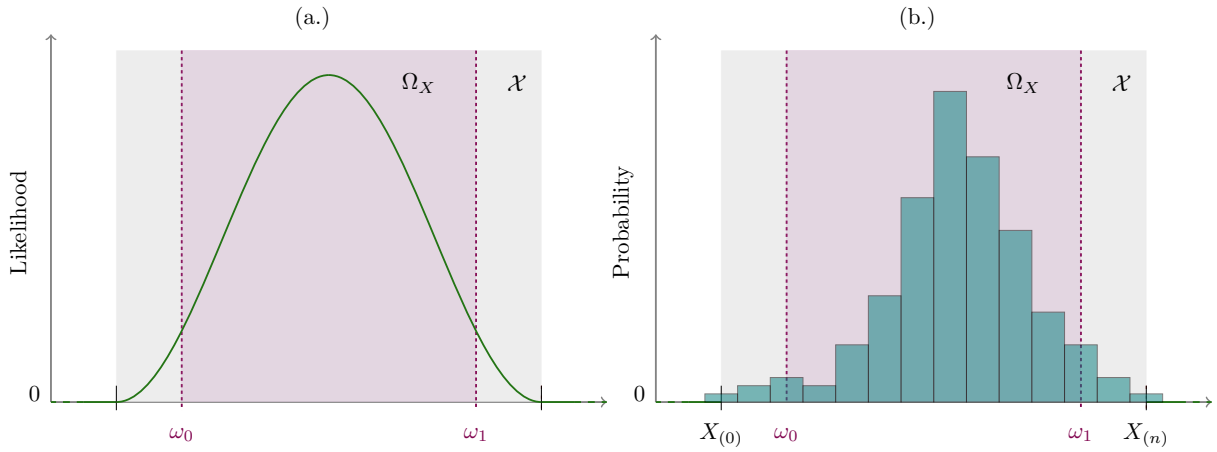

Figure E.1: Application domain Ω_X of P when P admits a density (a.) and when it is empirical (b.). In (a.), $\mathcal{X}$ is the support of the density (in grey), and the application domain Ω_X (in purple) is contained in $\mathcal{X}$. In (b.), $\mathcal{X}$ is the interval between the minimum and maximum observed values (in grey), and the application domain Ω_X (in purple) is also contained in $\mathcal{X}$. In both cases, Ω_X is chosen to be strictly included in $\mathcal{X}$, although it can be larger.

Remark E.1. In practice, the application domains of marginal distribution can be defined in many ways. For instance, if P is empirical, it can represent the range between the smallest and largest observed value of X_i in a specific dataset. If P is part of a parametric family, it can be defined using experts' opinions, usually enforced using truncation. These domains are usually subject to uncertainties in their bounds.

For instance, given a set $x_1, \ldots, x_n$ of training, validating or testing examples, its convex hull, or a broader span, are common choices of application domains Ω_X [224]. In other instances, Ω_X can be seen as the extrapolation domain where G is assumed to generalize well (e.g., paving of a compact subspace of $\mathbb{R}^d$ selected by tree-based classification [103], confidence measures or cross-validation schemes [106, 165, 220, 131]). In ML specifically, including out-of-distribution data in Ω_X remains an open problem [98, 220, 202].

E.1.2 Cumulative distribution functions and quantile functions

Recall the following classical definitions.

Definition E.1 (Cumulative distribution function). Let X be a random variable with induced probability measure P. The cdf of P is denoted and defined as

$$F_P(t) = \int_{(-\infty,t]} dP = P\big((-\infty, t]\big).$$

Definition E.2 (Space of cumulative distribution functions). The space of cdfs is the space denoted and defined as:

$$\mathcal{F} = \Big\{ F : \mathbb{R} \to [0,1] \mid F \text{ is right-continuous, non-decreasing} \\ \text{such that } \lim_{x\to\infty} F(x) = 1 \text{ and } \lim_{x\to-\infty} F(x) = 0 \Big\}. \tag{E.1}$$

The use of generalized quantile functions (gqf) is motivated by the fact that the marginal distributions P_i can be atomic. They are defined as one of the two generalized inverses of cdfs in $\mathcal{F}$. For a univariate probability measure P, one can define a left and right continuous generalized inverse, the former being usually called its gqf. They can be formally defined as follows [185, 63, 126].

Definition E.3 (Generalized quantile function). Let $P \in \mathcal{P}(\mathbb{R})$ with cdf F_P.

(i) The gqf of P is the unique left-continuous, non-decreasing generalized inverse of F_P, defined, for every $a \in (0,1)$, as:

$$\begin{aligned} F_P^{\leftarrow}(a) &= \sup\{t \in \mathbb{R} \mid F_P(t) < a\}, \\ &= \inf\{t \in \mathbb{R} \mid F_P(t) \geq a\}. \end{aligned} \tag{E.2}$$

(ii) The unique right-continuous non-decreasing generalized inverse $F_P^{\rightarrow}$ of F_P, almost-everywhere equal to $F_P^{\leftarrow}$, is defined, for every $a \in (0,1)$, as:

$$\begin{aligned} F_P^{\rightarrow}(a) &= \sup\{t \in \mathbb{R} \mid F_P(t) \leq a\}, \\ &= \inf\{t \in \mathbb{R} \mid F_P(t) > a\}, \\ &= F_P^{\leftarrow}\left(a^+\right) \end{aligned} \tag{E.3}$$

where $F_P^{\leftarrow}(a^+) = \lim_{x\to a^+} F_P^{\leftarrow}(x)$.

If the cdf F_P of a random variable X admits an inverse F_P^{-1} in the traditional sense (e.g., it is continuous, strictly increasing), then the following equality holds:

$$F_P^{-1} = F_P^{\leftarrow} = F_P^{\rightarrow}.$$

Furthermore, univariate probability measures are intrinsically linked to their gqf. Denote:

$$\mathcal{F}^{\leftarrow} = \Big\{ F^{\leftarrow} : (0,1) \to \mathbb{R} \mid F^{\leftarrow} \text{ is left-continuous and non-decreasing} \Big\}. \tag{E.4}$$

the space of gqfs. Recall that each probability measure in $\mathcal{P}(\mathbb{R})$ has a unique gqf in $\mathcal{F}^{\leftarrow}$ [185].
For a fixed $\alpha \in [0,1]$, an α-quantile of P is a number $p_\alpha \in \mathbb{R}$ such that, for $X \sim P$:

$$P(\{X < p_\alpha\}) \leq \alpha \quad \text{and} \quad P(\{X \leq p_\alpha\}) \geq \alpha.$$

In certain cases, α-quantiles are not unique. For instance, if P is purely atomic (e.g., an empirical measure), and its cdf F_P takes the constant value α on an open interval (t_0, t_1) (i.e., it is the case if t_0 and t_1 are both atoms of an empirical probability measure), then any $t \in (t_0, t_1)$ is an α-quantile. One can notice that $F^{\leftarrow}(\alpha)$ is the *infimum* of the α-quantiles of P, (i.e., $F_P^{\leftarrow}(\alpha) = t_0$), and $F^{\rightarrow}(\alpha)$ is the *supremum* of the α-quantiles of P (i.e., $F_P^{\rightarrow}(\alpha) = t_1$).

E.1.3 Wasserstein distance

Let p be a positive integer. The p-Wasserstein distance between two univariate marginals can be defined as follows [219]:

Definition E.4 (Wasserstein distance on the real line). Let $p \in \mathbb{N}^*$ and $P, Q \in \mathcal{P}_p(\mathbb{R})$ be two probability measures on $\mathbb{R}$ admitting F_P and F_Q as probability distribution functions, respectively. Then, the p-

Wasserstein distance between P and Q is:

$$W_p(P,Q) = \left(\int_0^1 \left| F_P^{\rightarrow}(x) - F_Q^{\rightarrow}(x) \right|^p dx \right)^{1/p}$$

In particular, for $p = 2$,

$$W_2(P,Q) = \sqrt{\int_0^1 \left(F_P^{\rightarrow}(x) - F_Q^{\rightarrow}(x) \right)^2 dx}.$$

E.2 Dependence modelling and copulas

E.2.1 Copulas

Dependencies between random variables can be modeled using copula-based representations [164]. Let $X = (X_1, \ldots, X_d) \sim P$ be a d-dimensional $\mathbb{R}^d$-valued random vector with marginal cdfs $F_{P_i}, i = 1, \ldots, d$, assumed to be continuous. Let $U_1, \ldots, U_d$ the random variables defined as:

$$U_i = F_{P_i}(X_i)$$

and denote $U = (U_1, \ldots, U_d)^\top \sim U_P$. For any $\mathbf{u} = (u_1, \ldots, u_d) \in [0,1]^d$, denote $H_{\mathbf{u}} = \bigtimes_{i=1}^d [0, u_i]$. The copula of X is the mapping from $[0,1]^d$ to $[0,1]$, denoted C_p defined as:

$$\begin{aligned} C_P(u) &= \mathbb{P}\left(U_1 \leq u_1, \ldots, U_d \leq u_d\right) \\ &= \int_{H_{\mathbf{u}}} dU_P \end{aligned}$$

If P is observed (and hence each F_{P_i} can jump), the notion of *empirical copula* characterizes the dependence structure between the inputs [164]. For $j \in \{1, \ldots, d\}$, denote $\{x_{j,i}\}_{1 \leq i \leq n}$ the jth marginal sample of observations. The empirical copula of X is defined as:

$$\hat{C}_P(u_1, \ldots, u_d) = \frac{1}{n} \sum_{i=1}^n \prod_{j=1}^d \mathbb{1}_{\left\{ \frac{R_{j,i}}{n} \leq u_j \right\}}(u_j), \tag{E.5}$$

where $R_{j,k}$ denotes the rank of $x_{j,k}$ in $\{x_{j,i}\}_{1 \leq i \leq n}$.

E.3 Computational details and code snippets

E.3.1 Moment matrix of the Lebesgue measure

The following `R` code snippet defines a function for computing the moment matrix of the Lebesgue measure on any interval $[a, b]$.

```
################################################
# Lebesgue Moment Matrix on an interval [a,b]
mom_mat<-function(a,b,d){
# a,b : Upper and lower bound of the interval
# d : Degree up to which the moment matrix is computed
    M<-matrix(NA, nrow=d+1, ncol=d+1)
    for(i in 1:(d+1)){
        for(j in 1:(d+1)){
            M[i,j] = (b^{i+j-1} - a^{i+j-1})/(i+j-1)
        }
    }
    return(M)
}
```

Listing E.1: Moment matrix of the Lebesgue measure on an interval.

E.3.2 Computing moment vector of arbitrary quantile functions

One wishes here at computing the vector described in Eq. (5.19). In the case where P is an empirical measure built on a n-sample, one has that for $[t_0, t_1] \in [0,1]$, $i = 0, \dots, p$:

$$\begin{aligned} r_i = & \frac{1}{i+1}\left[\sum_{j\in J} \frac{X_{(j)}}{n^{i+1}}\left((j+1)^{i+1} - j^{i+1}\right)\right. \\ & + X_{(\bar{j})}\left(t_1^{i+1} - \left(\frac{\bar{j}}{n}\right)^{i+1}\right) \\ & \left. + X_{(\underline{j}-1)}\left(\left(\frac{\underline{j}}{n}\right)^{i+1} - t_0^{i+1}\right)\right] \end{aligned}$$

where $J = \{i \in \mathbb{N} \mid \lfloor nt_0 \rfloor < i < \lfloor nt_1 \rfloor\}$, $\bar{j} = \lfloor t_1 n \rfloor$, $\underline{j} = \lfloor t_0 n \rfloor + 1$, and where $X_{(j)}$ denotes the j-th order statistic of the observe sample. In cases where $F_P^{\leftarrow}$ is continuous, it is possible to use numerical quadrature methods in order to evaluate each integral composing the elements r_i of r.

```
####################################################
# Moment vector of an empirical quantile function
mom_vec<-function(a,b,d){
# a,b : Upper and lower bound of the interval
# d : Degree up to which the moment vector is computed
# X : Dataset.

    # Setting-up the resulting vector
    r=rep(0, d+1)

    # Compute the weights
    weights_r<-function(j_down, J, j_up, i, n, a,b){
        w_J=sapply(J, function(x) (x+1)**(i) - x**i )
        w_J = w_J/(n**(i))
        w_jup<-(b**i - (j_up/n)**(i))
        w_jdown<-((j_down/n)**i - a**i)
        res=c(w_jdown, w_J, w_jup)/i
        return(res)
    }

    # Setting-up parameters
    n=length(X)
    X=sort(X)
    J=seq(floor(n*a)+1, floor(n*b)-1, 1)
    if(a==0){
        j_up=floor(b*n)
        j_down=1
    }else if (b==1){
        j_down=floor(a*n)+1
        j_up=n
    }else{
        j_up=floor(b*n)
        j_down=floor(a*n)+1
    }

    # Vector of relevant order statistics
    X_=X[c(j_down, J, j_up)]

    # Computing each element of r
    for(i in 1:(d+1)){
        wght_vec<-weights_r(j_down, J, j_up, i, n, a, b)
        r[i]=sum(X*wght_vec)
    }
    return(r)
```

```
}
```

Listing E.2: Moment vector of a empirical quantile function.

```
################################################
# Moment vector approximation of any function
mom_vec<-function(a,b,d,P){
# a,b : Upper and lower bound of the interval
# d : Degree up to which the moment vector is computed
# P : Function.

    # Setting-up the resulting vector
    r=rep(0, d+1)

    # Approximate using quadrature
    for(i in 1:(d+1)){
        f<-function(x,j){
            return((x^(j-1))*as.numeric(P(x)))
        }
        res<-integrate(f,
                       j=i,
                       lower=a,
                       upper=b)
        r[i]=res$value
    }
    return(r)
}
```

Listing E.3: Moment vector approximation of a function.

E.4 Proofs

Proof of Lemma 5.1.
Notice that if Eq. (5.3) is respected, then the constraints are non-decreasing. Then, there exists at least a function $F^{\leftarrow}$ in $\mathcal{F}^{\leftarrow}$ such that the constraints are respected (e.g., the linear interpolant of the constraints). So, there exists a probability measure with $F^{\leftarrow}$ as a generalized quantile function.

Proof of Lemma 5.2.
Since $[\eta_0, \eta_1]$ is bounded, one can define a standardized intensity parameter $\theta \in \Theta = [-1, 1]$ as:

$$\theta(b) = \frac{p_\alpha - b}{p_\alpha - \eta_1} \mathbb{1}_{\{b > p_\alpha\}}(b) + \frac{b - p_\alpha}{p_\alpha - \eta_0} \mathbb{1}_{\{b < p_\alpha\}}(b).$$

Equivalently, one can express b in terms of θ, which directly provides the expression of $b_\alpha(\boldsymbol{\eta}, \theta)$.

Proof of Lemma 5.3.
Preserving the midpoint of Ω_X while perturbing its width requires that, for any pair $(b_0, b_1) \in \mathbb{R}^2$, that

$$\begin{cases} \dfrac{b_0 + b_1}{2} = \dfrac{\omega_0 + \omega_1}{2} \\ b_1 - b_0 = \kappa(\omega_1 - \omega_0) \end{cases} \iff \begin{cases} b_1 = \dfrac{\omega_1(\kappa+1) - \omega_0(\kappa-1)}{2} \\ b_0 = \dfrac{\omega_0(\kappa+1) - \omega_1(\kappa-1)}{2} \end{cases}$$

where $\kappa \in [\frac{1}{\eta}, \eta]$. Using the transformation

$$\theta(\kappa) = \begin{cases} -\dfrac{\kappa - 1}{\frac{1}{\eta} - 1} & \text{if} \quad \dfrac{1}{\eta} \leq \kappa < 1 \\ 0 & \text{if} \quad \kappa = 1 \\ \dfrac{\kappa - 1}{\eta - 1} & \text{if} \quad 1 < \kappa < \eta \end{cases}$$

allows defining the formulas for b_0 and b_1 provided in the result's statement.

Proof of Lemma 5.4.
(i) Suppose that P is empirical. Notice that the empirical copula (see, Section E.2.1) only depends on the *ranks* of the observed data points. Since each $F_i^{\leftarrow}$ is strictly monotone increasing, the ranks between the initial and perturbed data points are preserved. Hence, the empirical copula between X and $\widetilde{X}$ is the same.
(ii) Let $F \in \mathcal{F}$, and recall that if $F^{\leftarrow}$ is strictly increasing then from [63], for all $u \in [0,1]$

$$(F \circ F^{\leftarrow})(u) = u$$

Now let $F_1, \dots, F_d \in \mathcal{F}$, such that $F_i^{\leftarrow}$ is strictly increasing, and denote:

$$\begin{array}{rcl} F : \mathbb{R} & \to & [0,1]^d \\ (u_1, \dots, u_d)^\top & \mapsto & (F_1(u_1), \dots, F_d(u_d))^\top \end{array}$$

One then has that:

$$F(T(X)) = F_P(X) \text{ a.s.}$$

and hence, X and $T(X)$ have the same copula.

Proof of Lemma 5.5.
Notice that, from Lemma 5.4, every probability measure in $\widetilde{Z}(P,\theta)$ has the same copula as P. Leveraging the work in [4] (Proposition 1.1), if P and Q share the same copula, one can rewrite their 2-Wasserstein distance as:

$$W_2^2(P,Q) = \sum_{i=1}^{d} W_2^2(P_i, Q_i) = \sum_{i=1}^{d} \int_0^1 \left(F_{P_i}^{\rightarrow}(x) - F_{Q_i}^{\rightarrow}(x)\right)^2 dx \tag{E.6}$$

Moreover, noticing that each marginal perturbation class $\mathcal{Q}_i(\theta)$ can be written as constraints on the generalized inverses of the cdf of Q_i. Hence, minimizing (E.6) entails minimizing each univariate transportation problem under marginal constraints. Finally, the perturbation map T is thus optimal between P and Q.

Proof of Proposition 5.1.
First, note that the intervals $A_i, i = 1, \dots, K$ are disjoint. Moreover for any $i = 1, \dots, K-1$, consider the four cases:

1. If $\alpha_i < \beta_i < \alpha_{i+1}$ and, then $A_i = (\alpha_i, \beta_i]$;
2. If $\beta_i < \alpha_i < \beta_{i+1}$ and, then $A_i = (\beta_i, \alpha_i]$;
3. If $\alpha_i < \beta_i$ and assume that $\alpha_{i+j} < \beta_{i+j-1}$ for $j = 1, \dots, m$ where $m \leq K - i$ is some non-negative integer, then $A_i = (\alpha_i, \alpha_{i+1}]$, additionally for $j = i+1, \dots, i+m-1$, $A_j = (\alpha_j, \alpha_{j+1}]$ and finally $A_{i+m} = (\alpha_{i+m}, \beta_{i+m}]$;
4. If $\beta_i < \alpha_i$ and assume that $\alpha_{i+j} < \beta_{i+j+1}$ for $j = 1, \dots, m$ where $m \leq K-i-1$ is some non-negative integer, then $A_i = (\beta_i, \alpha_i]$ and for $j = i+1, \dots, i+m$, $A_j = (\alpha_{j-1}, \alpha_j]$.

The integral can be decomposed as follows:

$$\int_0^1 \left(L(x) - F_P^{\rightarrow}(x)\right)^2 dx = \int_{\overline{A}} \left(L(x) - F_P^{\rightarrow}(x)\right)^2 dx + \sum_{i=1}^{K} \int_{A_i} \left(L(x) - F_P^{\rightarrow}(x)\right)^2 dx$$

where

$$\int_{\overline{A}} \left(L(x) - F_P^{\rightarrow}(x)\right)^2 dx \geq 0.$$

Since the quantile constraints are of the form:

$$L(\alpha_i) \leq b_i \leq L\left(\alpha_i^+\right).$$

one can always write $L(y) = b_i + h(y)$ for $y \in A_i$, and where h is an non-decreasing, left-continuous function. Moreover, note that:

- $h(y)$ is non-negative, and $F_P^{\rightarrow}(y) - b_i \leq 0$ if A_i falls in cases 2. and 4.
- $h(y)$ is non-positive, and $F_P^{\rightarrow}(y) - b_i \geq 0$ if A_i falls in cases 1. and 3.

Then one has:

$$\begin{aligned}\int_{A_i} \left(L(x) - F_P^{\rightarrow}(x)\right)^2 dx &= \int_{A_i} \left(L(x) - b_i - h(y)\right)^2 dx \\ &= \int_{A_i} \left(F_P^{\rightarrow}(x) - b_i\right)^2 dx + \int_{A_i} h(x)^2 dx \\ &\quad - 2\int_{A_i} h(x)\left(F_P^{\rightarrow}(x) - b_i\right) dx \\ &\geq \int_{A_i} \left(F_P^{\rightarrow}(x) - b_i\right)^2 dx\end{aligned}$$

since $h(x)$ and $F_P^{\rightarrow}(x) - b_i$ have different sign. Due to the constraint and the left-continuous non-decreasing nature of L, this bound is tight and is attained if and only if $h(y) = 0$ for all $y \in A_i$. Globally, this entails that

$$\int_0^1 \left(L(x) - F_P^{\rightarrow}(x)\right)^2 dx \geq \sum_{i=1}^{K} \int_{A_i} \left(F_P^{\rightarrow}(x) - b_i\right)^2 dx$$

and this tight bound is uniquely attained by the left-continuous non-decreasing function defined as

$$F_Q^{\leftarrow}(y) = \begin{cases} F_P^{\rightarrow}(y) & \text{if } y \in \overline{A} \\ b_i & \text{if } y \in A_i, \quad i = 1, \ldots, K. \end{cases}$$

E.5 Proof of Theorem 5.1

E.5.1 Ingredients

The proof of this theorem relies on the following results from [171, 172, 135], and further recalled in [205]. They involve sum-of-squares (SOS) polynomials, which can be defined as follows.

Definition E.5 (SOS polynomials). A polynomial S of even degree p is said to be a SOS polynomial if, for $m \in \mathbb{N}^*$, there exists $s_1, \ldots, s_m$ polynomials of degree at most equal to $\frac{d}{2}$, and such that, $\forall x \in \mathbb{R}$:

$$S(x) = \sum_{i=1}^{m} \left(s_i(x)\right)^2.$$

Theorem E.1. *Let $t_0, t_1 \in \mathbb{R}$ such that $t_0 < t_1$, and let $p \in \mathbb{N}^*$.*

(i) A univariate polynomial S of even degree $d = 2p$ is non-negative on $[t_0, t_1]$ if and only if it can be written as, $\forall x \in [t_0, t_1]$

$$S(x) = Z(x) + (x - t_0)(t_1 - x)W(x)$$

where Z is a SOS polynomial of degree at most equal to d, and W is an SOS polynomial of degree at most equal to $d - 2$.

(ii) An univariate polynomial S of odd degree $d = 2p + 1$ is non-negative on $[t_0, t_1]$ if and only if it can be written as, $\forall x \in [t_0, t_1]$

$$S(x) = (x - t_0)Z(x) + (t_1 - x)W(x)$$

where Z, W are SOS polynomials of degree at most equal to d.

It is important to note that Theorem E.1 is quite general in the sense that it allows for extensions to multivariate polynomials (i.e., polynomials taking values from $\mathbb{R}^d$). As pointed out in [58] (Thm. 1.4.2), nonnegative polynomials on compact intervals can also be defined as a linear combination of squared

polynomials. It may facilitate the identification of the nonnegative polynomials' coefficients, as done in [161] in the context of statistical learning. However, for the sake of potential future genericity, the direct powerful link between SOS polynomials and semi-definite positive matrices is leveraged, as expressed in the following theorem.

Theorem E.2. *Let S be a univariate polynomial of even degree $d = 2p$, with coefficients $s = (s_0, \dots, s_d)$, and denote x_p the usual monomial basis of polynomials of degree at most equal to p, i.e.,*

$$x_p = (1, x, x^2, \dots, x^{p-1}, x^p)^\top.$$

S is an SOS polynomial if and only if there exists a $(p \times p)$ symmetric semi-definite positive (SDP) matrix

$$\Gamma = \left[\Gamma_{ij}\right]_{i,j=1,\dots,p}$$

that satisfies, $\forall x \in \mathbb{R}$,

$$S(x) = x_p^\top \Gamma x_p.$$

Moreover, for $k = 0, \dots, d$, let $\mathbb{I}_k^p$ be the $(p \times p)$ matrix defined by, for $i, j = 1, \dots, p$:

$$\left[\mathbb{I}_k^p\right]_{i,j} = \mathbb{1}_{\{i+j=k+2\}}(i, j).$$

Then one additionally has that, for $i = 0, \dots, d$

$$s_i = \langle \mathbb{I}_i^p, \Gamma \rangle_F = \sum_{j+k=i+2} \Gamma_{j,k} \tag{E.7}$$

where, $\langle ., . \rangle_F$ denotes the Frobenius norm on matrices.

Theorem E.3. *Let $\mathbb{S}_n$ the subspace of real-valued symmetric matrices, in the vector space of square matrices. The set of symmetric SDP matrices Σ_N is a proper cone in $\mathbb{S}_n$, and thus is a closed convex set.*

A few results on the preservation of convexity of sets under transformations are also required. These lemmas can be found in [25].

Lemma E.1 (Linear maps preserve convexity). *Let V, W be two vector spaces over the same field F. Let $T : V \to W$ be a linear map, and let $\mathcal{C} \subset V$ be a convex set. Then the image of $\mathcal{C}$ under T, i.e., :*

$$T(C) = \{T(x) \in W \mid x \in C \subset V\}$$

is also a convex set.

Lemma E.2 (Cartesian product of convex sets is a convex set). *Let C_1 be a subset of $\mathbb{R}^m$ and C_2 be a convex subset of $\mathbb{R}^n$. Then, the Cartesian product $C_1 \times C_2$ is a convex subset of $\mathbb{R}^m \times \mathbb{R}^n$.*

Two additional results, proven beneath, are required before proceeding to the proof of Theorem 5.1.

Lemma E.3. *The mapping in (E.7), $V : \mathbb{S}_p \to \mathbb{R}^{2p}$, defined, for any $\Gamma \in \mathbb{S}_p$, as:*

$$V(\Gamma) = \left(\sum_{j+k=i+2} \Gamma_{j,k} \right)_{i=0,\dots,2p}$$

is linear.

Proof of Lemma E.3.
We need to show that:

- For $A, B \in \mathbb{S}_p$, $T(A + B) = T(A) + T(B)$;
- For $\alpha \in \mathbb{R}$, $\Gamma \in \mathbb{S}_p$, $T(\alpha\Gamma) = \alpha T(\Gamma)$.

First, one has, for $i = 0, \dots, 2p$:

$$\begin{aligned}\Big[T(A+B)\Big]_i &= \sum_{j+k=2p-i} \Big[A+B\Big]_{jk} \\ &= \sum_{j+k=i+2} A_{jk} + B_{jk} \\ &= \sum_{j+k=i+2} A_{jk} + \sum_{j+k=i+2} B_{jk} \\ &= \Big[T(A)\Big]_i + \Big[T(B)\Big]_i\end{aligned}$$

since it holds for $i = 0, \dots, 2p$, it entails:

$$T(A+B) = T(A) + T(B).$$

Moreover, one has, for $i = 0, \dots, 2p$:

$$\begin{aligned}\Big[T(\alpha\Gamma)\Big]_i &= \sum_{j+k=i+2} \alpha\Gamma_{jk} \\ &= \alpha\Big[T(\Gamma)\Big]_i\end{aligned}$$

and since it holds for $i = 0, \dots, 2p$, it entails:

$$T(\alpha\Gamma) = \alpha T(\Gamma).$$

Hence T is a linear map between $\mathbb{S}_p$ and $\mathbb{R}^{2p}$.

Lemma E.4. *Let S be a univariate polynomial of degree d and $s = (s_0, \dots, s_d)^\top \in \mathbb{R}^{d+1}$ its coefficients. Let S' be its derivative, i.e., a polynomial of degree $d-1$, with coefficients $\breve{s} = (s_1, \dots, s_d)^\top \in \mathbb{R}^d$. Let Z and W be SOS polynomials, with coefficients z and w, and assume that S' is non-negative on $[t_0, t_1]$ as a combination of Z and W as in Theorem E.1. Moreover, let*

$$D = \operatorname{diag}(1, 2, \dots, d)$$

be the $(d \times d)$ diagonal matrix with $(1, \dots, d)$ as a diagonal elements and denote the block-matrices

$$\overline{\mathcal{I}}_{i,d} = \begin{pmatrix} I_d \\ \mathbf{0}_{i,d} \end{pmatrix}, \quad \underline{\mathcal{I}}_{i,d} = \begin{pmatrix} \mathbf{0}_{i,d} \\ I_d \end{pmatrix}, \quad \overline{\underline{\mathcal{I}}}_{i,d} = \begin{pmatrix} \mathbf{0}_{i,d} \\ I_d \\ \mathbf{0}_{i,d} \end{pmatrix}$$

where $\mathbf{0}_{i,d}$ denotes the $(i \times d)$ matrix of zeros, and I_d is the $(d \times d)$ identity matrix. If d is odd, then $z \in \mathbb{R}^d$ and $w \in \mathbb{R}^{d-2}$ and furthermore

$$\breve{s} = Az + Bw$$

where A and B are $(d \times d)$ and $(d \times d-2)$ matrices, respectively. If the degree d of S is even, one has that $z, w \in \mathbb{R}^{d-1}$ and furthermore:

$$\breve{s} = Cz + Dw.$$

where C and D are $(d \times d-1)$ matrices. More specifically,

$$\begin{aligned} A &= \mathcal{D}_d^{-1}, & B &= \mathcal{D}_d^{-1}\left((t_0+t_1)\overline{\underline{\mathcal{I}}}_{1,d-2} - \underline{\mathcal{I}}_{2,d-2} - t_0 t_1 \overline{\mathcal{I}}_{2,d-2}\right), \\ C &= \mathcal{D}_d^{-1}\left(\underline{\mathcal{I}}_{1,d-1} - t_0\overline{\mathcal{I}}_{1,d-1}\right), & D &= \mathcal{D}_d^{-1}\left(t_1\overline{\mathcal{I}}_{1,d-1} - \underline{\mathcal{I}}_{1,d-1}\right). \end{aligned}$$

Proof of Lemma E.4.
First, assume that S is a polynomial of odd degree $d = 2p+1$, meaning that its derivative, S', is a polynomial of even degree $2p$. From Theorem E.1, one has that $S'(x)$ is positive on an interval $[t_0, t_1]$ if

and only if it can be expressed as :

$$S'(x) = Z(x) + (x - t_0)(t_1 - x)W(x)$$

where Z is an SOS polynomial of degree at most equal to $d-1$ and W is an SOS polynomial of degree at most equal to $d-3$. Denote $\breve{s} = (s_1, \dots, s_d) \in R^d$ the coefficients of S' and $z = (z_1, \dots, z_d) \in \mathbb{R}^d$ and $w = (w_1, \dots, w_{d-2)} \in \mathbb{R}^{d-2}$ the coefficients of Z and W respectively. One has that :

$$\begin{aligned} S'(x) &= \sum_{i=1}^{d} i s_i x^{i-1} \\ &= \sum_{j=0}^{d-1} (j+1) s_{i+1} x^i \end{aligned}$$

and if S' is assumed to be non-negative on $[t_0, t_1]$

$$\begin{aligned} S'(x) &= Z(x) + (x - t_0)(t_1 - x)W(x) \\ &= \sum_{j=0}^{d-1} z_{j+1} x^j + (-x^2 + (t_0 + t_1)x - t_0 t_1) \sum_{j=0}^{d-3} w_{j+1} x^j \end{aligned}$$

leading to the following identification :

$$\begin{cases} s_1 = z_1 - t_0 t_1 w_1 \\ s_2 = \frac{1}{2}\left(z_2 - t_0 t_1 w_2 + (t_0 + t_1) w_1\right) \\ s_i = \frac{1}{i}\left(z_i - t_0 t_1 w_i + (t_0 + t_1) w_{i-1} - w_{i-2}\right), \quad \text{for } i = 3, \dots, d-2 \\ s_{d-1} = \frac{1}{d-1}\left(z_{d-1} + (t_0 + t_1) w_{d-2} - w_{d-3}\right) \\ s_d = \frac{1}{d}\left(z_{d-1} - w_{d-2}\right), \end{cases}$$

or, written in a matrix form:

$$\breve{s} = \mathcal{D}_d^{-1}\left(z + \left((t_0 + t_1)\overline{\mathcal{I}}_{1,d-2} - \underline{\mathcal{I}}_{2,d-2} - t_0 t_1 \overline{\mathcal{I}}_{2,d-2}\right) w\right).$$

If S is assumed to be a polynomial of even degree $d = 2p$, S' is necessarily odd degree. From Theorem E.1, one has that $S'(x)$ is positive on an interval $[t_0, t_1]$ if and only if it can be expressed as :

$$S'(x) = (x - t_0)Z(x) + (t_1 - x)W(x)$$

where Z and W are SOS polynomials of degree at most equal to $d-2$ with $z = (z_1, \dots, z_{d-1}) \in \mathbb{R}^{d-1}$ and $w = (w_1, \dots, w_{d-1)} \in \mathbb{R}^{d-1}$ as coefficients, respectively. It leads to the following identification:

$$\begin{cases} s_1 = -t_0 z_1 + t_1 w_1 \\ s_i = \frac{1}{i}\left(z_{i-1} - t_0 z_i + t_1 w_i - w_{i-1}\right) \quad \text{for } i = 2, \dots, d-1 \\ s_d = \frac{1}{d}\left(z_{d-1} - w_{d-1}\right), \end{cases}$$

which can be written in matrix form as

$$\breve{s} = \mathcal{D}_d^{-1}\left(\left(\underline{\mathcal{I}}_{1,d-1} - t_0 \overline{\mathcal{I}}_{1,d-1}\right) z + \left(t_1 \overline{\mathcal{I}}_{1,d-1} - \underline{\mathcal{I}}_{1,d-1}\right) w\right).$$

E.5.2 Proof of the theorem

One can now proceed to prove Theorem 5.1.

Proof of Theorem 5.1.
This rationale can be broken down into two steps: **(a)** proving that the objective function in Eq. (5.17) can indeed be written in a quadratic form, and:**(b)** proving that the problem constraints form a feasible set in $\mathbb{R}^{d+1}$ which is closed and convex.

(a) Notice first that the initial objective function

$$\int_{t_0}^{t_1} (L(x) - F_P^{\rightarrow}(x))^2 dx$$

where $L \in \mathbb{R}[x]_{\leq d}$ with coefficients $s \in \mathbb{R}^{d+1}$, can be rewritten as:

$$\begin{aligned}
\int_{t_0}^{t_1} (F_P^{\rightarrow}(x) - L(x))^2 dx &= \int_{t_0}^{t_1} (\sum_{i=0}^{d} s_i x^i - F_P^{\rightarrow}(x))^2 dx \\
&= \int_{t_0}^{t_1} \left(\left(\sum_{i=0}^{d} s_i x^i \right)^2 + (F_P^{\rightarrow}(x))^2 - 2 \sum_{i=0}^{d} s_i x^i F_P^{\rightarrow}(x) \right) dx \\
&= \int_{t_0}^{t_1} \left(\sum_{i=0}^{d} s_i x^i \right)^2 dx - 2 \sum_{i=0}^{d} s_i \int_{t_0}^{t_1} x^i F_P^{\rightarrow}(x) dx \\
&\quad + \int_{t_0}^{t_1} (F_P^{\rightarrow}(x))^2 \, dx.
\end{aligned}$$

Note that

$$\begin{aligned}
\int_{t_0}^{t_1} \left(\sum_{i=0}^{d} s_i x^i \right)^2 dx &= \sum_{i=0}^{d} \sum_{j=0}^{d} s_i s_j \int_{t_0}^{t_1} x^{i+j} dx \\
&= s^\top M s
\end{aligned}$$

where M is the moment matrix of the Lebesgue measure on $[t_0, t_1]$, i.e., defined entry-wise, for $i, j = 1, \ldots, d+1$ as

$$M_{ij} = \int_{t_0}^{t_1} x^{i+j-2} dx = \frac{(t_1)^{i+j-1} - (t_0)^{i+j-1}}{i+j-1}.$$

and further notice that M is thus positive definite since, for any $u \in \mathbb{R}^{d+1}$,

$$u^\top M u = \int_{t_0}^{t_1} \left(\sum_{i=0}^{d} u_{i+1} x^i \right)^2 dx \geq 0$$

is always non-negative, and equal to 0 if and only if $u_i = 0, i = 1, \ldots, d+1$. Moreover, note that:

$$\sum_{i=0}^{d} s_i \int_{t_0}^{t_1} x^i F_P^{\rightarrow}(x) dx = s^\top r$$

where $r \in \mathbb{R}^{d+1}$ is the moment vector of G with respect to the Lebesgue measure on $[t_0, t_1]$, defined for $i = 0, \ldots, d$ as:

$$r_i = \int_{t_0}^{t_1} x^i F_P^{\rightarrow}(x) dx$$

Since a polynomial is completely characterized by its coefficients, searching for:

$$S^* = \underset{L \in \mathbb{R}[x]_{\leq d}}{\operatorname{argmin}} \int_{t_0}^{t_1} (L(x) - F_P^{\rightarrow}(x))^2 dx$$

is equivalent to finding the coefficients s^* of S^*, i.e.,

$$s^* = \underset{s \in \mathbb{R}^{p+1}}{\operatorname{argmin}} \, s^\top M s - 2 s^\top r$$

and thus proving the first part of the proposition.

(b) Notice that the interpolation constraints

$$\begin{cases} S(t_0) = b_0 \\ S(t_1) = b_1 \end{cases}$$

can be written as

$$\begin{cases} s^\top \mathbf{t_0}^d = b_0 \\ s^\top \mathbf{t_1}^d = b_1 \end{cases}$$

where, for $a \in \mathbb{R}$, one denote $\mathbf{a}^d$ the vector of powers of a up to d, i.e., $\mathbf{a}^d = (1, a, \dots, a^{d-1}, a^d) \in \mathbb{R}^{d+1}$. Moreover, by letting:

$$\mathbf{T} = \begin{pmatrix} \mathbf{t_0}^d \\ \mathbf{t_1}^d \end{pmatrix}, \quad b = \begin{pmatrix} b_0 \\ b_1 \end{pmatrix},$$

where $\mathbf{T}$ is a $(2 \times d + 1)$ block-matrix, the constraint can be written as:

$$Ts = b.$$

Furthermore, notice that

$$\mathcal{C}_0 = \{s \in \mathbb{R}^{d+1} \mid Ts = b\}$$

is a convex subset of $\mathbb{R}^{d+1}$, since the equality constraints are linear. Concerning the monotonicity constraint

$$S'(x) \geq 0, \quad \forall x \in [t_0, t_1],$$

from Lemma E.4 one can quite generically write

$$\begin{pmatrix} s_d \\ \vdots \\ s_1 \end{pmatrix} = T_0(z, w) := Az + Bw$$

where z and w are the coefficients of SOS polynomials of degrees depending on d. Additionally, notice that the mapping $T_0 : \mathbb{R}^d \times \mathbb{R}^{d-2} \to \mathbb{R}^d$ is linear. Next, let $V_1 : \mathbb{S}_p \to \mathbb{R}^{2p}$, and $V_2 : \mathbb{S}_q \to \mathbb{R}^{2q}$ be defined as in (E.7), where $p = d - 1/2$ and $q = d - 3/2$ if d is odd, or $p = d - 2/2$ and $q = d - 2/2$ if d is even, and note that both mappings are linear thanks to Lemma E.3.

Moreover, denote the following sets:

$$\mathcal{Z} = \{V_1(E) \mid E \in \Sigma_p\}, \quad \mathcal{W} = \{V_2(E) \mid E \in \Sigma_{p-1}\}$$

and notice the polynomial Z (resp. W) is SOS if and only its coefficients z (resp. w) are in $\mathcal{Z}$ (resp. $\mathcal{W}$) thanks to Theorem E.3. In addition again, notice that, thanks to Lemma E.1, and due to the fact that Σ_p is a closed convex set in $\mathbb{S}_p$ as per Theorem E.3, both $\mathcal{Z}$ and $\mathcal{W}$ are convex subsets of $\mathbb{R}^{2p}$ and $\mathbb{R}^{2q}$ respectively. Besides, thanks to Lemma E.2, the set $\mathcal{Z} \times \mathcal{W}$ is a convex subset of $\mathbb{R}^{2p} \times \mathbb{R}^{2q}$ as well. Moreover, let

$$\mathcal{C}_1 = \left\{ \begin{pmatrix} T_0(w, z) \\ x \end{pmatrix} \in \mathbb{R}^{d+1} \mid x \in \mathbb{R}, \quad (z, w) \in \mathcal{Z} \times \mathcal{W} \right\}$$

and note that it is a convex subset of $\mathbb{R}^{d+1}$ due to the fact that T_0 is a linear map.

Finally, since both $\mathcal{C}_0$ and $\mathcal{C}_1$ are convex sets, their intersection:

$$\mathcal{K} = \mathcal{C}_0 \cap \mathcal{C}_1$$

is as well, and note that any element $s \in \mathcal{K}$ are the coefficients of a polynomial respecting both equality and monotonicity constraints. In other words, $\mathcal{K}$ is the feasible set of coefficients of the initial optimization problem.

Appendix F

Additional use-cases

Contents

E.1 Some preliminaries **148**
E.1.1 Application domain 148
E.1.2 Cumulative distribution functions and quantile functions 148
E.1.3 Wasserstein distance 149
E.2 Dependence modelling and copulas **150**
E.2.1 Copulas 150
E.3 Computational details and code snippets **150**
E.3.1 Moment matrix of the Lebesgue measure 150
E.3.2 Computing moment vector of arbitrary quantile functions 151
E.4 Proofs **152**
E.5 Proof of Theorem 5.1 . . . . . **154**
E.5.1 Ingredients 154
E.5.2 Proof of the theorem 157

F.1 A COVID-19 epidemiological model

Remark. This use-case has been initially studied in:

M. Il Idrissi, V. Chabridon, and B. Iooss. Developments and applications of Shapley effects to reliability-oriented sensitivity analysis with correlated inputs. *Environmental Modelling and Software*, 143:105115, 2021. ISSN: 1364-8152. DOI: 10.1016/j.envsoft.2021.105115

In 2020, the COVID-19 crisis has raised major issues in the usefulness of epidemic modeling in order to give useful insights to public policy decision-makers. [193] have taken this example to insist on the essential use of GSA on such models, which claim to predict the potential consequences of intervention policies. A first study has been proposed by [148], in the context of COVID-19 in Italy, to assess the sensitivity of model outputs such as quarantined, recovered or dead people to inputs driving intervention policies. Another GSA has been performed in [48] in the French context of the first COVID-19 wave. By using data coming from this last analysis (thanks to the author's agreement), the goal of this section is to demonstrate how TSA can help to characterize the influence of various uncertain parameters on a real-scale model.

F.1.1 Model description

The deterministic compartmental model developed in [48] is representative of the COVID-19 French epidemic (from March to May) by taking into account the asymptomatic individuals, the testing strategies, the hospitalized individuals, and people going to Intensive Care Unit (ICU). Using several assumptions, it is based on a system of 10 ordinary differential equations, which are not developed here for the sake of conciseness (see [48] for more information).

Table F.1 presents the 20 input parameters with their prior distribution (chosen from literature studies), which form the inputs X, assumed to be mutually independent. For the present study, our variable of interest, which is a particular model output, then writes

$$U_{\max}^{\mathrm{p}} = \max_{v \in \text{time range}} \Big\{ U_v(X) \Big\} \tag{F.1}$$

where U_v is the number of hospitalized patients in ICU at time v. Note that the "p" in $U_{\max}^{\mathrm{p}}$ stands for "prior" as this quantity corresponds to the variable of interest before any calibration w.r.t. the available data.

In [48], after a first screening step allowing the removal of non-influential inputs, the model is calibrated on real data by using a Bayesian calibration technique. After the analysis of this step, the selected remaining inputs are

$$X_{\mathrm{sel}} = (p_a, N_a, N_s, R_0, t_0, \mu, N, I_0^-)^\top \tag{F.2}$$

and their distributions are obtained from a sample given by the calibration process. The non-influential inputs are fixed to their nominal values, and the posterior variable of interest becomes

$$U_{\max} = \max_{v \in \text{time range}} \Big\{ U_v(X_{\mathrm{sel}}) \Big\} \tag{F.3}$$

with $U_{\max}$ being the maximum number of hospitalized people in ICU who need special medical care on the studied temporal range, and U_v is the number of hospitalized patients in ICU at time v.

Note that the "p" in $U_{\max}^{\mathrm{p}}$ stands for "prior" as this quantity corresponds to the variable of interest before any calibration w.r.t. the available data.

In [48], after a first screening step that allows for suppressing non-influential inputs, the model is calibrated on real data by using a Bayesian calibration technique. After the analysis of this step, the selected remaining inputs are

$$X_{\mathrm{sel}} = (p_a, N_a, N_s, R_0, t_0, \mu, N, I_0^-)^\top \tag{F.4}$$

and their distributions are obtained from a sample given by the calibration process. The non-influential inputs are fixed to their nominal values and the posterior variable of interest becomes

$$U_{\max} = \max_{v \in \text{time range}} \Big\{ U_v(X_{\mathrm{sel}}) \Big\} \tag{F.5}$$

Input	Description	Prior distribution
p_a	Conditioned on being infected, the probability of having light symptoms or no symptoms	$\mathcal{U}(0.5, 0.9)$
p_H	Conditioned on being mild/severely ill, the probability of needing hospitalization (H or U)	$\mathcal{U}(0.15, 0.2)$
p_U	Conditioned on going to hospital, the probability of needing ICU	$\mathcal{U}(0.15, 0.2)$
p_{HD}	Conditioned on being hospitalized but not in ICU, the probability of dying	$\mathcal{U}(0.15, 0.25)$
p_{UD}	Conditioned on being admitted to ICU, the probability of dying	$\mathcal{U}(0.2, 0.3)$
N_a	If asymptomatic, number of days until recovery	$\mathcal{U}(8, 12)$
N_s	If symptomatic, number of days until recovery without hospital	$\mathcal{U}(8, 12)$
N_{IH}	If severe symptomatic, number of days until hospitalization	$\mathcal{U}(8, 12)$
N_{HD}	If in H, number of days until death	$\mathcal{U}(15, 20)$
N_{UD}	If in ICU, number of days until death	$\mathcal{U}(8, 12)$
N_{HR}	If hospitalized but not in ICU, the number of days until recovery	$\mathcal{U}(15, 25)$
N_{UR}	If in ICU, number of days until recovery	$\mathcal{U}(15, 25)$
R_0	Basic reproducing number	$\mathcal{U}(3, 3.5)$
t_0	Starting date of epidemics (in 2020)	$\mathcal{U}(01/25, 02/24)$
μ	Decaying rate for transmission (after social distanciation and lockdown)	$\mathcal{U}(0.03, 0.08)$
N	Date of effect of social distanciation and lockdown	$\mathcal{U}(20, 50)$
λ_1	Type-1 testing rate	$\mathcal{U}(1e-4, 1e-3)$
p_{HU}	Conditioned on being hospitalized in H, the probability of needing ICU	$\mathcal{U}(0.15, 0.2)$
N_{HU}	If in H, number of days until ICU	$\mathcal{U}(1, 10)$
I_0^-	Number of infected undetected at the start of epidemics	$\mathcal{U}(1, 100)$

Table F.1: Model inputs and their prior distribution. H is the number of hospitalized individuals with severe symptoms. U is the number of hospitalized individuals in ICU.

with $U_{\max}$ being the maximum number of hospitalized people in ICU who need special medical care on the studied temporal range, and U_v is the number of hospitalized patients in ICU at time v.

F.1.2 Importance quantification for ICU bed shortage

The central question of this study would be to determine which inputs influence the event of a country experiencing a shortage of ICU bed capacity during the time period. For that purpose, one can introduce a threshold k, which represents the total number of ICU beds in the country, which is assumed to be constant during the studied time period. The new variable of interest would then be $\mathbb{1}_{\{U^{\mathrm{P}}_{\max}>k\}}(X)$ for the full compartmental model (preliminary study) and $\mathbb{1}_{\{U_{\max}>k\}}(X_{\mathrm{sel}})$ for the model with selected

inputs (post-calibration study). Two input-output samples of size $n = 5000$ are available. The first one (preliminary study) includes all the inputs following their prior distribution (see Table F.1) and the corresponding output $U^P_{\max}$ of the compartmental model. The second one (post-calibration study) is composed of a sample of X_{sel} after the Bayesian calibration, and the corresponding output $U_{\max}$ of the compartmental model with the non-selected inputs fixed to their nominal values.

Five different thresholds are studied on $U^P_{\max}$: $5 \cdot 10^3$, 10^4, $5 \cdot 10^4$, 10^5 and $2 \cdot 10^5$, with respectively 58.1%, 47.7%, 22%, 10.1% and 2.2% of the total output samples being in a failure state. This illustrates the behavior of the target Shapley effects when the failure probability decreases. The threshold of 6300 has been chosen for $U_{\max}$, with 10.9% of the total output samples being above this threshold. Figure F.1 illustrates two different thresholds and the corresponding estimated failure probability on the histogram of both outputs.

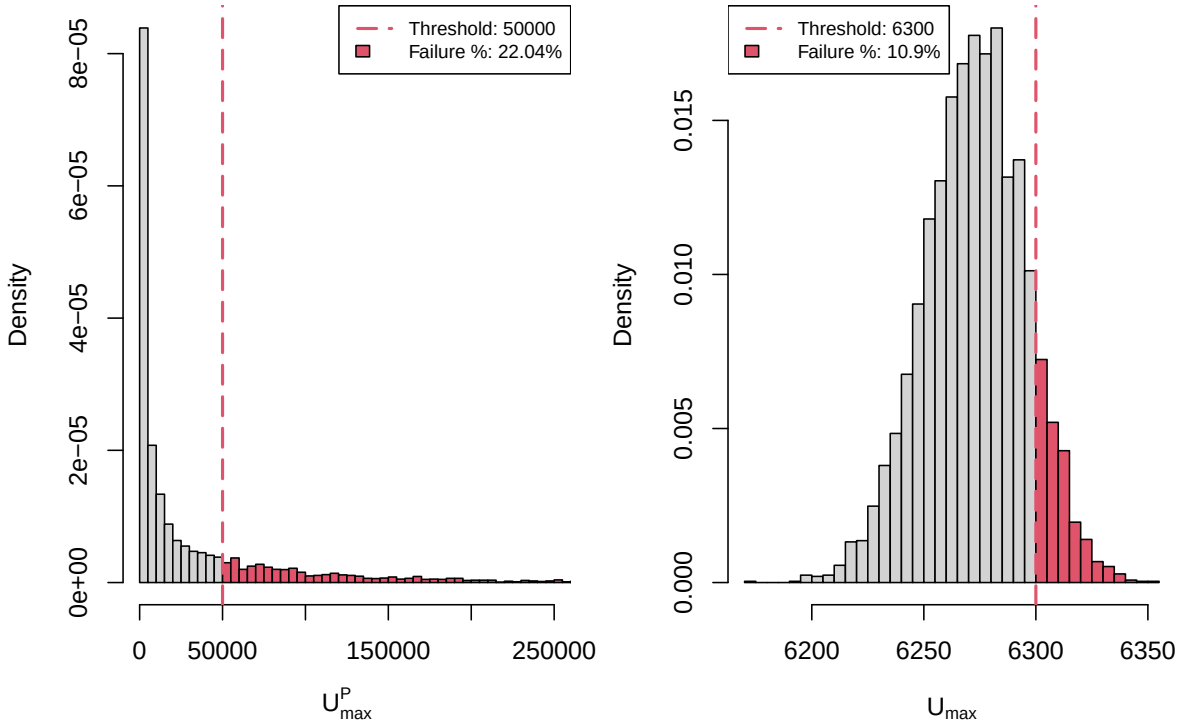

Figure F.1: Illustration of thresholds on the histograms of $U^P_{\max}$ (left) and $U_{\max}$ (right).

The target Shapley effects have been estimated using a variant of the nearest-neighbor estimation scheme, with a fixed number of random permutations of 10^3, and with a number of neighbors set to 3, following the rule of thumb guideline of [33], due to the sheer complexity of this estimation algorithm. Since the compartmental model is deterministic, the target Shapley effects have been forced to sum up to one. Figure F.2 presents the main results for $U^P_{\max}$, with the red dotted line being the average influence of an input, in the case of similar importance (i.e., $\frac{1}{20}$). One can notice that for less restrictive thresholds (i.e., threshold for which the failure probability is high), the input N, the effective date of lockdown/social distanciation measures, seem to be the most influential, reaching more than 50% of the TSA variable of interest's variance. However, as soon as the threshold becomes more and more restrictive (i.e., the failure probability becomes lower and lower), the effect of N decreases and the effects of the other inputs increase accordingly, in order to reach what seems to be an equilibrium at the value $\frac{1}{20}$. This behavior can be explained by two main reasons:

- The nature of a restrictive TSA variable of interest induces high interaction between the inputs;
- The Shapley allocation system, when applied to variance as a production value, redistributes the interaction effects equally between all inputs (there is no correlation between inputs in this prior study).

One can argue that, as soon as k becomes very restrictive, the combined interaction effects outweigh the

effect of N itself, and since these effects are equally distributed among all the inputs, their share will tend to go towards $\frac{1}{20}$.

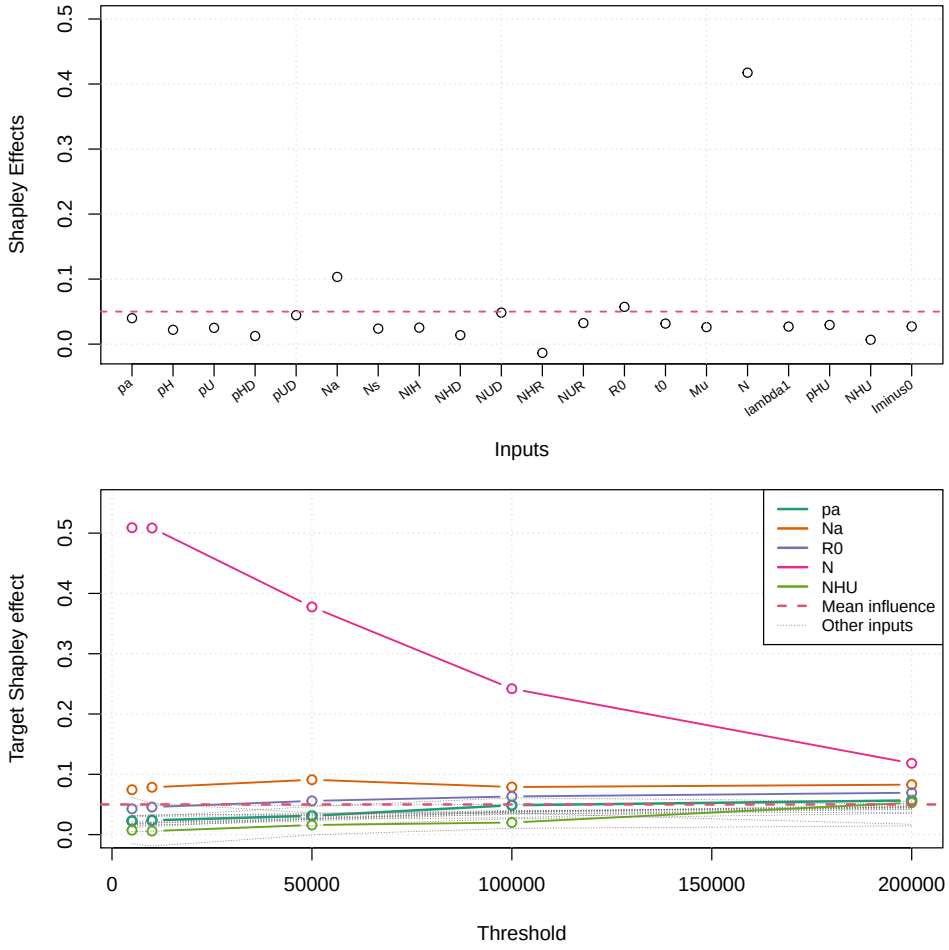

Figure F.2: Shapley effects (top) and target Shapley effects for different thresholds (bottom) for $U^P_{\max}$.

For the post-calibration study, some selected inputs X_{sel} are linearly correlated (see Figure F.3 - top. This is typically the case for N and μ, with an estimated correlation coefficient $\hat{\rho}(N, \mu) = 0.69$, and for R_0 and N with an estimated correlation coefficient of $\hat{\rho}(N, R_0) = -0.66$. This correlation structure does not allow for interpretable Sobol' indices, which encourages the use of Shapley-inspired indices. The Shapley effects and the target Shapley effects of X_{sel} for $U_{\max}$ have been computed using the nearest-neighbor procedure, with a fixed number of neighbors of 3, and forced to sum to one because of the deterministic nature of the model.

In Figure F.3 (bottom), one can notice that N_a, the number of days until recovery, seem to be the most important input in explaining the number maximum number of ICU patients on the studied time range, with a Shapley effect of around 35% of the output variance. The inputs p_a, N_s, R_0 and N seem to present average effects, that is around $\frac{1}{8}$, while t_0, μ and I_0^- seem to be less influential, with around 5% of explained variance each.

However, focusing on the occurrence of an ICU bed shortage, one can notice that the target Shapley effect of N_a is lower (around 22%), with the influence of N being higher (around 15%) than their Shapley effects. Moreover, t_0, μ and I_0^- present higher TSA effects, i.e., slightly under 10%, due to the interaction induced by the indicator function. One can also remark that the influence of N_s is higher than that of R_0 in the TSA setting, which was the inverse for the Shapley effects. This would indicate that N_s, the number of days until recovery for a symptomatic patient without hospitalization, has more influence on the event of a bed shortage than the basic reproducing number of the virus, R_0.

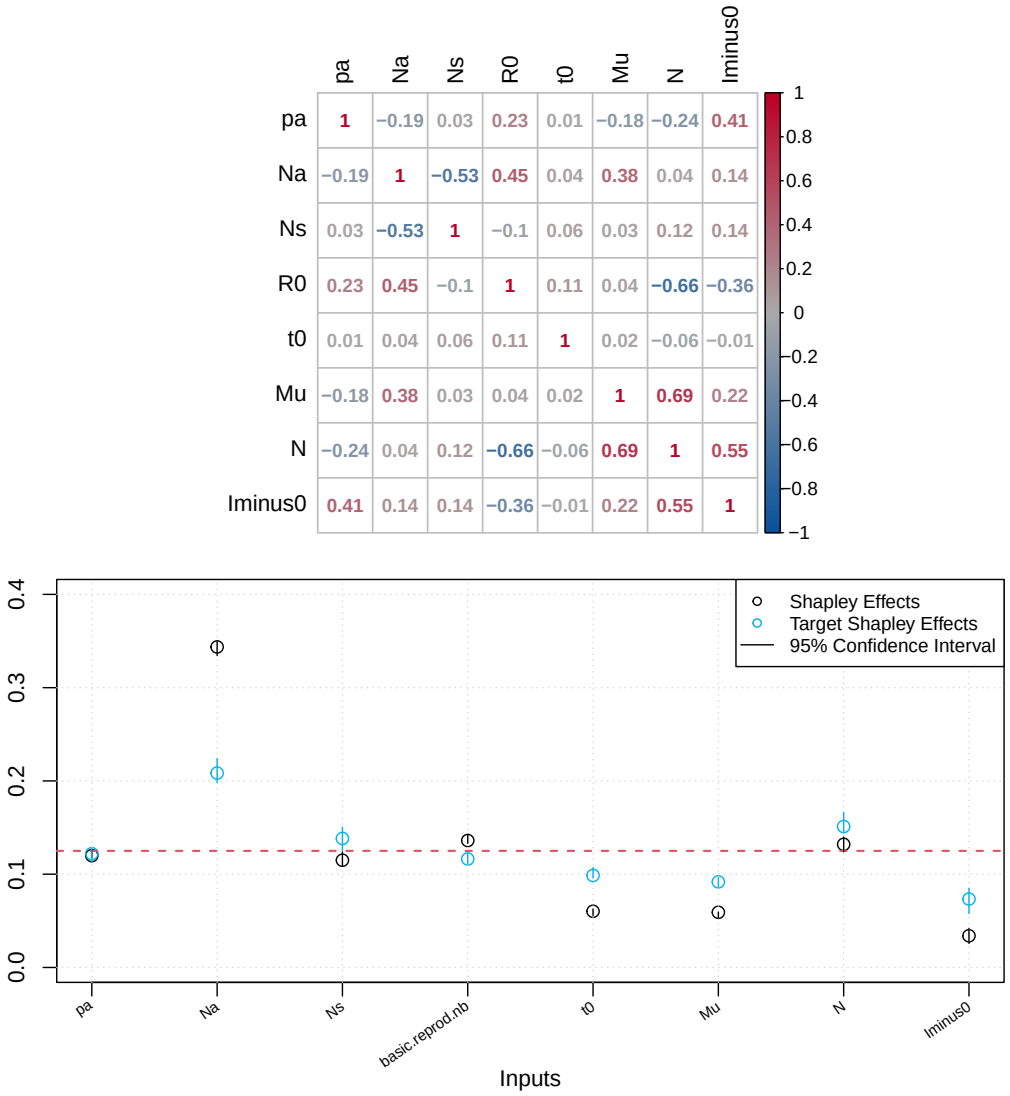

Figure F.3: Input correlation matrix (top), Shapley effects for $U_{\max}$ and target Shapley effects (bottom) for $\mathbb{1}_{\{U_{\max}>t\}}(X_{\text{sel}})$. The 95% confidence intervals have been computed by uniformly selecting 80% of the observations, for 100 repetitions, without replacement.

F.2 Ultrasonic non-destructive control of a weld

Remark. This use-case has been initially studied and presented in

M. Herin. Proportional values: an alternative to Shapley values in sensitivity analysis. Msc Internship Dissertation, EDF R&D, EDF Lab Chatou, 2021

M. Il Idrissi, M. Hérin, and V. Chabridon. Cooperative game theory and global sensitivity analysis. École Thématique sur les Incertitudes en Calcul Scientifique (ETICS), Erdeven, France, 2021. URL: https://www.gdr-mascotnum.fr/etics.html

This industrial application considers the ultrasonic non-destructive control of a weld containing manufacturing defects. This use-case is presented in-depth in [191] and has been used as a means for illustration of the Shapley Effects in [121].

F.2.1 Model description

Ultrasonic non-destructive control allows for detecting defects in certain industrial installations. For example, the ability to reliably detect defects in welds is a crucial problem for power plant operators.

However, certain types of welds allow for complex materials, which can result in disturbances in the wave propagation during the ultrasonic control (e.g., deviation/division of the beam, attenuation of the wave amplitude...). As a consequence, the interpretation of such analysis can be challenging, due to those complex phenomena.

In order to solve this problem, EDF has developed a simulation tool called `ATHENA2D`, which is dedicated to simulating elastic wave propagation in heterogeneous and anisotropic materials, which is the case of defective welds. As depicted in Figure F.4, in this use-case, a weld is decomposed in 7 different regions (hence the heterogeneity), each of which has a different columnar grain orientation (hence the anisotropy). The seven domains are assumed to be homogenous (i.e., the grain orientation is the same throughout the domain). The `ATHENA2D` tool relies on finite-element code, and, in this case, takes in eleven inputs: four of which are elastic coefficients relative to the welding material, and the remaining seven serve for describing the orientation of the grain in each region of the defective weld. The output of this numerical model is the amplitude of the defect echoes resulting from an ultrasonic inspection.

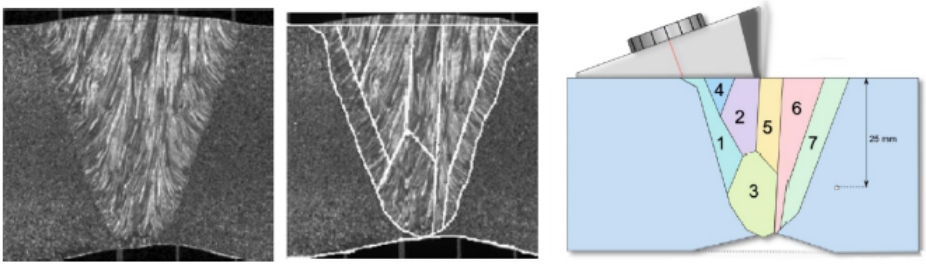

Figure F.4: Metallographic picture of a defective weld (left), description of the weld in seven homogeneous domains (middle), inspection configuration [121].

In [191], a probabilistic modeling of the inputs, with an independence assumption between the grain orientations, has been made in order to perform a sensitivity analysis. In [121], this assumption has been questioned, and another probabilistic model has been proposed, allowing for dependency between the inputs. Following this paper, it has been estimated using physical models that the correlation matrix between the 7 different grain orientations $(Or_1, \dots, Or_7)$ is:

$$\Sigma = \begin{pmatrix} 1 & 0.8 & 0.74 & 0.69 & 0.31 & 0.23 & 0.20 \\ 0.8 & 1 & 0.64 & 0.53 & 0.59 & 0.51 & 0.46 \\ 0.74 & 0.64 & 1 & 0.25 & 0.60 & 0.57 & 0.54 \\ 0.69 & 0.53 & 0.25 & 1 & -0.25 & -0.35 & -0.33 \\ 0.31 & 0.59 & 0.60 & -0.25 & 1 & 0.96 & 0.84 \\ 0.23 & 0.51 & 0.57 & -0.35 & 0.96 & 1 & 0.95 \\ 0.20 & 0.46 & 0.54 & -0.33 & 0.84 & 0.95 & 1 \end{pmatrix} \tag{F.6}$$

Traditional GSA (mainly, the Sobol' indices) fail to give meaningful insights due to the absence of independence. This linear dependency between the inputs motivates the use of the cooperative games' framework, in order to produce interpretable indices in such a complex setting. The following results aim at leveraging the power of the given-data estimation procedures. No definitive probabilistic model is used, but the different samples are approximated using the KNN method.

The inputs are defined as $X = (C_{11}, C_{13}, C_{33}, C_{55}, Or_1, \dots Or_7) \in \mathbb{R}^{11}$. The first four inputs correspond to the elastic coefficients (GPa) of a 316L Stainless Steel weld[1], and the remaining seven inputs correspond to the orientation of the grain (degrees) of each of the seven homogenous domains. Let $G(X)$ be the output of the `ATHENA2D` model.

Remark F.1. Due to the time-consuming nature of the `ATHENA2D` code, and to the restrictions relative to the timing of this study, an accurate approximation through the use of a meta-model is going to be used as a placeholder of the original numerical model.

A Gaussian process has been fitted using a sample of input-outputs from the `ATHENA2D` model, following the methodology presented in [121]. An i.i.d. sample of inputs of size $m = 10000$ has been simulated

[1] Expressed using Voigt notation.

with respect to the joint law of the inputs, which are assumed to be Gaussian, with the correlation matrix depicted in Equation F.6. The elastic coefficients of the 316L stainless steel are assumed to be independent of each other and of the rest of the inputs.

In the following, the predictions made by the meta-model are assumed to be sufficiently close to the real values that the `ATHENA2D` would provide to interpret the results as if they were produced directly from the numerical model.

F.2.2 Importance quantification

[121] estimated the Shapley effects of the input variables to measure the influence of the inputs on the output's variance and ranked them into three groups:

- Or_1 and Or_3 with a share of more than 20% of the output's variance;
- Or_2 with 11%;
- C_{11} and Or_4 to Or_7, with effects ranging between 6 and 8%;
- C_{33}, C_{55} and C_{13} with effects lower than 3%.

In order to compare this classification with the one provided by the PMEs, both have been computed on 5 simulated datasets of size $m = 10000$ composed with i.i.d observations. Then, the means and standard errors of the effects have been computed. The estimated importance attributions are displayed in Figure F.5.

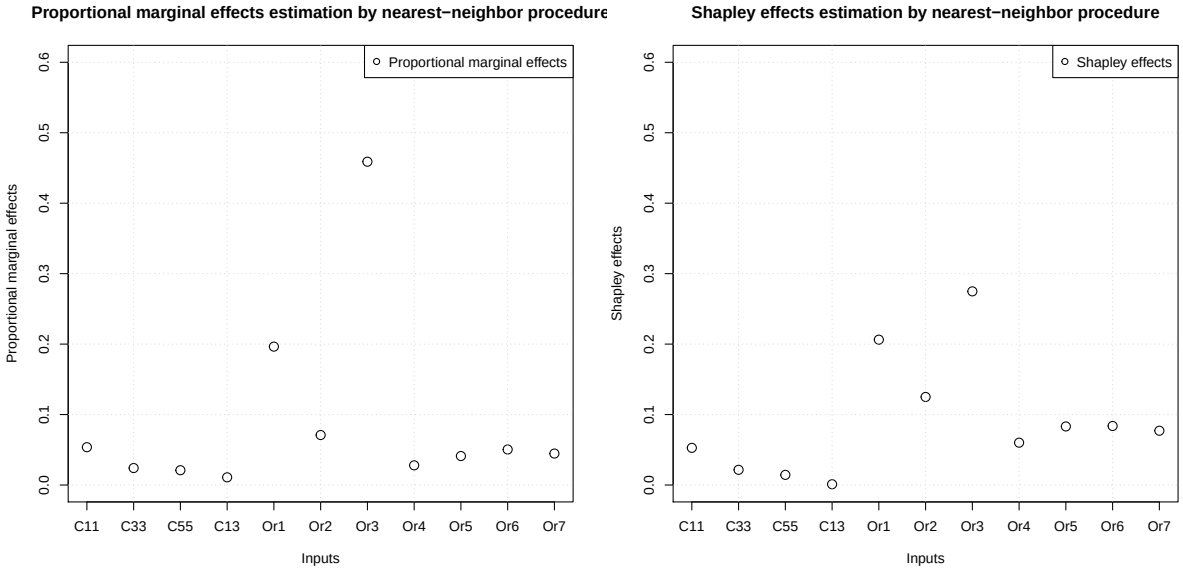

Figure F.5: Proportional marginal and Shapley effects.

One can notice the difference between the Shapley effects importance ranking and the one based on the proportional marginal effects:

- O_{r3} with a share of more than 45 %;
- O_{r1} with a share of more than 20 %;
- C_{11}, O_{r2} with a share around 5 %;
- $C_{13}, C_{33}, C_{55}, O_{r4}, O_{r5}, O_{r6}, O_{r7}$ with a share of less than 5 %.

The main difference between both rankings is that the PM ranking is more discriminative. On the other side, the Shapley ranking is more uniform, i.e., the variance is distributed more equitably between all

inputs. More precisely, one can see that O_{r1} is granted almost half of the variance for the PM effects against only 20% for the Shapley effects. However, the two most important variables remain unchanged (O_{r1} and O_{r3}). The difference between both effects is mainly visible for the inputs O_{r2},O_{r3},O_{r4},O_{r5}, O_{r6} and O_{r7}. One can notice that this group of inputs, with O_{r1}, is the correlated group of inputs. More precisely, in comparison to Shapley effects, the PME increases largely the importance of O_{r2} at the expense of O_{r3},O_{r4},O_{r5}, O_{r6} and O_{r7}. Again, in the correlated case, the PME allows highlighting some inputs, where the Shapley values tend to standardize the importance throughout correlated inputs.

F.3 Robot arm model

Remark. This use-case has been initially studied in the supplementary material of:

M. Herin, M. Il Idrissi, V. Chabridon, and B. Iooss. Proportional marginal effects for global sensitivity analysis. *SIAM/ASA Journal on Uncertainty Quantification*, 2024. URL: https://hal.science/hal-03825935. In press

In this use-case, a model of the position (on the two-dimensional plane) of a robot arm with four segments is studied [6].

F.3.1 Model description

The arm shoulder is fixed at the origin, and the robot's segments have lengths L_i ($i = 1, \ldots, 4$). Each segment is positioned at an angle A_i ($i = 1, \ldots, 4$) with respect to the horizontal axis. While, in the original model, the inputs are assumed to be independent, statistical dependence is introduced here between the angles and between the segment lengths. The probabilistic structure of the inputs can be described as follows:

- The angles A_i ($i = 1, \ldots, 4$) follow a uniform distribution over $[0, 2\pi]$. They are pairwise correlated by the way of a Gaussian copula with a correlation parameter equal to 0.95;
- The lengths are sequentially built: L_1 follows a uniform distribution over $[0, 1]$, while L_i ($i = 2, \ldots, 4$) follows a uniform distribution over $[0, L_{i-1}]$. These inequality constraints create strong correlation between the lengths.

The model's output is the distance from the end of the robot arm to the origin, and writes:

$$Y = \left\{ \left[\sum_{i=1}^{4} L_i \cos\left(\sum_{j=1}^{i} A_j \right) \right]^2 + \left[\sum_{i=1}^{4} L_i \sin\left(\sum_{j=1}^{i} A_j \right) \right]^2 \right\}^{1/2}. \tag{F.7}$$

In this model, A_1 is an exogenous variable: it is intuitive if we think about the mechanism of the robot arm because the angle between the origin and the first arm cannot have any effect on the distance output. Moreover, if Eq. (F.7) is developed using elementary trigonometric formulas, A_1 does no longer appear in the formula.

A unique i.i.d. sample of size 2000 of these 8 inputs has been simulated, on which the output of the model has been computed. Figure F.6 illustrates this data sample by the way of the pairwise scatterplots, the marginal distributions of each input by means of histograms, and the dependence structure with estimated correlation coefficients. One can also notice first-order tendencies of the different inputs on the output Y (last row).

F.3.2 Importance quantification

Since, in this scenario, only an i.i.d. sample is available, the Shapley effects and PMEs have been computed using the nearest-neighbor procedure (with an arbitrarily chosen number of nearest neighbors equal to 6). Figure F.7 presents the Shapley effects and PMEs estimates with a 90%-confidence intervals computed on 100 replications of estimated effects by random selection of 80% of the dataset's observations. According to both effects, the most influential input is L_1, with a Shapley effect around 35%

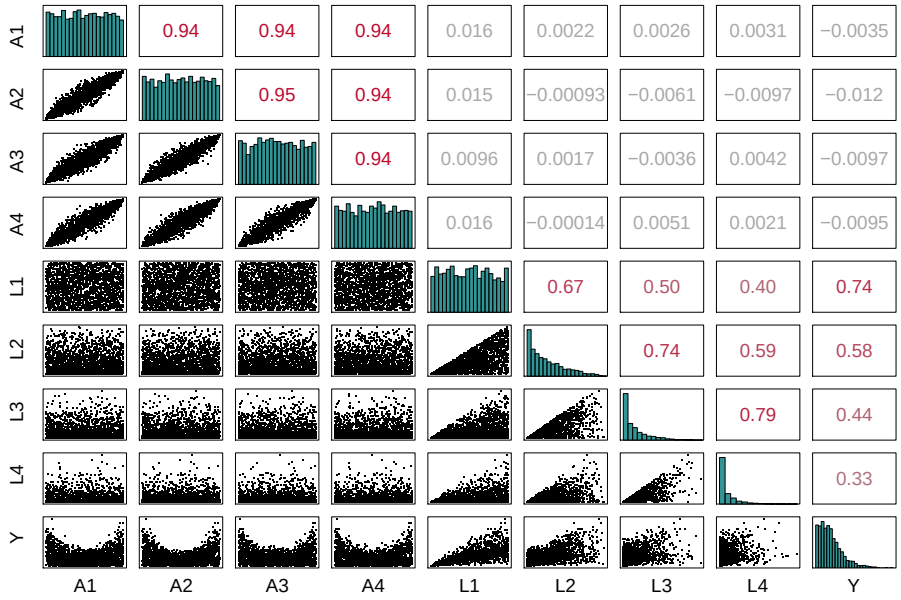

Figure F.6: Pairwise plots of the 2000-size i.i.d. sample for the robot arm model. Histograms of the variables (diagonal), scatterplots (lower part), and correlation coefficients (upper part).

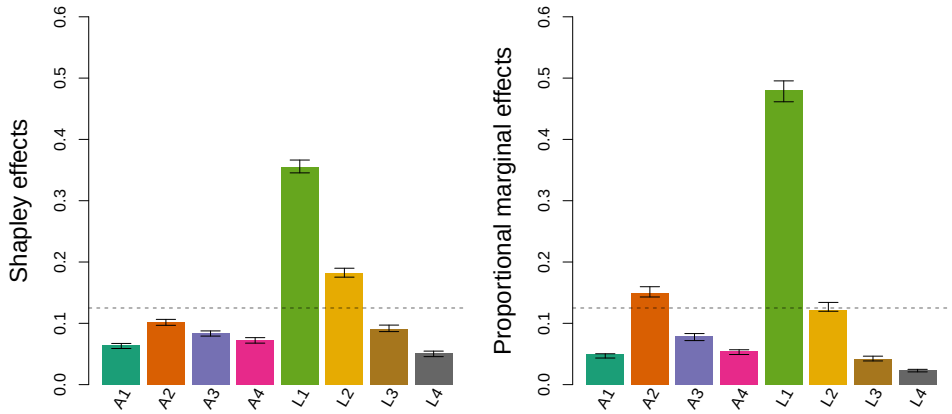

Figure F.7: Shapley effects and PMEs estimators by nearest-neighbor procedure for the robot arm model. The vertical error bars represent the 5% and 95% quantiles of the estimates computed on 100 repetitions where 80% of the initial data is randomly selected (without replacement). The horizontal grey dashed bar represents the average importance of $1/8$.

and PME around 48% of the output's variance. While both effects seem to agree on the most influential input, they offer a fairly different influence hierarchy, as depicted in Table F.2. This different influence hierarchy can be explained by the fairly high correlation between the inputs. For instance, L_2 has Shapley effects around 18% while having a linear correlation coefficient with L_1 equal to 0.67, whereas it has a PME of around 12.1%. Additionally, focusing on the angle inputs, which are very linearly correlated, one can notice that their Shapley effects are relatively equal, varying between 6% and 10%. On the other hand, their PME grants A_2 nearly 15%, with reduced influence of the other angles. Only using the Shapley effects did not consider A_2 as an above-average influential variable, while the PMEs consider it as important. This highlights the more discriminating power of the PME for influence ranking in situations of highly correlated inputs, where the Shapley effects typically grant a similar output variance share to each correlated input.

Moreover, one can notice that the exogenous input A_1 has a low but non-zero PME. This illustrates the bias induced by the nearest-neighbor procedure used to perform estimations [33] and thus, the detection of exogenous inputs is not guaranteed if this estimation method is used. However, in this case, the Monte Carlo estimation method cannot be used.

This use-case illustrates the more discriminating power of the PMEs compared to the Shapley effects,

	Shapley effects		PMEs	
Influence Rank	Input	Value	Input	Value
1	L_1	35.4%	L_1	48%
2	L_2	18.2%	A_2	14.9%
3	A_2	10.2%	L_2	12.1%
4	L_3	9%	A_3	8%
5	A_3	8.4%	A_4	5.5%
6	A_4	7.2%	A_1	4.9%
7	A_1	6.4%	L_3	4.2%
8	L_4	5.1%	L_4	2.4%

Table F.2: Influence hierarchy between the inputs w.r.t. the Shapley effects and the PMEs, on the robot arm use-case.

in cases of highly correlated inputs. Overall, the PMEs favor the already most influential inputs at the expense of the inputs they are correlated with. This behavior is particularly interesting in a screening setting, along with the ability of the PMEs to detect exogenous inputs, while maintaining a meaningful interpretation as shares of variance.

Appendix G

Résumé étendu

Contents

F.1 A COVID-19 epidemiological model 162
F.1.1 Model description 162
F.1.2 Importance quantification for ICU bed shortage 163
F.2 Ultrasonic non-destructive control of a weld 166
F.2.1 Model description 166
F.2.2 Importance quantification 168
F.3 Robot arm model 169
F.3.1 Model description 169
F.3.2 Importance quantification 169

G.1 Contexte et motivation

Il apparaît comme naturel, lorsqu'on se confronte à l'étude de phénomènes physiques, d'initier des expériences répétées. Traditionnellement, les investigations en ingénierie se fondent sur ces expériences afin d'extraire des connaissances en testant différentes configurations, telles que la modification de l'environnement ou des conditions initiales. Les résultats sont ensuite soumis à une analyse comparative. Cette approche constitue l'essence même du domaine de la *physique expérimentale*. Cependant, à mesure que l'innovation s'est développée et que l'ambition industrielle a pris de l'ampleur, la réalisation de telles expériences s'est rapidement révélée trop coûteuse, dangereuse, complexe, voire impossible. Plus récemment, l'*ingénierie moderne* a adopté une approche novatrice en remplaçant les configurations expérimentales par des *modèles physiques*, conduisant à des *simulations numériques* des phénomènes étudiés.

Ces *modèles numériques* ont démontré leur utilité en transformant les pratiques industrielles contemporaines. Ces avancées en modélisation permettent, par exemple, d'anticiper la rentabilité des parcs éoliens en fonction de leur emplacement ou d'améliorer la conception des centrales nucléaires pour assurer leur sûreté et leur résilience face à des événements rares tels que des catastrophes naturelles. Électricité de France (EDF), et plus spécifiquement sa branche de recherche et développement (EDF R&D), joue un rôle essentiel dans le développement, la certification et la diffusion de ces simulateurs pour la production et la distribution d'électricité[1]. L'objectif de ces modèles numériques est de simuler des systèmes pouvant être considérés comme *critiques*, leur fiabilité étant d'autant plus cruciale dans un contexte industriel pour la prise de décisions.

A mesure que ces outils ont gagné en popularité, proportionnellement à l'ampleur des besoins industriels, certaines simulations sont devenues *trop complexes pour être étudiées analytiquement*. La réalisation de simulations des phénomènes physiques étudiés, malgré l'accès à d'importantes capacités de calcul, *se révèle particulièrement chronophage*. De plus, certains modèles physiques englobent des équations complexes (par exemple, les équations de Navier-Stokes) qui sont souvent résolues numériquement. En raison de leur complexité intrinsèque, ces modèles numériques peuvent être considérés comme des *boîtes noires*.

Les simulations réalisées au fil du temps ont été consignées dans des bases de données, tout comme les mesures terrain effectuées sur les sites industriels. Cette abondance de données, combinée à la démonstration de l'efficacité des méthodes d'apprentissage supervisé pour modéliser des phénomènes complexes, a naturellement suscité la question suivante : *Comment ces méthodes de modélisation basées sur les données peuvent-elles améliorer les processus industriels ?*

En particulier, ces méthodes promettent de fournir une solution aux simulations chronophages des modèles numériques en offrant des substituts rapides à évaluer. Elles permettent également d'exploiter les données issues de capteurs pour modéliser des phénomènes complexes qui ne peuvent pas encore être simulés numériquement, voire qui ne peuvent pas l'être. La fiabilité de ces méthodes doit être évaluée pour favoriser leur adoption en tant que composante intrinsèque de la modélisation des systèmes critiques.

Cependant, les progrès récents en intelligence artificielle ont conduit à la résurgence de modélisations surparamétrées. Ces dernier sont très efficaces, mais sont également considérés comme des boîtes noires dû à leur complexité. Leur fiabilité est donc difficile à attester de manière analytique.

La principale difficulté réside dans le fait que les systèmes critiques sont généralement soumis à des incertitudes. Ces incertitudes peuvent provenir de diverses causes (par exemple, manque de connaissance, erreurs de mesure, ou aléas intrinsèques aux phénomènes étudiés). Comprendre et contrôler les effets de ces incertitudes sur les systèmes critiques est crucial pour la prise de décision industrielle. Contrôler ces incertitudes est un défi en ingénierie industrielle ayant à faire aux modèles numériques boîtes noires, mais aussi en intelligence artificielle.

Le travail présenté dans cette thèse s'intéresse principalement aux incertitudes entourant les modèles boîtes noires (numériques ou appris à partir de données). La section suivante vise à offrir une perspective sur le "quoi, pourquoi et comment" les incertitudes peuvent être prises en compte en ingénierie industrielle ainsi que dans le domaine de l'apprentissage automatique.

[1] https://www.edf.fr/groupe-edf/inventer-lavenir-de-lenergie/rd-un-savoir-faire-mondial/nos-offres/nos-logiciels-et-codes-de-calcul

G.2 Analyse de sensibilité et interprétabilité post-hoc

Un parallèle est établi entre deux domaines des mathématiques appliquées : la *quantification des incertitudes* (UQ) et *l'apprentissage automatique* (machine learning). Tandis que le premier est lié à l'étude des incertitudes propagées dans les modèles numériques, le second peut être vu comme une hybridation entre l'apprentissage statistique et les sciences informatiques. Bien que leurs objectifs peuvent être fondamentalement différents, ils partagent nombre de similitudes lorsqu'il s'agit de d'interpréter une modélisation (numérique ou apprise). En particulier, de nombreux objectifs de l'*analyse de sensibilité* (SA) et de l'*interprétabilité post-hoc*, un sous-domaine de l'intelligence artificielle explicable (XAI) [182, 16], sont partagés, comme décrit ci-dessous.

G.2.1 Analyse de sensibilité

En s'inspirant de [182], l'analyse de sensibilité peut être résumé ainsi :

> «L'étude de la manière dont les sorties d'un système sont liées et influencées par ses entrées.»

Historiquement, l'analyse de sensibilité fait partie de la méthodologie UQ [57], dont l'objectif principal est d'extraire des enseignements à partir de modèles informatiques boîte noire. Ces modèles sont souvent spécifiés pour simuler des phénomènes physiques, tels que la modélisation thermo-mécanique pour l'analyse structurelle des processus de fabrication[2], ou l'évaluation de la sécurité des installations industrielles[3]. Ils sont souvent composés d'une série d'opérations mathématiques complexes (par exemple, les solveurs d'équations différentielles, les modèles à éléments finis) conçues par des experts du domaine pour approcher au mieux le comportement des phénomènes physiques. Ces modèles sont cruciaux dans les études industrielles car ils offrent une alternative moins coûteuse, plus sûre et complémentaire aux expériences répétées.

Ces modèles numériques peuvent être considérés comme des systèmes entrée-sortie, où les entrées représentent des conditions initiales, ou des quantités physiques associées (par exemple, la température ambiante, la pression, l'humidité). Dans la méthodologie UQ, ces entrées sont considérées comme *incertaines*, soit en raison d'un manque de connaissance réductible (i.e., incertitude épistémique) soit en raison d'incertitudes contrôlées (par exemple, les erreurs de mesure).

Les incertitudes sont identifiées et quantifiées, et les entrées sont ensuite dotées d'une structure probabiliste (par dires d'experts du domaine ou par observation du phénomène). Les sorties du système deviennent alors également aléatoires : c'est ce que l'on appelle communément l'étape de *propagation des incertitudes* dans la méthodologie UQ.

C'est là que l'analyse de sensibilité entre en jeu. L'analyse de sensibilité cherche à mettre en lien des entrées aléatoires avec une sortie aléatoire d'un modèle. En particulier, quatre enjeux sont d'intérêt [48] :

- **Exploration du modèle :** exploration de la relation entrée-sortie dans le contexte incertain pour mieux comprendre le comportement du modèle ;
- **Détection d'entrée à effets négligeables :** détection des entrées non importantes (i.e., dont les incertitudes ont un impact limité sur l'incertitude de la sortie) pour les exclure de l'étude d'incertitude (en les considérant comme constantes) ;
- **Priorisation :** identification des entrées les plus importantes, i.e., celles dont l'incertitude affecte le plus l'incertitude de la sortie (ou d'une quantité d'intérêt) ;
- **Robustesse à la distribution des entrées :** étude des variations de la distribution de la sortie sortie (ou d'une quantité d'intérêt) par rapport aux changements dans la structure probabiliste choisie pour les entrées.

Ces enjeux peuvent être abordés soit d'un point de vue local (i.e., au voisinage d'une valeur d'entrée particulière), soit d'un point de vue global (i.e., sur l'ensemble du domaine des entrées) [158]. L'analyse

[2] par exemple, l'utilisation du code informatique `code_aster` pour la fabrication additive [101].
[3] par exemple, l'utilisation du code numérique `CATHARE2` pour les incidents de perte de réfrigérant dans les centrales nucléaires [3].

de sensibilité fournit de nombreuses méthodes statistiques pour proposer des outils pratiques permettant de répondre à ces quatres grandes questions [120, 30, 48]. Ces outils fournissent des *diagnostics* au praticien. Ces diagnostics peuvent être compris comme des estimations des quantités que la méthode d'analyse de sensibilité vise à quantifier. Ils sont une aide à la découverte scientifique (par exemple, l'amélioration de la compréhension des phénomènes étudiés) ou aux applications industrielles (par exemple, le soutien aux processus décisionnels). La Figure G.1 illustre la manière et la place des analyses de sensibilité pour tirer des enseignements des modèles numériques.

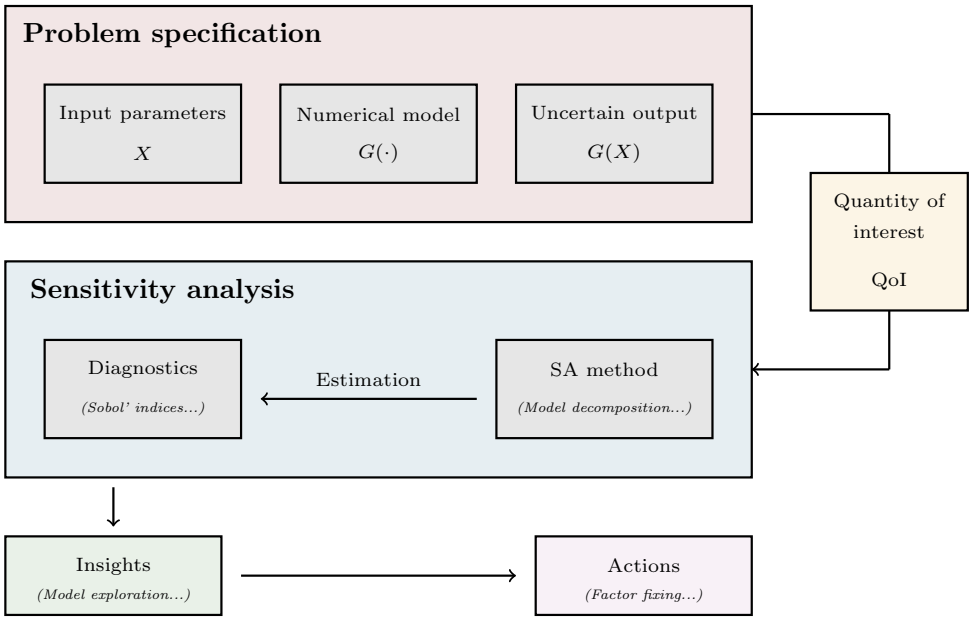

Figure G.1: Analyse de sensibilité pour mieux comprendre les modèles numériques.

G.2.2 Interprétabilité post-hoc

Selon [16], l'interprétabilité post-hoc peut être résumée comme suit :

«La capacité à expliquer et à fournir des raisons pour le comportement d'un modèle d'apprentissage automatique donné.»

L'objectif de l'apprentissage automatique est de proposer des outils pour modéliser divers phénomènes à partir de données observées. L'objectif est de produire un modèle capable d'ajuster au mieux la sortie en fonction des entrées, permettant d'approcher le phénomène observé en se basant sur un ensemble d'observations de variables d'entrée (i.e., caractéristiques) et de variables de sortie (i.e., cibles) formant un *jeu de données* observé. D'un point de vue statistique, les variables d'entrée et de sortie sont supposées être aléatoires, et le jeu de données est composé de réalisations de ces variables aléatoires [214]. Cependant, la structure probabiliste génératrice des entrées et des sorties est souvent inconnue et seulement observée. La véritable relation liant les entrées aux sorties est également inconnue. La conception d'un modèle d'apprentissage suppose que cette relation peut être approchée à l'aide d'un modèle appartenant à une certaine famille (par exemple, un modèle linéaire, un processus auto-régressif, un réseau de neurones). Cette famille est souvent caractérisée par des paramètres (par exemple, coefficients linéaires, coefficients auto-régressifs, poids et biais des neurones). Le processus d'apprentissage peut être décrit comme l'exploitation de données pour trouver les meilleures valeurs pour ces paramètres, dans le sens où elles minimisent une erreur (empirique) entre les valeurs cibles observées et les valeurs prévues par le modèle [97].

L'apprentissage automatique peut être utilisé pour deux objectifs connexes, mais fondamentalement différents :

- **Exploration de la relation entrée-sortie :** déterminer s'il existe une relation significative entre l'entrée et la sortie et, le cas échéant, sa nature (par exemple, linéaire, non linéaire) ;

- **Performance prévisionnelle :** construire le modèle le plus performant permettant d'accomplir une tâche de prévision spécifique avec une grande précision.

Dans le domaine de la statistique, les modèles supervisés ont longtemps été considérés comme un outil pour étudier des liens multivariés, constituant une avancée par rapport aux études statistiques univariées et bivariées traditionnelles [222]. La modélisation supervisée a, en particulier, permis de nombreuses avancées dans plusieurs domaines de recherche appliqués (par exemple, l'économie [7], la biologie [153], la médecine [43], les processus industriels [132]). L'objectif principal de ces études statistiques était alors l'exploration. Cependant, ces méthodes permettent également, par processus d'approximation, de répondre à des problématiques de prévision. La recherche de performance prédictive connaît un essor [124], notamment avec l'introduction d'approches d'apprentissage profond [90], qui obtiennent des scores de prévision presque parfaits sur des tâches très complexes (par exemple, la reconnaissance de chiffres [139], la classification d'images [130]).

A mesure que la nature des tâches de prévision devient de plus en plus diversifiée, la complexité intrinsèque des modèles les plus performants augmente, notamment car ils sont dotés d'un nombre important de paramètres. Ces modèles à haute performance sont considérés comme des boîtes noires. Bien que les aspects théoriques du processus d'apprentissage sont bien établis [97], la raison mathématique pour laquelle ces modèles surparamétrés montrent une performance aussi impressionnante est encore peu maîtrisée [163]. Il est facile de montrer qu'un modèle fonctionne, mais beaucoup plus difficile de comprendre pourquoi.

Malgré cette maîtrise théorique limitée, ces modèles restent attractifs pour la modélisation de systèmes critiques grâce à l'abondance de données récoltées ainsi qu'à l'efficacité croissante de la puissance de calcul disponible. Leur compréhension est cruciale pour leur adoption : le processus décisionnel doit reposer sur un raisonnement scientifique, dont la garantie est assurée par des études théoriques.

Le domaine de l'intelligence artificielle explicable (XAI) émane de ce besoin de mieux comprendre ces algorithmes boîte noire [16]. Ce nouveau champs d'étude englobe l'entièreté des aspects du processus d'explication de l'intelligence artificielle, du développement d'outils adaptés, à l'étude de l'interaction entre l'architecte de modèles et les experts du domaine. L'*interprétabilité post-hoc* fait partie du domaine de l'XAI. L'adjectif post-hoc fait référence au fait que le modèle d'apprentissage étudié est déjà entraîné : l'accent est mis sur la tentative d'extraire des enseignements sur le comportement d'un modèle spécifique (i.e., avec un ensemble fixe de paramètres) plutôt que de développer de nouvelles familles de modèles interprétables. Parmi les aspects auxquels l'interprétabilité post-hoc vise à répondre, on compte [16] :

- **La fiabilité :** la confiance quant à la manière dont un modèle agit face à un problème donné ;
- **La transférabilité :** élucidation des limites qui pourraient affecter un modèle, permettant une meilleure compréhension et mise en œuvre sur des données non observées ;
- **L'informativité :** extraction d'informations sur les relations internes d'un modèle ;
- **La confiance :** garantir la robustesse et la stabilité d'un modèle qui doit être fiable ;
- **L'équité :** évaluer si un modèle est influencé par des entrées protégées, qui pourraient entraîner à des traitements injustes ou discriminants.

Ces aspects peuvent être abordés soit d'un point de vue local (i.e., sur une instance de prévision particulière) soit d'un point de vue global (i.e., sur l'ensemble du domaine des entrées) [156]. De nombreuses méthodes ont été proposées dans la littérature [16]. Elles sont souvent justifiées de manière empirique, par le biais de benchmarks. La Figure G.2 présente une schématisation de la manière dont l'interprétabilité post-hoc peut être réalisée dans un processus de prise de décision.

L'interprétabilité post-hoc et l'analyse de sensibilité partagent de nombreux aspects. Leurs similitudes ont déjà été soulignées dans la littérature [182, 29, 142]. Le travail présenté dans cette thèse se situe à la croisée de ces deux domaines. Il poursuit l'objectif suivant : *développer des méthodes fondées sur des bases théoriques pour interpréter la modélisation par le biais de boîtes noires de systèmes critiques afin de justifier leur adoption dans la pratique*. Les méthodes proposées dans ces travaux sont *model-agnostic* (i.e., elles ne dépendent pas du type de modèle étudié). En effet, les boîtes noires rencontrées en ingénierie moderne peuvent prendre des formes différentes. Les méthodes proposées *s'inscrivent dans une démarche théoriquement fondée*, les preuves empiriques n'étant pas suffisantes pour garantir l'adoption des modèles boîtes noires pour la modélisation de systèmes critiques. Leur caractérisation repose sur des fondements

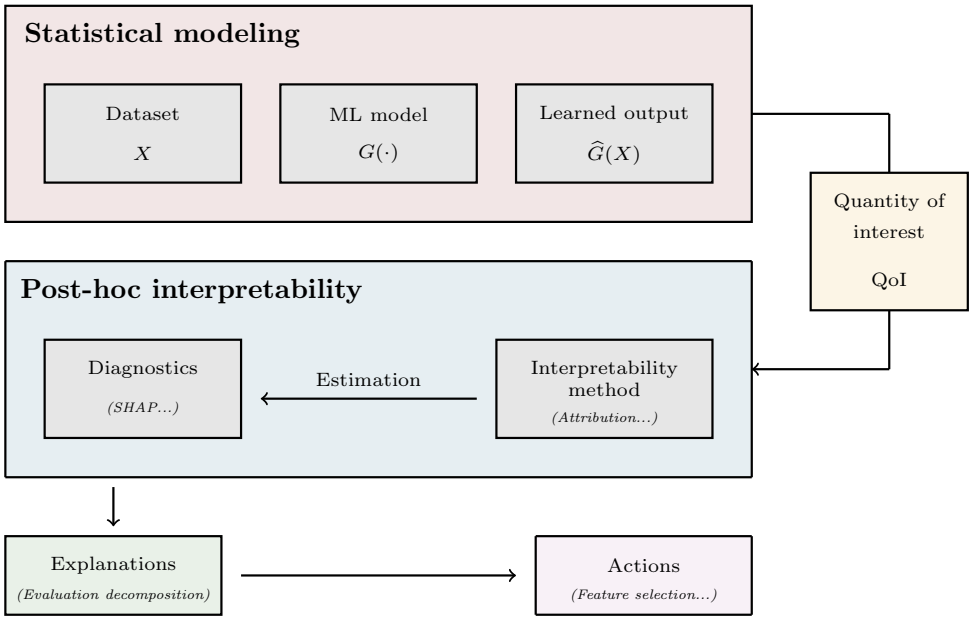

Figure G.2: Interprétabilité post-hoc pour mieux comprendre les modèles de machine learning.

théoriques solides, leurs propriétés sont étudiées, leurs limites sont mises en évidence, et la signification des connaissances qu'elles apportent est clairement énoncée et maîtrisée.

On adopte le cadre théorique de l'analyse de sensibilité comme point de départ, en raison de son succès historique dans la promotion de l'adoption de modèles numériques (boîtes noires) pour les études d'ingénierie. On propose un cadre mathématique permettant d'unifier entre l'analyse de sensibilité et l'interprétabilité post-hoc : *l'interprétabilité des modèles*. Pour répondre au mieux aux besoins pratiques en XAI, deux verrous doivent être levés : la prise en compte de la **dépendance entre les entrées** et le **processus de génération de données inconnu** (i.e., le praticien n'ayant généralement accès qu'à un ensemble de données observées). Ces deux contraintes sont au cœur des développements réalisés dans cette thèse.

G.3 Un cadre mathématique pour l'interprétabilité des modèles

Cette section présente le cadre mathématique de l'interprétabilité des modèles, ainsi que le premier ensemble de notations qui sont utilisées dans le manuscrit. Cette méthodologie probabiliste repose sur des notions de théorie de la mesure assez générales. On peut se référer à [122] et à l'appendice A pour quelques définitions préliminaires et résultats pertinents. Les éléments suivants sont introduits, définis et discutés :

- **Entrées aléatoires :** elles représentent les entrées incertaines des modèles numériques, ou les caractéristiques observées liées à un modèle d'apprentissage automatique. Dans ce cadre, les entrées aléatoires prennent la forme de vecteurs d'*éléments aléatoires* ;
- **Modèles boîtes noires :** ils représentent les boîtes noires utilisées pour modéliser des systèmes (potentiellement critiques). Ils peuvent représenter un modèle numérique d'un phénomène physique ou un modèle d'apprentissage automatique entraîné à partir de données. Dans ce cadre, les modèles boîtes noires prennent la forme de fonctions faisant correspondre deux espaces ;
- **Sortie aléatoire :** une sortie aléatoire est la composition d'un modèle boîte noire avec ses entrées aléatoires, devenant ainsi un élément aléatoire à valeur dans l'image (co-domaine) du modèle boîte noire (c.f., propagation des incertitudes en UQ) ;
- **Quantité d'intérêt :** elle représente une quantité significative liée à la sortie aléatoire (par exemple, une prévision d'un modèle, sa variance). Les quantités d'intérêt sont définies comme des

opérations entre le co-domaine du modèle boîte noire et un certain espace ;

- **Méthodes d'interprétabilité :** elles représentent des moyens de résoudre un *conundrum* clairement énoncé, i.e., une question pratique clé pour laquelle une réponse est attendue, afin de prendre une décision.

Dans le reste du manuscrit, les notations suivantes sont adoptées : $\subset$ indique une inclusion propre (stricte) entre deux ensembles, tandis que $\subseteq$ indique que l'égalité entre les deux ensemble est possible. De plus, pour un ensemble non-vide A, l'ensemble $\{B : B \subseteq A\}$ des sous-ensembles de A ne contient pas l'ensemble vide $\emptyset$.

G.3.1 Entrées aléatoires

Pour permettre le lien entre l'XAI et l'analyse de sensibilité, de nombreux types d'entrées doivent être considérés. Ces dernières ne sont pas nécessairement à valeur réelle (e.g., des données non tabulaires telles que le texte, les images et les séries temporelles). En s'inspirant de [48], les entrées aléatoires sont définies en utilisant la notion très générale d'*élément aléatoire* et de *vecteurs d'éléments aléatoires* (Appendix A), qui généralisent les variables aléatoires et de vecteurs aléatoires (qui sont intrinsèquement à valeur dans $\mathbb{R}$ ou $\mathbb{R}^d$).

Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace probabilisé abstrait, soit d un entier positif, et soit $(E_1, \mathcal{E}_1), \ldots, (E_d, \mathcal{E}_d)$ une collection d'espaces mesurables Boréliens. Pour tout $A \subset D := \{1, \ldots, d\}$, on note :

$$E_A := \bigtimes_{i \in A} E_i, \quad \mathcal{E}_A := \bigotimes_{i \in A} \mathcal{E}_i, \quad \text{and} \quad E := \bigtimes_{i \in D} E_i, \quad \mathcal{E} := \bigotimes_{i \in D} \mathcal{E}_i$$

où $\bigtimes$ désigne le produit cartésien entre ensembles et $\bigotimes$ désigne le produit de tribus ([150], Section 2.4.2). Il est intéressant de remarquer que pour tout $A \subset D$, $(E_A, \mathcal{E}_A)$ est également un espace mesurable Borélien standard et $(E, \mathcal{E})$ aussi ([126], Lemme 1.2).

Les *entrées aléatoires* sont représentées par une application a valeur dans E, et mesurable par rapport à $\mathcal{F}$, notée $X = (X_1, \ldots, X_d)^\top$ (i.e., un vecteur d'éléments aléatoires). Pour tout $A \subset D$, le *sous-ensemble d'entrées* relatif à A, i.e., le vecteur d'éléments aléatoires à valeur dans E_A, est défini par $X_A := (X_i)_{i \in A}$.

La *tribu engendrée par les entrées aléatoires* (Definition A.2) est notée σ_X, et pour tout $A \subset D$, la tribu engendrée par le sous-ensemble d'entrées X_A est notée σ_A. Ces tribus engendrées sont traditionnellement interprétées comme l'*information* apportée par un élément aléatoire.

La *distribution jointe des entrées* est la mesure de probabilité induite par l'application mesurable X (Definition A.5), notée P_X. Pour tout $A \subset D$, la *distribution marginale de l'ensemble d'entrées* X_A est la mesure de probabilité induite par l'application mesurable X_A, notée P_{X_A}.

G.3.2 Modèle boîte noire

Les modèles boîtes noires sont définis de manière abstraites, pour accommoder à la fois les modèles numériques et les modèles d'apprentissage automatique, et en particulier, la variété de sorties possibles (par exemple, les maillages, le texte, les sorties réelles ou binaires).

Soit $(Y, \mathcal{Y})$ un espace mesurable Borélien standard. Un *modèle boîte noire* est représenté par une application mesurable notée $G : E \to Y$.

G.3.3 Sortie aléatoire

De manière assez naturelle, et conformément à la *propagation des incertitudes*, la sortie aléatoire fait référence à la composition des entrées aléatoires et du modèle boîte noire. Elle peut être interprétée comme la représentation du modèle du système dans son ensemble, en prenant en compte toutes les incertitudes auxquelles il est soumis.

La sortie aléatoire est notée par la fonction mesurable $G(X) := G \circ X : \Omega \to Y$, i.e., un élément aléatoire à valeur dans Y. Il est important de noter que les sorties aléatoires sont nécessairement des fonctions mesurables par rapport à σ_X. On note $\mathcal{G}_X$ *l'espace des sorties aléatoires* défini comme :

$$\mathcal{G}_X = \{f : \Omega \to Y : f \text{ est } \sigma_X\text{-mesurable}\}.$$

De plus, pour tout $A \subseteq D$, on note $\mathcal{G}_A$ le sous-ensemble de $\mathcal{G}_X$ de fonctions à valeur dans Y, mesurables par rapport à σ_A (ici, $\sigma_D = \sigma_X$). On note également $\mathcal{G}_\emptyset$ l'espace des fonctions mesurables par rapport à la *la tribu* $\mathbb{P}$*-triviale* (Definition A.6), notée $\sigma_\emptyset$.

Figure G.3 illustre les relations entre les entrées aléatoires, le modèle boîte noire et la sortie aléatoire. Le cadre proposé s'appuie sur les *relations fonctionnelles* entre ces trois notions. Dans le contexte de l'interprétabilité des modèles boîtes noires, il est important de noter que le modèle est considérée comme étant *fixe*. L'objectif principal ne porte pas sur *la modélisation de phénomènes en utilisant des boîtes noires*, mais plutôt sur *l'extraction d'informations issues de ces boîtes noires après qu'elles aient été modélisées*.

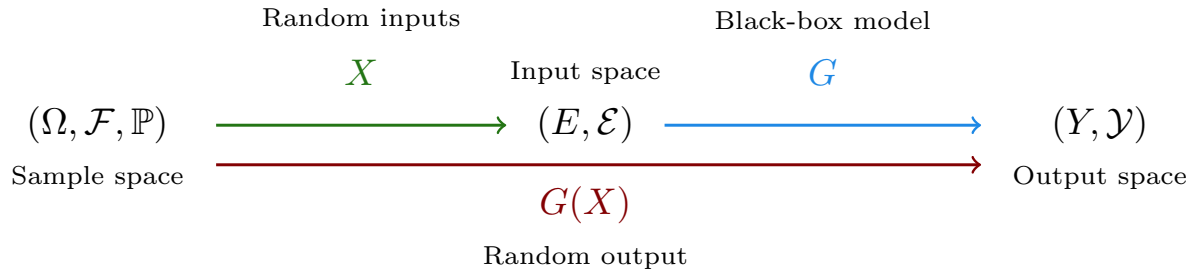

Figure G.3: Modélisation boîte noire : Relation entre les entrées et les sorties.

G.3.4 Quantité d'intérêt

Les QoIs sont primordiales dans le cadre de l'interprétabilité des modèles. Elles doivent avoir *un sens pratique pour les experts du domaine* : elles sont vectrices d'informations clés sur le *conundrum* (voir la section suivante) que l'étude d'interprétabilité cherche à traiter. Ces QoIs sont également considérées comme aléatoires, bien que dans la plupart des cas pratiques, elles sont déterministes. En résumé, cette définition des QoIs élargit la notion homonyme en analyse de sensibilité [48] pour prendre en compte différents types de situations.

Soit $(Q, \mathcal{Q})$ un espace mesurable, appelé l'*espace des QoIs*. Soit $\mathrm{QoI} : \mathcal{G}_X \to Q$ un opérateur. La QoI réfère à l'élément aléatoire résultant de sa composition avec la sortie aléatoire, i.e., $\mathrm{QoI}\,(G\,(X))$.

Par exemple, une QoI peut être la sortie aléatoire elle-même, une *évaluation* (observation) de la sortie aléatoire, i.e., , pour un $\omega \in \Omega$, la quantité $G(X(\omega))$ [149], dans le cas des sorties à valeur dans $\mathbb{R}$, son *espérance*, i.e., $\mathbb{E}\,[G(X)]$, sa *variance*, i.e., $\mathbb{V}\,(G(X))$ [204], et dans le cas des sorties à valeur dans $\mathbb{R}^d$, sa matrice de covariance [85]. Il est important de noter que dans les deux derniers exemples, l'opérateur intégral est pris par rapport à la mesure de probabilité fixe $\mathbb{P}$ sur Ω.

G.4 Conundrums et méthodes d'interprétabilité

En général, les méthodes d'interprétabilité peuvent être comprises comme des *transformations significatives de la QoI*. Bien que cette définition ne soit pas très formelle, elle reste suffisamment générale pour englober le large éventail de méthodes proposées dans la littérature (par exemple, [16]). Une *transformation de la QoI* peut être comprise comme la réalisation d'une étude méthodologique de la QoI. Le terme *significatif* met l'accent sur l'objectif de la méthode d'interprétabilité, qui est incarné par un *conundrum* [9], i.e., une question précise de compréhension reliée à un modèle boîte noire. Une (ou plusieurs) méthode d'interprétabilité est choisie pour résoudre un conundrum, i.e., définir les quantités théoriques pertinentes pour répondre à la question. L'estimation de ces quantités conduit à des *diagnostics*, qui peuvent ensuite être interprétés.

G.4.1 Conundrums

Les conundrums se matérialisent sous forme de *questions pratiques*. Dans la modélisation abstraite des *jeux d'explication* [9], deux joueurs, un *explicant* (par exemple, un expert du domaine, un ingénieur) et un *expliqué* (par exemple, un décideur, une autorité de régulation), interagissent afin de résoudre un *conundrum*, i.e., une question émanant de l'expliqué, dont la réponse est donnée par l'explicant. Pour résoudre

un conundrum, l'explicant doit fournir une *explication* à l'expliqué, qui décide alors si l'explication est *suffisante*. L'étude des interactions entre l'explicant et l'expliqué se situe à la croisée de nombreuses disciplines scientifiques telles que la logique, la théorie des jeux non coopératifs, la psychologie, l'ergonomie et la microéconomie. Ces domaines font partie intégrante de l'XAI pour étudier les aspects sociaux de *l'acceptation de l'intelligence artificielle* [16].

Par exemple, les conundrums peuvent prendre la forme de ces questions:

- *Pourquoi le modèle fournit-il cette prévision précise pour ce point de données d'entrée précis ?*
- *Quelles entrées sont responsables de l'incertitude du système modélisé ?*
- *Quel est l'impact du manque de connaissance des entrées sur le système modélisé ?*

Le travail présenté dans cette thèse se concentre sur un aspect de ce processus d'interaction complexe : fournir des outils à l'explicant pour construire des explications pertinentes pour certains conundrums spécifiques. Afin d'assurer la pertinence des informations fournies par ces outils, ils doivent être maîtrisés théoriquement. Ces outils sont appelés *méthodes d'interprétabilité*, et sont introduits ci-dessous.

G.4.2 Méthodes d'interprétabilité

Dès lors qu'un conundrum est établi, la première étape consiste à identifier des QoIs pertinentes. Ce sont des indicateurs clés, relatifs au conundrum en jeu. Les méthodes d'interprétabilité consistent à trouver une méthodologie pour résoudre le conundrum en étudiant les QoIs choisies. Par exemple, si la question principale porte sur les raisons derrière une prévision du modèle, une QoI appropriée peut être la prévision elle-même, en utilisant des méthodes adaptées axées sur la causalité (par exemple, les méthodes contrefactuelles [159, 9]) ou des approches d'extraction de règles [18]. Si le conundrum concerne l'identification des entrées qui affectent le plus une QoI, les méthodes de décomposition peuvent être considérées (par exemple, les décompositions coalitionnelles [111], les méthodes d'attribution [149]). Si l'expliqué s'interroge sur le comportement du modèle pour des données en dehors de la distribution initiale des entrées, les méthodes de perturbation d'entrées peuvent être pertinentes (par exemple, projections de mesure de probabilité [141, 13, 113], perturbations basées sur la géométrie de l'information [86, 127]).

De nombreuses méthodes d'interprétabilité ont été proposées dans la littérature de l'XAI, par exemple, se référer à [16, 143, 207] pour un aperçu et une taxonomie de ces méthodes. Trois contraintes sont introduites dans les développements présentés dans cette thèse pour répondre à l'objectif principal de l'étude des modélisations de systèmes critiques :

- **Pertinence** : La capacité des méthode d'interprétabilité choisies à répondre à un conundrum doit être justifiée ;
- **Fondement théorique** : Les méthodes d'interprétabilité choisies doivent être basée sur un cadre théorique solide, avec des hypothèses clairement énoncées, leurs propriétés étudiées et leurs limites mises en évidence ;
- **Cohérence pratique** : Outre la théorie, les méthode d'interprétabilité choisies doivent être étudiées empiriquement et validées sur des cas d'études.

G.5 Deux méthodes d'interprétabilité

Deux méthodes d'interprétabilité sont présentées dans cette thèse :

- **Les décompositions de QoI** : L'étude de la manière dont les QoIs peuvent être décomposées. Cette méthode est particulièrement adaptée pour les conundrums liés à la quantification de l'influence ;
- **Les perturbations des entrées** : L'étude de la manière dont la distribution des entrées peut être perturbée, ainsi que la propagation de ces perturbations sur le modèle étudié. Cette méthode propose de répondre à certains conundrums relatifs à la robustesse des modèles boîtes noires.

G.5.1 Décomposition de QoI pour évaluer l'influence

Les méthodes de décomposition de QoI, impliquent de pouvoir écrire $\mathrm{QoI}\left(G\left(X\right)\right)$ comme une somme d'éléments de l'espace Q, à condition que la QoI soit dotée d'une opération d'addition appropriée (i.e., $(Q,+)$ forme groupe abélien, voir Appendix B). Par exemple, les *méthodes d'attribution additives* [149] étudient les décompositions suivantes :

$$\mathrm{QoI}\left(G\left(X\right)\right) = \phi_{\emptyset} + \sum_{i \in D} \phi_i,$$

où pour chaque $i \in D$, $\phi_i \in Q$, et où chaque ϕ_i correspond à un *effet de l'entrée* X_i. Les *décompositions coalitionnelles de QoI* diffèrent des méthodes d'attribution car la somme est prise sur l'ensemble des *parties de* D (i.e., l'ensemble des sous-ensembles de D, y compris $\emptyset$), noté $\mathcal{P}_D$. Une décomposition coalitionnelle de $\mathrm{QoI}\left(G\left(X\right)\right)$ consiste en la somme :

$$\mathrm{QoI}\left(G\left(X\right)\right) = \sum_{A \in \mathcal{P}_D} \phi_A,$$

où pour chaque $A \in \mathcal{P}_D$, $\phi_A \in Q$, et où chaque ϕ_A correspond à un *effet du sous-ensemble d'entrées* X_A. Le terme coalition vient du fait que les effets des *coalitions* (i.e., des sous-ensembles) d'entrées sont pris en compte dans la décomposition, contrairement aux méthodes d'attribution qui se concentrent uniquement sur les effets *individuels* (i.e., univariés).

Le paradigme principal derrière l'idée de la décomposition de QoI repose sur le fait que les éléments qui forment les décompositions doivent contenir une certaine information sur *l'influence qu'une entrée (ou un sous-ensemble des entrées) peut avoir sur la QoI*. Si Q est doté d'un ordre total naturel, la comparaison des magnitudes de ces effets peut exprimer un éventuel classement d'influence sur les entrées.

Dans la littérature, de nombreuses techniques reposant sur ces méthodes ont été proposées. Des exemples de méthodes d'attribution pour la décomposition d'évaluation des modèles de régression (i.e., $Y = \mathbb{R}$, et $\mathrm{QoI}\left(G\left(X\right)\right)\left(G(X)\right) = G(X(\omega))$ pour certains $\omega \in \Omega$) sont LIME [186] ou SHAP [149]. Les effets Shapley [169] ou les effets marginaux proportionnels [100] peuvent être utilisés pour la décomposition de la variance de la sortie (dans ce cas, $Y = \mathbb{R}$ et $\mathrm{QoI}\left(G\left(X\right)\right) = \mathbb{V}\left(G(X)\right)$). Les indices de Sobol' peuvent être vus comme une décomposition coalitionnelle de la variance [204], qui sont au coeur du domaine de l'analyse de sensibilité.

G.5.2 Perturbations des entrées pour l'évaluation de la robustesse des modèles

Les méthodes de perturbation des entrées consistent à modifier, de manière contrôlée, la distribution des entrées. Une fois que la distribution modifiée est obtenue, ces méthodes permettent d'étudier le comportement de QoIs *sous la distribution perturbée*. Formellement, pour des entrées aléatoires initiales X et des entrées perturbées $\widetilde{X}$, cela revient à comparer $\mathrm{QoI}\left(G\left(X\right)\right)$ et $\mathrm{QoI}(G(\widetilde{X}))$, et mettre en évidence leurs différences en fonction de la nature de la perturbation. Étudier ces différences permet ainsi d'évaluer le comportement d'un modèle (à travers ses QoIs) via une structure probabiliste de ses entrées différente de celle initialement prévue. L'étude de la différence de comportement du modèle permet ensuite *d'évaluer sa robustesse une fois soumis à ces perturbations*. Cette méthode d'interprétabilité peut être utilisée pour des études prospectives, des analyses exploratoires, ou pour garantir la cohérence du modèle avec les connaissances des experts du domaine.

Formellement, soit $\mathcal{C}$ une *classe de perturbation*, i.e., un ensemble particulier de mesures de probabilité induites par des entrées aléatoires à valeur dans E, et $\mathcal{D}$ est une *mesure de discrépance* entre mesures de probabilité (i.e., ce ne doit pas nécessairement être une distance). Le *problème de perturbation* peut être formulé comme le problème d'optimisation sous contrainte suivant :

$$\begin{aligned} P_{\widetilde{X}} \in \underset{P}{\operatorname{argmin}} \quad & \mathcal{D}\left(P_X, P\right) \\ \text{s.t.} \quad & P \in \mathcal{C}. \end{aligned}$$

Dans la littérature, plusieurs choix de discrépances et de classes de perturbations ont été étudiés. En s'appuyant sur le travail pionnier de [47] sur les projections entropiques, le choix de la divergence de Kullback-Leibler (KL) a été étudié par [141] en analyse de sensibilité et par [13] en XAI, où $\mathcal{C}$ est défini via des contraintes sur les moments généralisés de la distribution des entrées. Dans [86, 127], les auteurs

proposent d'étudier des familles paramétriques de distributions, où la disparité est choisie par le biais de la métrique de Fisher sur l'espace des paramètres de familles de densités de probabilité paramétrées, conduisant à des classes de perturbations naturelles composées de séquences de distributions perturbées le long de géodésiques. Dans [113], le choix de la distance de Wasserstein est motivé conjointement avec des classes de perturbations préservant les copules dont les contraintes sont formulées sur les quantiles des entrées.

G.6 Articulation du manuscrit

L'objectif général de ce manuscrit est d'explorer le concept de l'interprétabilité des modèles boîtes noires et de proposer une première approche d'un cadre mathématique. Ce cadre permet de justifier l'utilisation et de guider le développement des méthodes d'interprétabilité afin d'améliorer la fiabilité des modèles boîtes noires des systèmes critiques et de tendre vers leur acceptation par les instances réglementaires.

Dans le **Chapitre** 2, les origines algébriques de la question fondamentale de la mesure de l'influence sont mises en évidence. Ce lien découle de l'hypothèse que les sous-ensembles d'entrées peuvent être classés par rapport à leur influence. Cela ouvre naturellement la voie à l'étude des décompositions coalitionnelles de QoIs, afin de produire des mesures d'influence dont l'objectif principal est d'exprimer cet ordre qui supposé exister. Ces mesures d'influences sont définies, de manière intrinsèque, comme étant des fonctions ensemblistes, dont le domaine est un ensemble puissance (*power-set*). De fait, le lien avec le domaine de la combinatoire est plutôt direct. En particulier, la généralisation par Rota de la formule d'inversion de Möbius, lorsqu'elle est appliquée aux power-set, permet de définir deux approches pour définir des mesures d'influence : l'approche input-centric (centrée sur les entrées), qui nécessite la définition d'une mesure de valeur qui quantifie l'influence totale de chaque sous-ensembles d'entrées, et l'approche model-centric (centrée sur le modèle), qui nécessite une décomposition intrinsèque de la sortie aléatoire. Ces deux approches sont illustrées par le problème de quantification de l'importance, i.e., la décomposition de la variance de la sortie aléatoire.

Le **Chapitre** 3 plonge plus profondément dans l'approche input-centric. Cette question a été étudiée sous le paradigme de la théorie des jeux coopératifs, par analogie entre les joueurs et les entrées d'un modèle boîte noire. Cette analogie permet de produire des allocations, qui sont généralement construites sur une mesure d'influence input-centric, connue sous le nom de dividendes de Harsanyi. Dans ce cadre, des attributions d'importance peuvent être définies (i.e., qui décomposent la variance entre les entrées elles-mêmes, plutôt qu'entre chaque sous-ensemble d'entrées), telles que la redistribution égalitaire des dividendes proposée par les effets de Shapley. Ces derniers présentent un inconvénient : les entrées qui ne sont pas dans le modèle, mais corrélées avec des entrées dans le modèle, peuvent se voir accorder une certaine importance. Ce problème est résolu avec les effets proportionnels marginaux, qui reposent sur une redistribution proportionnelle des dividendes. Ces deux méthodes offrent deux façons différentes de quantifier l'importance. Elles sont comparées et illustrées sur des cas d'études. Enfin, le choix de la fonction de valeur est discuté, car celui-ci reste arbitraire.

Dans le **Chapitre** 4, l'approche model-centric est abordée, en étudiant le problème de la décomposition de la sortie aléatoire. Ce problème possède déjà une solution lorsque les entrées sont mutuellement indépendantes, connue sous le nom de décomposition de Hoeffding. Bien que de nombreux développements aient été proposés dans la littérature, aucune réponse définitive n'a été proposée concernant sa généralisation potentielle pour des entrées dépendantes sous des hypothèses peu restrictives. Aborder cette généralisation comme un problème de décomposition en somme directe d'espaces de Lebesgue (i.e., espaces de Hilbert), permet de montrer que cette décomposition reste vraie pour des entrées dépendantes, sous deux hypothèses raisonnables : l'absence de dépendance fonctionnelle parfaite et la non dégénérescence de la dépendance stochastique. Ces développements découlent de l'étude des sous-espaces intrinsèques des espaces de Lebesgue générés par chaque sous-ensemble des entrées. En particulier, l'étude des relations entre ces espaces se fait par le biais des angles de Dixmier et Friedrichs. Le résultat obtenu est intuitif et repose sur des considérations géométriques, et en particulier sur la notion de projection oblique. Enfin, cette approche permet de définir plusieurs mesures d'influence, qui peuvent être justifiées théoriquement, et dont les propriétés sont présentées. Cette décomposition est illustrée au moyen d'un cas d'étude analytique.

Enfin, le **Chapitre** 5 est consacré à l'étude du problème de l'évaluation de la robustesse des modèles boîtes noires. En particulier, l'étude du comportement d'un modèle lorsque la distribution de ses en-

trées est perturbée. Une vision formalisée et générale de ce problème est proposée, en le modélisant comme un problème d'optimisation sur des espaces de mesures de probabilité. Le processus de définition des perturbations est discuté, et quatre critères de sont proposés. Sur la base de ces critères, on explore le choix de la distance de Wasserstein comme moyen de comparer les mesures de probabilité, et les perturbations basées sur les quantiles, tout en préservant la structure de dépendance initiale. Le problème peut être résolu analytiquement, mais cette solution n'est pas adaptée aux études pratiques. Des contraintes de régularité sont introduites, en plus des perturbations des quantiles et de la préservation de la dépendance. L'utilisation de polynômes interpolants isotoniques est étudiée, ce qui conduit à un problème d'optimisation avec une solution unique. La méthode présentée est illustrée et discutée sur des cas d'études. Elle ouvre la voie à de nouvelles méthodes de validation des modèles d'apprentissage, allant au-delà des métriques classiques.

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

Université de Toulouse

Titre : Développement de méthodes d'interprétabilité en apprentissage automatique pour la certification des intelligences artificielles reliées aux systèmes critiques

Mots clés : Interprétabilité, Analyse de sensibilité, Apprentissage automatique

Résumé : Les algorithmes d'apprentissage automatique, qui ont énormément contribué à l'essor de l'intelligence artificielle (IA) moderne, ont démontré à maintes reprises leur haute performance pour la prévision de tâches complexes. Cependant, malgré le gain manifeste évident lié à l'utilisation de ces méthodes pour l'accélération et l'amélioration de la performance de tâches d'ingénierie variées (mise en relation d'informations collectées par des capteurs, détection de signaux rares, etc.), incluant en particulier la modélisation de systèmes critiques industriels (temps de calcul, valorisation de données récoltées, hybridation entre la physique et les données expérimentales), la modélisation par apprentissage automatique n'est toujours pas largement adoptée dans les pratiques d'ingénierie moderne. Les résultats empiriques des modèles appris sur certains jeux de données (benchmarks) ne suffisent pas à convaincre les instances de sûreté et de contrôle en charge des activités industrielles. Cette thèse a pour but de développer des méthodes permettant la validation de l'usage de modèles boîtes-noires (dont les IA) par le biais de l'étude des incertitudes. Un formalisme mathématique global est proposé pour l'étude théorique des méthodes d'interprétabilité des modèles boîtes noires. Ce travail méthodologique permet de rapprocher deux domaines très proches : l'analyse de sensibilité (SA) des modèles numériques et l'interprétabilité post-hoc. Deux thématiques concrètes sont au cœur des travaux de cette thèse : la quantification d'influence et l'étude de robustesse face aux perturbations probabilistes. Une attention particulière est portée au cadre et aux propriétés théoriques des méthodes proposées dans le but d'offrir des outils convaincants allant au-delà des considérations empiriques. Des illustrations de leur utilisation, sur des cas d'études issues de problématiques réelles, permettent d'étayer la cohérence de leur utilisation en pratique. La situation d'entrées dépendantes, c'est-à-dire lorsque les entrées du modèles boîte-noire ne sont pas supposées mutuellement indépendantes, prennent une place importante dans les travaux menés. Cette considération a permis la généralisation de méthodes existantes en SA et en intelligence artificielle explicable (XAI). Au-delà de leurs propriétés théoriques pertinentes, ces nouvelles méthodes sont davantage cohérentes avec les études pratiques, où les données récoltées sont souvent corrélées. En particulier, un stratagème de perturbation probabiliste conservant cette dépendance fondé sur des méthodes de transport optimal est proposé. De plus, une généralisation sous des hypothèses peu restrictives de la décomposition fonctionnelle d'Hoeffding est également démontrée. Elle permet d'étendre à un contexte non mutuellement indépendant une multitude de méthodes déjà existantes et utilisées en pratique. Les travaux présentés sont en lien étroit avec différents domaines mathématiques : statistiques, probabilités, combinatoire algébrique, optimisation, transport optimal, analyse fonctionnelle et théorie des jeux coopératifs. Plusieurs liens entre ces disciplines sont établis afin d'offrir une vision générale de l'étude d'interprétabilité des modèles boîtes-noires.

Title: Development of interpretability methods for certifying machine learning models applied to critical systems

Key words: Interpretability, Sensitivity Analysis, Machine Learning

Abstract: Machine learning algorithms, which have significantly contributed to modern artificial intelligence (AI) advancement, have repeatedly demonstrated their performance in predicting complex tasks. However, despite the potential benefits of using these methods for modeling critical industrial systems (computation time, data value, hybridization between physics and experimental data), these algorithms have not yet been widely adopted in modern engineering practices. Empirical results on benchmark datasets do not seem sufficient to convince safety and control authorities responsible for industrial activities. This thesis aims to develop methods for validating the use of black-box models (particularly those embedded in AI systems) through the study of uncertainties. A general and comprehensive mathematical formalism is proposed for the theoretical study of black-box model interpretability methods. This methodological work unifies two closely related research areas: sensitivity analysis (SA) of numerical models and post-hoc interpretability. Two central themes to this thesis are influence quantification and robustness to probabilistic perturbations. Special attention is paid to the framework and theoretical properties of the proposed methods to provide compelling tools that go beyond empirical considerations. Illustrations of their use on real-world problem cases support the consistency of their practical use. The consideration of dependent inputs, i.e., when the inputs of the black-box models are not assumed to be mutually independent, plays a significant role in the research conducted. This consideration has allowed the generalization of existing methods in SA and explainable artificial intelligence (XAI). Beyond their relevant theoretical properties, these new methods are more consistent with practical studies, where collected data is often correlated. In particular, a probabilistic perturbation strategy that preserves this dependence based on optimal transport methods is proposed. Furthermore, a generalization under non-mutually independent assumptions of the Hoeffding functional decomposition is also demonstrated. It allows the extension of a multitude of existing methods used in practice. The presented work is closely related to various mathematical domains: statistics, probability, algebraic combinatorics, optimization, optimal transport, functional analysis, and cooperative game theory. Several connections between these disciplines are established to provide a global and comprehensive view of black-box model interpretability research.