Sensitivity analysis and algorithmic fairness for machine learning and artificial intelligence

Clément Benesse

THÈSE

En vue de l'obtention du
DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par l'Université Toulouse 3 - Paul Sabatier

Présentée et soutenue par

Clément BENESSE

Le 16 décembre 2022

Analyse de sensibilité et Équité Algorithmique pour le Machine Learning et l'Intelligence Artificielle

---

Ecole 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
Fabrice GAMBOA et Jean-Michel LOUBES

Jury

M. Sébastien GAMBS, Rapporteur
Mme Mathilde MOUGEOT, Rapporteure
Mme Béatrice LAURENT-BONNEAU, Examinatrice
M. Erwan SCORNET, Examinateur
M. Fabrice GAMBOA, Directeur de thèse
M. Jean-Michel LOUBES, Co-directeur de thèse

On the links between Global Sensitivity Analysis and Algorithmic Fairness for eXplainable and Fair Machine Learning

Clément Bénesse

2019 – 2022

Remerciements

S'il n'est que quelques mots, en ce lieu déposés,
Auxquels vous devriez prêter attention,
Ceux-ci vont pour nombre, j'en suis sûr, s'avérer
Parmi les plus pertinents et les plus aimés.

En effet, admirez l'avènement du temps
Non de quelque rictus, de je-ne-sais quels chants
Mais bien des louanges, autres remerciements
qui ne seront pour vous qu'éloges et encens!

Alors souffrez là cette tirade anormale!
Car oui! Je nourris en mon for intérieur
L'espoir que du souvenir de cette clameur,
Vous n'aurez pas moins qu'une pensée amicale!

Comment commencer autrement de tels propos
Que par la famille? Elle a toujours été là:
Lorsque tout va bien comme lorsque rien ne va.
Peu importe la suite, elle continuera.

Continuons par ceux qui ont choisi de gré
De partager leur quotidien à mes côtés
Sébastien, Athé, la Bande des Lyonnais
Jamais ne pourrais-je assez vous remercier.

On continue, viennent là tous les toulousains,
Autres taroteurs, cobureaux, le beau gratin!,
Ceux qui ont rendu ma vie dans la Belle Rose
Un plaisir, je remercie jusqu'à l'overdose!

Enfin, je terminerai ces quelques pensées
Par ceux qui m'ont accompagné ces années tierces
Merci de m'avoir montré ce qu'est la recherche:
Un meilleur sujet, je n'aurai pu espérer.

Quant à celles et ceux dont j'ai pu faire fi
N'y voyez pas là le signe d'une infamie
Mais simplement un malheureux, facheux oubli.
Vous êtes tout de même dans ma vie. <3

Contents

Introduction générale

Ce manuscrit est composé de cinq chapitres dont le détail est donné ci-dessous, ainsi que d'annexes contenant les preuves ou certains résultats techniques. Il est fortement recommandé de parcourir dans l'ordre des chapitres présentés.

- Le premier chapitre (en français) a pour but de donner un aperçu général des diverses notions qui seront abordées tout au long de cet écrit. Nous verrons ainsi que ce doctorat s'ancre dans une dynamique sociale récente et de grande ampleur, exhibant le lien existant entre Explicabilité et Équité Algorithmique. Un état de l'art de chacun de ces domaines est donné de manière indépendante, quand bien même la structure laisse deviner le sujet du chapitre suivant.

- Dans le deuxième chapitre (en anglais), nous formalisons un certain nombre de définitions. Puis, nous explicitons le lien qui existe entre l'Équité de Groupe et l'Analyse de Sensibilité Globale. Ce chapitre correspond à l'article [Bén+21], publié dans l'édition spéciale du journal *Machine Learning* consacrée à l'étude de l'Équité Algorithmique.

- Le troisième chapitre (en anglais) est consacré à l'étude de métriques, définies auparavant, dans le cas d'utilisation de métamodèles. Il étend un certain nombre de résultats précédemment présents dans la littérature et fournit une trame générale pour la résolution d'un type de problèmes récent que sont les audits.

- Le quatrième chapitre (en anglais) regroupe deux résultats indépendants l'un de l'autre. Tout d'abord seront développées les propriétés statistiques d'un nouvel estimateur des indices de Sobol' – brique élémentaire utilisée tout au long de ce manuscrit. Puis, nous évoquerons le sujet de la robustesse distributionnelle dans le cas de l'Analyse de Sensibilité Globale puis de la traduction de cette notion dans un cadre d'Équité Algorithmique

- Enfin, le cinquième chapitre (en français) récapitule le propos des chapitres précédents et fournit des pistes et perspectives pour des travaux futurs.

Chapter 1

Une histoire d'influences

1.1 Motivations historiques, sociales et juridiques

Au cours des dernières années, le déploiement d'algorithmes utilisant du *Machine Learning* (*ML*) a explosé. Il est désormais presque impossible de trouver un pan de la société qui n'ait été touché par cette vague, un domaine qui n'ait vu l'arrivée de l'*Intelligence Artificielle* (*IA* en français, *Artificial Intelligence* ou *AI* en anglais). Que ce soit dans nos objets connectés de la vie courante ou dans les voitures pour les applications civiles, dans les usines ou en phase de conception pour l'industrie ou encore dans le traitement de données et la création de plans d'expérience dans la recherche, ce raz-de-marée technologique a changé nombre de paradigmes en l'espace de deux décennies, amenant par le biais du *Big Data* ce qui est désormais considéré par beaucoup de spécialistes comme la Quatrième Révolution Industrielle. La promesse est la suivante: les algorithmes automatiseront la plupart des tâches et, pour peu que l'on ait assez de données, feront bien mieux que l'être humain.

Ce changement de système, passant d'une prise de décision humaine à une automatisation algorithmique, commence néanmoins à montrer ses limites. L'image de cette technologie en apparence omnipotente et omniprésente commence à s'effriter et à montrer de nombreux problèmes socialement inacceptables. Que ce soit le coût écologique d'une technologie vorace et insatiable ; l'absence de contrôle combiné avec la confiance absolue de l'utilisateur en un système supposé «infaillible» ; ou encore les exemples de plus en plus nombreux de dérives systémiques, ce que l'on pourrait nommer de «contrat algorithmique» montre de manière criante ses faiblesses. Autant de limitations qui ont amené la communauté scientifique et la société civile à passer au crible les algorithmes, délaissant l'idéologie du «accuracy above all», cette course peu précautionneuse aux résultats au prix de plus de données. Désormais, l'incitation est à «remettre l'humain dans la boucle» («put humans back in the loop», en référence à la boucle algorithmique) dans l'espoir de récupérer un certain nombre de caractéristiques faisant cruellement défaut à l'apprentissage statistique moderne. Cette prise de

conscience globale a été le déclencheur de ce qui tend désormais à être englobé par le terme de *Cinquième Révolution Industrielle*, c'est-à-dire une recherche de propriétés «socialement acceptables et éthiques» pour les algorithmes utilisés.

Il est important de noter qu'au delà de l'objet de recherche, les concepts étudiés ici le sont en ce qu'ils permettent de répondre à un besoin de la société civile. La dissonance conceptuelle de notre société entre la nécessité de compréhension, de contrôle et de maîtrise de ses outils, et le confort d'une technologie «clé en mains» qui n'a besoin pour fonctionner que de données toujours plus massives, n'était tenable qu'en l'absence d'accrocs. Maintenant qu'une prise de conscience des écueils évoqués précédemment s'est opérée, il semble inenvisageable de refermer cette boîte de Pandore. C'est pour cette raison que de nombreuses démarches provenant des sphères politico-juridiques cherchent à encadrer et codifier, à réguler, ce que sera l'Intelligence Artificielle de demain, gageure impossible sans l'expertise de la communauté scientifique et technique. C'est précisément le côté extrêmement adaptable du Machine Learning, ce qui en a fait sa force, qui oblige les différentes communautés à converger l'une vers l'autre.

D'un point de vue concret, de nombreux textes pionniers ont été produits récemment par la Commission Européenne qui se veut le point focal de cette régulation. Parmi ceux-ci, nous pouvons noter le «Livre Blanc sur l'Intelligence Artificielle», l'«Artificial Intelligence Act», ainsi que le duo de lois que sont le «Digital Market Act» et le «Digital Service Act». Ce quatuor de textes a pour but de définir le paysage technologique, juridique et social dans lequel se positionnent ces outils afin de proposer une nomenclature sur les Systèmes d'Intelligence Artificielle et les contextes particuliers pouvant nécessiter une attention accrue à l'égard de ces algorithmes *data*-vores. *In fine* sont demandés des outils statistiques, des normes techniques et des audits permettant de vérifier que les vertus éthiques et autres caractéristiques mentionnées ont bien été prises en compte.

Mais si la question «Pourquoi l'IA a besoin d'être régulée et encadrée?», ainsi que sa suite logique «Par quel type de moyens l'IA peut-elle être régulée?», sont du ressort de la société, nous allons nous intéresser à la problématique qui sera la clé de voûte de ce manuscrit: «Quelle partie de l'IA voulons-nous encadrer et de quelle manière pouvons-nous le faire?»

Parmi la longue liste des défauts potentiels d'une IA «naïve» (e.g. *accès aux données privées/privacy* ou encore coût écologique), nous allons ici nous préoccuper de deux sujets particuliers que sont la *Loyauté* ou *Équité Algorithmique* (*Algorithmic Fairness* ou tout simplement *Fairness* en anglais) qui se préoccupe des biais discriminants exhibés par les algorithmes; et l'*Explicabilité*, au moyen de l'*Analyse de Sensibilité* (*Sensitivity Analysis* en anglais) que l'on retrouve dans la littérature sous l'acronyme *«XAI»* (*eXplainable AI*). Une emphase sera notamment mise sur la version distributionnelle de ces notions, c'est-à-dire en se plaçant dans un cadre probabiliste qui sera défini par la suite. Chacune sera développée ci-dessous dans une section qui lui est propre, et elles le seront de manière indépendante l'une de l'autre. La structure sera la même dans les deux cas, commençant par les motivations sous-jacentes propres à ces concepts, puis

une analyse des outils proposés par la communauté scientifique pour répondre à ces problématiques. Bien que chaque section se suffise à elle-même, il devrait être clair à la fin de ce chapitre que les problématiques, sans être deux faces d'une même pièce, sont assez proches pour être profondément connectées.

1.2 Équité algorithmique

Une des premières préoccupations dans ce but d'insérer des «valeurs humaines dans l'IA» est l'Équité Algorithmique. Ce sujet est probablement celui qui a été le plus mis en avant par la société ces dernières années et le plus documenté par les médias. Cela vient principalement du fait que les "défauts d'équité", c'est-à-dire les situations dans lesquelles un algorithme fait preuve de discriminations, sont particulièrement visibles. Ainsi, que ce soit en analyse de langage naturel automatisée [De-+19; GG19], en reconnaissance faciale [KMM15], en assurance [BC22] ou encore en justice prédictive [FBL16] (et les exemples ne sont pas restreints aux seuls domaines cités ici), un algorithme dicté uniquement par la recherche de performance maximale peut parfois exhiber des comportements inacceptables.

Notons que la notion d'équité dans la prise de décision sociale n'est pas une idée nouvelle. De nombreux philosophes ont pensé la société au prisme de l'équité. Parmi les plus contemporains d'entre eux, nous pouvons par exemple citer J.Rawls et son ouvrage "Justice as Fairness" [Raw04], écrit en 1985, qui affirme que la société et la justice sont une affaire d'équilibre et d'équité entre les plus favorisés et les plus défavorisés. De manière similaire, en restant dans le domaine juridique, la règle dictée aux États-Unis au début des années 1970 du "80%" (cf. le *California's Fair Employment and Housing Act*) énonce que les ratios d'embauches entre sous-populations défavorisés et favorisées devraient être supérieurs à 80%, sans quoi les recruteurs peuvent être poursuivis pour discrimination. Pourtant, force est de constater que les discriminations peuvent encore survenir. L'automatisation prenant place dans toutes les parts de notre société peut amener autant de potentielles discriminations, avec leurs lots de conséquences désastreuses qui ont été documentées dans un certain nombre de situations (e.g. l'analyse de l'algorithme «COMPAS» développé par l'entreprise Northpointe [Was19]). Ce simple fait est le moteur principal d'une littérature foisonnante spécialisée dans la prise de décision algorithmique équitable, l' *Équité algorithmique*.

Il est important de noter qu'il n'est pas pour autant question ici de simple "traduction" de concepts juridiques et sociaux en termes algorithmiques. Des spécialistes ont mis en avant le fait que les discriminations dues à un algorithme inéquitable sont différentes des discriminations faites par un humain, et plus complexes à déceler, à quantifier et à corriger [WMR17; Dwo+20]. En effet, la discrimination humaine est le plus souvent une discrimination "directe" au sens où l'on constate une lien quasi-causal entre un attribut sensible (appartenance à un groupe discriminé) et une conséquence (comme une embauche ou l'accès à un prêt par exemple). Au contraire, l'algorithme, de par sa nature et son fonctionnement,

a tendance à prendre en compte un très grand nombre de paramètres dans sa prise de décision. Ainsi, ce n'est peut-être pas directement l'attribut sensible qui pousse le changement de décision mais l'effet de cette appartenance sur un grand nombre d'autres paramètres qui, de manière cumulée, va entraîner la discrimination, d'où le terme de discrimination "indirecte". C'est en particulier ce fait qui explique l'inefficacité de la "fairness through unawereness" [Dwo+11] qui prônait le fait de simplement supprimer les attributs sensibles des jeux de données. L'idée peut sembler de prime abord efficace mais ne l'est en fin de compte que très peu, un attribut sensible étant le plus souvent corrélé avec d'autres variables et l'algorithme parvenant ainsi à le «deviner». De plus, la notion d'équité peut dépendre du contexte dans lequel on l'applique. Avec des agents humains (e.g. juges), cette adaptation au contexte est appliquée de fait mais ce n'est *a priori* pas le cas pour une machine [Dwo+20].

Nous avons mentionné précédemment qu'une forte demande d'outils à ce sujet provenait de la société civile. Nous allons donc expliciter ici comment la communauté scientifique répond à ce besoin en décrivant un certain nombre de méthodes. La vaste majorité du propos sera ici dévolue aux méthodes dites distributionnelles, bien qu'un aparté sur le point de vue individuel sera présent. Nous mettrons en exergue comment les métriques d'équité utilisées reviennent le plus souvent à la recherche d'une égalité d'indicateurs à travers tous les groupes et toutes les sous-populations existantes. Un grand pan de la littérature concerne donc, dans un ordre croissant de difficulté, la détection, la quantification et la mitigation de ces biais discriminants. Au fur et à mesure, il devrait paraître de plus en plus clair que l'équité algorithmique est avant tout une question de compromis et de choix éclairé. Bien qu'offrant des outils, aucune réponse n'est purement décorrélée du contexte dans lequel elle est apportée, de sorte qu'un dialogue avec les utilisateurs finaux de l'algorithme sera toujours de mise.

1.2.1 Notations et hypothèses courantes

Le cadre utilisé en Équité Algorithmique est le suivant. Les variables d'entrée utilisées par un algorithme sont considérées comme étant des variables aléatoires. Elles sont données en entrée à un algorithme déterministe qui est assimilé à une boîte noire (c'est-à-dire qu'il est impossible de connaître quoi que ce soit sur la structure de cet algorithme) et ne seront connus que les couples (Variables d'entrée – Sortie). Dans ce cadre, les variables d'entrée sont de deux types. Tout d'abord se trouvent les variables non-protégés, que l'on dénote par $\mathbf{X} = (X_1, \ldots, X_p)$, qui sont les variables ou caractéristiques que l'on retrouve dans le cadre classique en Machine Learning. Le second type sont les variables dites *sensibles* ou *protégées*, que l'on dénote par $\mathbf{S} = (S_1, \ldots, S_m)$. Ces dernières représentent des marqueurs sociaux ou biologiques pouvant être source de biais. En général, ces variables sensibles sont binaires et marquent l'appartenance à une partie de la population (e.g origine ethnique, sexe biologique, etc) mais des exemples de variables sensibles continues existent, notamment en biologie et en médecine (e.g. poids, taux de glucose dans le sang, marqueur d'handicap, etc). Enfin, nous avons un algorithme f sur lequel aucune hypothèse n'est faite (i.e f

Symbole mathématique	Signification
$\mathbf{X} = (X_1, \ldots, X_p)$	Liste des variables non-protégés en entrée de l'algorithme
$\mathbf{x} = (x_1, \ldots, x_p)$	Réalisation de la variable aléatoire $\mathbf{X}$
p	Nombre de variables non-protégés utilisés
$\mathbf{S} = (S_1, \ldots, S_m)$	Liste des variables sensibles en entrée de l'algorithme
$\mathbf{s} = (s_1, \ldots, s_m)$	Réalisation de la variable aléatoire $\mathbf{S}$
m	Nombre de variables sensibles utilisés
f	Algorithme utilisé
$Y := f(\mathbf{X}, \mathbf{S})$	Variable aléatoire correspondant à la sortie de f
$\mathcal{F}(\cdot)$	Mesure de Fairness générique appliquée à un modèle.

Table 1.1: Différentes notations utilisées dans le cadre «Fairness»

est une *boîte noire*). Si besoin est, nous dénoterons par une lettre minuscule une réalisation du vecteur aléatoire dénoté par la même lettre en majuscule (e.g $\mathbf{x}$ pour une réalisation de $\mathbf{X}$).

Dans la majorité de la littérature, $\mathbf{S}$ est supposé unidimensionnelle (i.e. il n'y a qu'une seule variable sensible). Dans ce cas, nous dénotons par S cette variable protégée. Le cas multidimensionnel est lié à la notion d'intersectionalité, voir Remarque 1.

Pour compléter le cadre, l'Équité se quantifie à l'aide d'une «mesure de Fairness». Cette quantité a pour but de quantifier à quel point un modèle est équitable en un certain sens. Lorsque nécessaire, nous noterons par $\mathcal{F}(\phi)$ une mesure de Fairness générique qui serait appliquée à l'algorithme ϕ (par rapport à la variable S qui sera le plus souvent sous-entendue). Des exemples plus précis seront fournis lorsque nous étudierons les différentes définitions d'Équité.

Remarque 1 (Intersectionalité)**.** *Bien qu'une part conséquente de la littérature soit composée de l'analyse de biais provoqués par une seule variable protégée, il est possible pour un individu d'être discriminé de manière particulièrement véhémente en raison d'une appartenance à plusieurs minorités en même temps [Fou+19; Bén+21]. Ce fait a par exemple été documenté pour la reconnaissance faciale. Dans ce domaine, les personnes ayant une peau foncée peuvent être moins bien reconnue que celles possédant une peau claire. De manière similaire, les modèles performent parfois moins bien sur les femmes que sur les hommes (en considérant cette variable binaire). Des analyses ont montrée que les performances sur des femmes de couleurs pouvaient parfois être encore plus dégradées qu'un simple cumul des biais «homme-femme» ou ceux dûs à la couleur de peau.*

Cette discrimination «majorée» est prise en compte dans ce que l'on appelle l'équité intersectionnelle. Ces biais apparaissent lorsqu'un individu appartient à l'intersection de plusieurs groupes minoritaires, dans un cadre avec plusieurs variables sensibles. Il est possible de quantifier ce biais additionnel de la même manière qu'un biais classique mais l'expérimentateur est alors rapidement confronté à un problème d'échantillonage. Trouver quelqu'un appartenant à diverses minorités est une tache de plus en plus ardue au fur et à mesure que le nombre

de variable augmente. L'estimation de ces biais est donc de plus en plus complexe et sensible au bruit.

Nous verrons dans le prochain chapitre des outils permettant, à défaut d'estimer précisément la force de ces iniquités intersectionnelles, de les majorer sans coût additionnel.

1.2.2 Définitions classiques de métriques de Fairness

Équité de groupe

L'idée fondatrice de l'équité de groupe (*Group Fairness* dans la littérature anglo-saxonne) est la suivante: il faut que le modèle ait des comportements similaires au sein de la classe majoritaire et de la classe minoritaire (dans le cas binaire $S = 0$ ou $S = 1$, qui est le plus courant). Si tel est le cas, alors l'algorithme sera considéré comme étant équitable. Dans le cas contraire, l'appartenance à une de ces deux sous-populations est le marqueur d'un changement de décision de la part de l'algorithme et celui-ci présente donc des biais discriminants. Le lecteur intéressé peut trouver plus de détails dans [Dwo+11; Cho17; BGL20].

Mais qu'entendons-nous par le terme de «comportements similaires»? Nous allons voir que le comportement visé correspond à une *métrique de Fairness* et encode ici une information en moyenne ou une information distributionnelle sur les deux lois de probabilités que sont $\mathbb{P}_{f(\mathbf{X},S=0)}$ et $\mathbb{P}_{f(\mathbf{X},S=1)}$. Nous présentons ici les métriques les plus classiques mais le fond reste toujours le même: l'équité parfaite correspond à l'égalité de ces métriques à travers toutes les sous-populations.

Métriques en moyenne L'exemple le plus classique d'équité de groupe est probablement le *Disparate Impact* [Bes+20]. Si l'on suppose que l'on a un classifieur binaire ($f(\mathbf{X}, S) \in \{0, 1\}$) et une variable protégé elle-même binaire (e.g. appartenance à une minorité), alors l'équité est atteinte au sens du Disparate Impact lorsque la probabilité de succès est la même, quel que soit la valeur de l'attribut sensible: $\mathbb{P}(f(\mathbf{X}, S) = 1|S = 0) = \mathbb{P}(f(\mathbf{X}, S) = 1|S = 1)$. Néanmoins, cette condition peut être trop restrictive et donc relâchée afin d'obtenir une équité approximative, voir Remarque 3. Nous utiliserons notamment dans la suite la relaxation $\left(\mathbb{P}(f(\mathbf{X}, S) = 1|S = 0) - \mathbb{P}(f(\mathbf{X}, S) = 1|S = 1)\right)^2 \leq \varepsilon$ avec ε un seuil limite. Dans la littérature, une autre relaxation est souvent utilisée en prenant le ratio de ces deux quantités. Ainsi, il faut que le *Disparate Impact*, noté DI et défini par

$$\mathrm{DI}(f) := \frac{\mathbb{P}(f(\mathbf{X}, S) = 1|S = 0)}{\mathbb{P}(f(\mathbf{X}, S) = 1|S = 1)}, \tag{1.1}$$

soit le plus proche possible de 1. Si le Disparate Impact est égal à 1, alors les deux groupes ont les mêmes chances de succès. Notez que la règle des "80%" évoquée un peu plus tôt correspond au fait que le Disparate Impact soit supérieur à $0,8$.

Remarque 2. *Le Disparate Impact a été défini à l'origine dans le cadre d'un seul attribut sensible binaire. Parce que l'on considère la condition $S = 0$ en*

numérateur et $S = 1$ *au dénominateur, avec l'hypothèse implicite que la probabilité de succès est plus faible pour la sous-population discriminée, le Disparate Impact sera toujours compris entre* 0 *et* 1. *Si l'attribut sensible est catégoriel, alors cette quantité peut tout de même être défini en comparant le minimum de la probabilité de succès pour toutes les sous-populations sur le maximum de la probabilité de succès, c'est-à-dire*

$$\mathrm{DI}(f) := \frac{\min_s \mathbb{P}(f(\mathbf{X}, S) = 1 | S = s)}{\max_{s'} \mathbb{P}(f(\mathbf{X}, S) = 1 | S = s')}. \tag{1.2}$$

Cette métrique, malgré (et à cause de) sa simplicité, présente un certain nombre de limites. Cela a amené la communauté à développer des variantes, souvent basées sur l'égalité pour tous les sous-groupes d'autres quantités en moyenne relatives au classifieur. C'est par exemple le cas des métriques «*Avoiding Disparate Treatment*», «*Avoiding Disparate Treatment*», ou encore «*Equality of Odds*». Le lecteur trouvera dans la Table 2.2 la définition mathématique de ces métriques.

Métriques distributionnelles Un point commun entre toutes ces techniques est la recherche d'une égalité entre deux distributions, celle des sorties données par l'algorithme pour la sous-population discriminée $\mathbb{P}_{f(\mathbf{X},0)}$ et celle des sorties données par l'algorithme pour la sous-population non-discriminée $\mathbb{P}_{f(\mathbf{X},1)}$. Or, comme nous le verrons dans la section suivante relative à l'explicabilité, il existe diverses distances ou divergences pour également quantifier ce type d'égalités distributionnelles. Ce constat a entraîné son lot de nouveaux indicateurs d'équité. Parmi ceux-ci, nous pouvons par exemple citer

1. l'utilisation du transport optimal [Bar+18], comparant ainsi la distribution des sorties données à la classe majoritaire à la distribution des sorties données à la classe minoritaire ;

2. l'utilisation des *HSIC* (Hilbert-Schmidt Independence Criterion), critère basé sur la notion de *RKHS* (Reproducing Kernel Hilbert Space) fréquemment utilisée pour les techniques à noyaux, voir [Gre+; Li+19] ; ou encore Shapley, critère provenant de la théorie des jeux (voir [FRF; HDV20]) ; tous deux explicités dans la prochaine section ;

3. ou encore *la corrélation maximale HGR (Hirschfeld-Gebelein-Rényi)* qui est une généralisation du coefficient de corrélation de Pearson avec l'utilisation de fonctions test, voir [MCK].

Cette liste est loin d'être exhaustive et nous recommandons au lecteur intéressé de parcourir les références suivantes, toutes développant une nouvelle métrique quantifiant une notion de distance ou de divergence distributionnelle: [GKK18; BGW19; Chi19; ODP19; Ris+21]. Ces dernières années ont vu l'arrivée d'un grand nombre de méthodes quantifiant des marqueurs et couvrant des définitions spécifiques de l'Équité de Groupe, et le chapitre suivant sera un moyen de mettre un cadre unificateur sur ces méthodes.

Remarque 3 (Équité parfaite ou approximative)**.** *L'équité algorithmique est parfaitement atteinte lorsque les quantités d'intérêt mentionnées au-dessus (taux de succès dans le cas le plus simple) sont égales pour toutes les sous-population potentiellement discriminées. Néanmoins, cette égalité s'obtient parfois au prix d'une dégradation des performances, de sorte que le praticien soit confronté à un compromis entre un algorithme informatif et répondant à un objectif potentiellement biaisé, et un algorithme parfaitement équitable. Il est courant que ce compromis soit particulièrement complexe à respecter et que la contrainte d'équité soit trop dure pour être implémentée de manière réaliste. Dans ce cas, il est nécessaire de «relâcher» celle-ci et de demander non pas une parfaite égalité à travers tous les groupes, mais simplement «suffisamment peu de différences». Cette équité approximative nécessite donc un seuil à fixer au préalable, seuil correspondant au niveau de laxisme autorisé au modèle utilisé. Ce niveau dépend du contexte et provient des contraintes conjoncturelles (e.g. normes techniques imposées par un organisme). Comme mentionné plus tôt, ce concept n'a pas été inventé pour répondre aux problématiques modernes de Machine Learning puisque l'on retrouve des équivalents directs dans les lois déjà en vigueur.*

Remarque 4 (De la nécessité d'un choix)**.** *Nous l'avons évoqué précédemment, le choix d'une métrique de fairness est primordial afin de cibler un type de discrimination précis que l'on veut réduire ou faire disparaître. Mais sommes-nous réellement obligés de faire ce choix?*

Malheureusement, la réponse est oui. Il est nécessaire de faire un choix, bien que l'idée de forcer un modèle à être équitable selon toutes les définitions d'équité semble intéressante et, in fine, l'idéal. C'est pourtant impossible car ces définitions sont parfois incompatibles entre elles. Alors même que forcer un algorithme à être équitable par rapport à une métrique revient à faire un compromis, en rajouter d'autres peut imposer plus de contraintes et mener l'algorithme à ne plus être informatif. De plus, certaines conceptions de l'équité sont intrinsèquement opposées et incompatibles, de telle sorte qu'il soit impossible de les réaliser en même temps. Ainsi, bien qu'il soit tentant d'éviter le choix d'une équité à respecter, cette démarche n'est pas une solution [KMR16].

Équité locale

Un autre penchant de l'équité algorithmique a été développé non pas au niveau distributionnel («Est-ce que le modèle possède un biais générique contre telle minorité?») mais plutôt au niveau individuel («Est-ce que tel individu en particulier a été victime de discrimination?»). Cette vision, connue sous le nom d'équité locale (*Local Fairness*), englobe l'ensemble des techniques permettant de répondre à cette dernière question. Un des méthodes les plus courantes consiste en la recherche de ce qui est appelé dans la littérature de Machine Learning des *contrefactuels* [Kus+18; Lar+21b; WMR17]. Pour un individu donné, appartenant à une minorité, sera créé un alter-ego en tout point semblable mais appartenant à la majorité. Si l'individu de la minorité a moins de chances de succès que son équivalent dans la majorité, alors c'est là la preuve d'une

discrimination. Néanmoins, des travaux récents ont mis en exergue les difficultés que peut engendrer la recherche de ce doppelganger. Pour visualiser le cœur du problème, prenons un exemple simple. Supposons que Bob soit un homme mesure 1m85, ce qui le place aux environs des 10% des hommes les plus grands. Si nous sommes à la recherche d'Alice, un alter ego de Bob de sexe féminin, faut-il que celle-ci mesure 1m85 également? Ou serait-il préférable qu'elle ait la taille qui la placerait au niveau des 10% des femmes les plus grandes (soit aux alentours de 1m70)? La réponse à cette question, une fois de plus, peut dépendre du contexte (e.g. le modèle cherche-t-il une taille absolue à dépasser ou le simple fait d'être parmi les plus grands?). Mais au delà de ce fait, la notion même de quantile, que nous avons utilisé dans ce cas simple univarié, n'est pas bien définie pour plusieurs variables, ce qui rend l'analogie plus complexe en multivarié. Pour cette raison, certains travaux ont été menés afin d'utiliser la notion de transport optimal, permettant de connaître pour tout individu son équivalent dans l'autre sous-population.

Remarque 5 (Graphes causaux)**.** *Un moyen de rapidement visualiser les potentielles sources d'iniquités est l'utilisation de graphes causaux [Pea09; Sch19]. Ces DAGs (Directed Acyclic Graphs) montrent les influences existantes entre toutes les variables d'entrée, et entre ces variables et la sortie du modèle. Par le prisme de cet outil, il est donc possible de cibler une manière spécifique qu'aurait une variable sensible de provoquer des discriminations dans la décision algorithmique en bout de chaîne [BGW19; Chi19]. Il est à noter que chacune de ses manières se traduit directement en Équité globale en terme d'indicateurs vus précédemment. Que ce soit par un lien direct de la variable protégée vers la sortie, par agrégat à travers tous les chemins possibles ou encore par le biais d'un chemin spécifique, il est possible de quantifier exactement le poids de cette iniquité ciblée. Cela se traduit également d'un point de vue local, individuel, ou au niveau du modèle par l'utilisation de contrefactuels, aisément représentés et visualisables sur un graphe causal.*

Néanmoins, bien que ces méthodes soient extrêmement pratiques de par leur transversalité, elles ne sont pas toujours utilisables. En effet, au delà de potentiels problèmes d'identifiabilité, leur coût devient souvent prohibitif lorsque le nombre de variables considérées croît, en raison de l'explosion de la combinatoire associée [Lar+21b].

1.2.3 Mitigation des biais

Après avoir détecté un biais discriminant dans une décision algorithmique et quantifié l'ampleur de ce défaut, encore faut-il être capable de le corriger! Pour cela, il existe trois familles de procédés permettant d'opérer, connues sous le nom de «*pré-processing*», «*in-processing*» et «*post-processing*».

1. Le *pré-processing* consiste à travailler sur les données fournies en amont au modèle. L'idée est de donner à l'algorithme non pas les données réelles mais une version modifiée ou transformée de telle manière que l'algorithme renvoie des réponses «fair» à la fin. Un avantage direct de ce type de

méthode est qu'aucune opération n'est faite directement sur le modèle. Il n'y a pas besoin d'un nouvel apprentissage, ni d'aucune modification sur la structure ou les paramètres de l'algorithme. Cela en fait une technique efficace lorsque ce dernier est fixé une bonne fois pour toute sans retouche possible, pour des raisons d'accès, de coûts, etc. En revanche, la transformation à faire en amont sur les données pour garantir un modèle équitable n'est *a priori* pas générique et dépend du modèle. Elle peut également être difficile à calculer et devra en permanence être utilisée, à la manière d'une surcouche sans laquelle la garantie d'équité tomberait.

2. Le *in-processing*, deuxième moyen de mitiger les biais discriminants, opère directement lors de la phase d'apprentissage de l'algorithme. L'idée est de rajouter à l'objectif que le modèle essaie d'atteindre au mieux possible un terme de pénalisation correspondant aux potentielles iniquités. Il est à noter que ce sont des techniques courantes en apprentissage et que de telles pénalisations sont des outils classiques pour forcer un modèle à respecter une contrainte voulue par l'expérimentateur. L'in-processing est en général la méthode privilégié et de nombreux outils (différentiation automatique, pertes adaptatives, choix automatique et par validation des hyperparamètres) ont été développés dans ce but. En revanche, et c'est là le défaut principal de cette méthode, il est nécessaire de ré-entraîner le modèle, ce qui engendre des coûts supplémentaires et n'est pas toujours faisable lorsque le modèle est déjà fixé.

3. Le *post-processing*, agissant en fin de chaîne, est très proche conceptuellement du pré-processing. Au lieu de modifier les données avec lesquelles on nourrit notre algorithme, l'idée est ici d'en changer les réponses. D'une certaine manière, le post-processing consiste à passer derrière le modèle en étant conscient de ses biais et de ses spécificités, et d'agir tel un correcteur en aval («Le robot a dit X mais en réalité, ce qu'il voulait dire est plutôt...»). De la même manière que pour le pré-processing, on retrouve des avantages et inconvénients similaires. Aucune modification ou retouche du modèle n'est nécessaire. En revanche, chaque algorithme ayant ses spécificités, il incombe au praticien la totalité de la charge: trouver le bon filtre correspondant à la métrique de fairness voulue, savoir de quelle manière le faire agir et s'assurer qu'il est toujours appliqué aux réponses algorithmiques.

Pour résumer chacune de ces techniques, il est possible de faire une analogie entre un modèle que l'on veut rendre équitable, et quelqu'un à qui l'on veut apprendre à ne pas jouer avec un couteau. Le pré-processing, c'est s'assurer qu'il n'y ait aucun couteau dans la pièce et qu'il n'y en aura jamais. L'in-processing, c'est apprendre à la personne qu'un couteau peut être dangereux et de lui demander de faire attention. Le post-processing, c'est potentiellement laisser des couteaux accessibles, et passer derrière les éventuels blessures pour les panser avant qu'elles n'affectent quelqu'un. La première et la troisième méthode ne nécessitent aucun apprentissage mais impose une vigilance accrue. La seconde en est l'opposé.

1.3 Quantification d'incertitudes

L'explicabilité des décisions algorithmiques est également au cœur des préoccupations, et son absence peut être un défaut majeur rencontré. Les meilleurs performances sont souvent atteintes pour des architectures de type "réseaux de neurones". Pour ces structures, les résultats théoriques sont encore peu nombreux – malgré une accélération impressionnante de la quantité de travaux à ce sujet. De plus, il est extrêmement délicat, dans ces situations, de comprendre une fois les phases d'entraînement terminées, quelles sont les variables responsables des décisions ou réponses proposées par un algorithme [Sam+21]. Néanmoins, ce problème n'est pas l'apanage seul des techniques modernes d'apprentissage statistique En effet, cet écueil peut également survenir dans le cas de codes expérimentaux, par exemple lorsque les codes utilisés sont la concaténation de nombreuses recherches, menant à une impossibilité de connaître en détail l'ensemble des opérations prenant place dans le code utilisé [FKL21]. Ainsi, pour diverses raisons, il arrive qu'il soit extrêmement difficile et onéreux, voire tout simplement impossible, de comprendre la totalité de ce qu'il se passe à l'intérieur de ces codes ou algorithmes.

Face à ce constat, la communauté statistique apporte tout un panel d'outils qui sont inclus sous la dénomination de *Quantification d'incertitudes* pour obtenir de l'explicabilité et de la compréhension dans le fonctionnement interne de ces algorithmes. Nous mettrons tout particulièrement l'emphase sur ce qui est appelé l'*Analyse de Sensibilité* [Da 15; Da +21]. L'objectif principal de cette littérature est de répondre à la question suivante: *Parmi toutes mes variables d'entrées, laquelle ou lesquelles sont les plus influentes sur la sortie donnée par l'algorithme que j'ai à ma disposition?* Nous allons voir que, de manière similaire à l'équité, des points de vue locaux et globaux peuvent être adoptés afin de comprendre au mieux le fonctionnement d'une *boîte noire*, c'est-à-dire un algorithme pour lequel il est impossible de connaître la structure interne. Comme précédemment, nous porterons tout particulièrement notre attention sur le point de vue distributionnel. Pour cela, nous serons amenés à définir un certain nombre d'indices qui seront des indicateurs de ces influences. Par la suite, nous mentionnerons des travaux portant sur l'estimation de ces indices, observant l'apparition de nouvelles techniques et schémas d'estimation. Enfin, nous aborderons ce qu'il se passe lorsque l'on change la distribution des entrées; ou encore lorsque l'on utilise non pas l'algorithme directement mais plutôt un *méta-modèle*, c'est-à-dire une approximation de celui-ci.

1.3.1 Notations et hypothèses courantes

De manière similaire à la situation précédente, en Analyse de Sensibilité, les variables d'entrée sont considérées comme étant des variables aléatoires. Elles sont données en entrée à un algorithme déterministe qui est une boîte noire (c'est-à-dire qu'il est impossible de connaître quoi que ce soit sur la structure de cet algorithme) et ne seront connus que les couples (Variables d'entrée – Sortie). En revanche, contrairement au cadre "Équité", il n'y a pas de variable sensible

Symbole mathématique	Signification
$\mathbf{X} = (X_1, \dots, X_p)$	Liste des variables en entrée de l'algorithme
$\mathbf{x} = (x_1, \dots, x_p)$	Réalisation de la variable aléatoire $\mathbf{X}$
p	Nombre de variables utilisées
f	Algorithme utilisé
$Y := f(\mathbf{X})$	Variable aléatoire correspondant à la sortie de f
$GSA_Z(f)$ par rapport à une variable aléatoire Z.	Indice d'Analyse de Sensibilité Globale générique, appliqué pour un algorithme f et

Table 1.2: Différentes notations utilisées dans le cadre GSA

en Analyse de Sensibilité. Nous avons donc en notre possession un vecteur de variables aléatoires en entrée $\mathbf{X} = (X_1, \dots, X_p)$, le plus souvent à valeurs réelles. Est également donné, un algorithme déterministe f que l'on considère de carré intégrable, c'est-à-dire pour lequel la quantité $\mathrm{Var}_{\mathbf{X}}(f)$ est finie. Si besoin est, nous dénoterons par $\mathbf{x}$ une réalisation du vecteur aléatoire $\mathbf{X}$.

Une hypothèse classique de cette littérature est l'indépendance des variables d'entrées. Cela se traduit par l'égalité $\mathbb{P}_{\mathbf{X}} = \mathbb{P}_{X_1} \times \cdots \times \mathbb{P}_{X_p}$. Cette hypothèse vient des origines historiques du domaine puisque souvent réalisées dans le cas de codes expérimentaux, dans lequel les variables génératrices sont souvent uniformes entre deux valeurs nominales, et indépendantes les unes des autres. Néanmoins, elle est très peu réaliste dans le cas d'applications réelles, ce qui a mené l'émergence de travaux visant à relâcher cette condition. Nous les évoquerons plus en détail lorsque nous parlerons des indices de Sobol'.

1.3.2 Une question de distances

En réalité, lorsque l'on s'intéresse à l'influence d'une variable aléatoire sur une autre, ou sur une sortie d'algorithme, nous sommes en fin de compte en train de nous intéresser à la notion intuitive de "quasi-indépendance". Même si ce terme n'a, *a priori*, pas de définition claire d'un point de vue mathématique, il semble intuitif de dire que moins la variable aléatoire X ne va avoir d'influence sur la variable aléatoire Y, plus ces deux variables sembleront indépendantes l'une de l'autre.

En se souvenant de la définition d'indépendance, c'est-à-dire le fait que X et Y sont indépendantes si et seulement si les lois $\mathbb{P}_{X,Y}$ et $\mathbb{P}_X \times \mathbb{P}_Y$ sont égales, une généralisation de cette définition à la quasi-indépendance semble immédiate en terme de distances. Ainsi, il semble naturel de dire qu'une variable aléatoire X_1 a moins d'influence sur Y que X_2 si la distance entre les distributions $\mathbb{P}_{X_1,Y}$ et $\mathbb{P}_{X_1} \times \mathbb{P}_Y$ est plus petite que la distance entre $\mathbb{P}_{X_2,Y}$ et $\mathbb{P}_{X_2} \times \mathbb{P}_Y$. En prenant le cas extrême de l'égalité, nous retombons sur la propriété d'indépendance: la variable aléatoire X_1 pourrait être supprimée ou mise à une valeur nominale sans aucune répercussion sur Y; elle est donc non-influente.

C'est cette idée générale qui guide l'Analyse de Sensibilité Globale [Da 15].

En regardant cette distance $d(\mathbb{P}_{(X_i,Y)}, \mathbb{P}_{X_i} \times \mathbb{P}_Y)$ pour l'ensemble des variables d'entrée X_i, on obtient ainsi une idée de la variable ayant le plus d'influence (celle pour laquelle la distance est la plus élevée), la deuxième la plus influente, etc, jusqu'à celle enfin qui est le moins informative sur la sortie du modèle. Pour cette dernière, si la distance calculée entre la loi jointe et la loi produit est nulle, il est possible de la supprimer des données sans aucune répercution sur l'efficacité de l'algorithme [AC21; Cha20; Big+21]. Cela fournit donc un premier ensemble d'indices permettant d'expliquer certaines caractéristiques du modèle.

Remarque 6 (Analyse de sensibilité locale)**.** *Bien que la vaste majorité des méthodes présentées ici rentre dans le cadre de l'Analyse de Sensibilité Globale, un pan de la littérature d'intéresse, de manière similaire à l'Équité, au point de vue local ou individuel [GSW96; RSG16]. L'idée est de fournir des indicateurs répondant à la question «Pour tel individu donné, quelle est la variable pour laquelle un petit changement serait le plus à même de faire varier la sortie algorithmique?» La méthode naïve pour apporter une réponse à ce type de question consiste à faire un développement de Taylor du modèle autour du point correspondant à l'individu, et donc à calculer les dérivées partielles du modèle. En contrepartie, cela nécessite des hypothèses supplémentaires sur la régularité, hypothèses parfois non vérifiées. Nous verrons néanmoins que cette idée a inspiré la création d'indices d'explicabilité globale: ce sont les* indices basés sur la dérivée *[Rou+14; Lüt+21].*

Alors que le concept semble clair, un certain nombre de problèmes arrivent néanmoins lorsque l'on essaye de le mettre en pratique, et notamment le choix de la distance. Il existe une infinité de distances ou de mesures de dissimilarités définies entre distributions. Parmi les plus connues et les plus utilisées par la communauté statistique, nous pouvons, par exemple, mentionner

1. la dissimilarité de Kullback-Liebler, trouvant ses fondements dans la théorie de l'information;
2. la distance en variation totale, utilisée principalement en Probabilités;
3. la distance de Wasserstein, remise au goût du jour ces dernières années de par les récents développements computationnels et son adéquation avec certains algorithmes de Machine Learning;
4. ou encore les $\varphi-$divergences, introduites par [Csi67], offrant tout une gamme de pseudo-distances possibles.

Bien que des liens puissent être tirés entre ces distances et divergences, elles ne sont pas équivalentes, ayant chacune ses spécificités. Cela rend le choix de l'une plutôt que l'autre subjectif et assez complexe au néophyte ou non-expert.

Et quand bien même une distance particulière serait privilégiée, les quantités d'intérêt évoquées sont coûteuses et complexes à calculer . Bien qu'étant les objets qui, *in fine*, nous intéressent, ce cumul de désavantages a poussé la communauté à délaisser les distances entre distributions pour se tourner vers des quantités ou des outils moins intrinsèques mais plus aisément accessible.

Remarque 7. *Dans la littérature, il est possible de parfois rencontrer la distance entre $\mathbb{P}_Y$ et $\mathbb{P}_{Y|X}$ plutôt que les deux distributions mentionnées au dessus. L'idée reste la même. En effet, il suffit d'écrire $\mathbb{P}_{(X,Y)}$ sous la forme $\mathbb{P}_X\mathbb{P}_{Y|X}$ pour se rendre compte que l'on regarde dans un cas $d(\mathbb{P}_X\mathbb{P}_{Y|X}, \mathbb{P}_X \times \mathbb{P}_Y)$ et dans l'autre simplement $d(\mathbb{P}_{Y|X}, \mathbb{P}_Y)$.*

À l'interface entre ces méthodes distributionnelles et celles développées autour de la variance que nous verrons après se trouvent les HSIC (*Hilbert-Schmidt Independence Criterion*) [Gre+]. Basés sur une approche de type *noyaux*, l'idée est de généraliser la notion de covariance, par la notion d'opérateur de covariance croisée. Cet opérateur quantifie les corrélations aux différents ordres arrivant lorsque l'on regarde les données au prisme d'un noyau (ou *kernel*) choisi. L'avantage de cette technique est de proposer un cadre commun dans lequel l'estimation de ces indices est aisée et dans lequel le praticien peut «mettre sa patte» par le choix du noyau. En revanche, afin d'être réellement indicatif, il est nécessaire que les noyaux soient *caractéristiques*. Dans ce cas, et uniquement si les noyaux sont caractéristiques, la nullité de l'indice HSIC traduit réellement une indépendance des variables et non pas simplement une absence de corrélation. De plus, en l'absence de normalisation et de bornes sur les valeurs possibles, l'analyse peut être délicate, à moins d'utiliser de récents résultats au prix d'une contrainte de plus. Le cumul de ces deux contraintes sont le moteur du développement des indices basés sur la décomposition de la variance.

1.3.3 Indices basés sur la décomposition de variance

Les indices les plus utilisés sont les indices de Sobol' [Sob; Sal+10]. Basés sur la décomposition d'Hoeffding, équivalent de l'ANOVA fonctionnelle que l'on applique à la sortie de l'algorithme, chaque indice est associé à une variable ou à un groupement de variable et permet de quantifier l'influence sur la sortie, par le biais d'un pourcentage de variabilité expliquée. Parmi les avantages présentés par cette méthode, ce «pourcentage de variabilité expliquée» permet de traquer les influences et leurs origines, et donc de répondre aisément à des questions telles que «*Quelle est la variable la plus déterminante dans le résultat?*» (i.e. quelle est la variable avec l'indice de Sobol' le plus élevé?) ou encore «*Peut-on mettre cette variable à une valeur nominale, voire la supprimer, sans que cela n'ait de conséquence sur le résultat?*» (i.e. cette variable a-t-elle un indice de Sobol' suffisamment faible). C'est par exemple un élément de décision possible dans un cas de *screening* ou de réduction de dimensions [AC21; Big+21]. De plus, par le biais de comparaison entre l'indice de Sobol' d'une variable et ceux des regroupements de variables dans lesquels elle apparaît, il est possible d'inférer si une entrée est influente par elle-même ou de manière jointe avec d'autres, ce qui peut-être une information précieuse. Néanmoins, malgré tous ces avantages, la pierre angulaire de cette technique est cette décomposition d'Hoeffding, qui n'est valable que lorsque les différents entrées sont indépendantes les unes des autres. Il va de soi que cette hypothèse n'est pas toujours réalisée. Afin de contourner cette limitation, de nombreux travaux ont vu le jour, principalement proposant

de ramener le cas non-indépendant au cadre classique. Pour cela, on peut utiliser ou bien des décompositions hiérarchiques alternatives à celle d'Hoeffding [Gra15; CGP11] ou alors passer par des transformations isoprobabilistes [MTA15] comme la transformation de Lévy-Rosenblatt [Lév55; Ros52] (d'une distribution multivariée quelconque vers des distributions uniformes indépendantes) ou la transformée de Nataf (d'une distribution multivariée quelconque vers une copule gaussienne). Ces derniers travaux, menant à la définition des indices de Sobol' étendus, permettent alors de prendre en compte un troisième type d'influence (les deux premiers étant l'effet direct et l'effet joint avec d'autres variables) qui n'apparaît pas dans un cadre indépendant: l'effet rebond. Dans cette situation, une variable est influente sur la sortie par l'effet qu'elle a sur une variable intermédiaire qui, à son tour, agira sur la sortie. Ces différents types d'influence sont représentées de manière plus visuelles sous forme de graphe causal, voir Figure 2.1.

Cependant, la famille des indices de GSA basés sur une décomposition de la variance est loin de se limiter aux seuls indices de Sobol'. En effet, cette dernière décennie a vu apparaître d'autres ensembles d'indices, tous liés à cette brique élémentaire décrite dans le paragraphe précédent ; chacun répondant à une limitation présentée par les indices de Sobol'. Parmi ceux-ci, nous mentionnons ici les indices de Shapley [Owe14; OP17; IP17], les indices de Cramér-von-Mises [GKL18] et les plus récentes généralisations qui en découlent pour des espaces métriques quelconques [FKL21].

Les indices de Shapley ont été proposé comme une première technique s'affranchissant de l'hypothèse d'indépendance des variables. Provenant de la théorie des jeux, ils quantifient originellement la contribution d'un joueur à la complétion d'un objectif complexe, en prenant en compte les contributions individuelles et coalitionnelles de ce joueur avec d'autres groupements de joueurs [Sha16; Deh17]. En transférant cet outil dans un cadre GSA, les joueurs deviennent des variables d'entrée, l'objectif complexe devient la sortie d'un algorithme «boîte-noire» et la contribution devient la part d'influence. Ces indices sont extrêmement intéressants et ont rapidement trouvé leurs lettres d'or dans une utilisation industrielles de par leur simplicité d'estimation et leur caractère informatif en tant qu'information agrégée rendant compte des différentes sources d'influences possibles pour une variable [ICI21]. En revanche, une attention particulière reste de mise dans leur analyse afin de ne pas leur faire dire plus qu'ils ne disent réellement.

Les indices de Cramér-von-Mises, quant à eux, répondent à une autre limitation: une décomposition de la variance ne traduit *a priori* qu'une «information $\mathbb{L}^2$», c'est-à-dire une information de second moment sur la distribution de la sortie. Pour répondre à ce problème, les indices de Cramér-von-Mises [GKL18] ont vu le jour, pouvant être vus de manière équivalente comme l'utilisation de la distance de Cramér-von-Mises plutôt que la variance ; ou alors comme l'application de la décomposition d'Hoeffding aux différentes lignes de niveaux de la fonction de répartition de la loi de sortie, qui correspond à une information distributionnelle. Pour des considérations pratiques, ces valeurs obtenus sur les différentes lignes de niveaux sont moyennées afin d'obtenir *in fine* des indices

intelligibles prenant leurs valeurs entre 0 et 1. Il est à noter que ces indices de Cramér-von-Mises utilisent implicitement les indices de Sobol' – même si appliqués à un autre objet que dans le cadre classique – et peuvent donc bénéficier de certaines extensions mentionnées précédemment, au prix d'une analyse plus complexe et moins directe. Ce sont également la porte d'entrée vers la définition de généralisations pour des espaces métriques quelconque [Gam+20], étendant encore le champ d'application possible de ces indices.

Remarque 8. *Les quelques indices présentés ici forment la base des indices basés sur la décomposition de variance, mais pas la totalité des outils que le lecteur intéressé peut rencontrer dans cette littérature. Ainsi, nous pouvons par exemple mentionner ici quelques autres indices développés au cours des années pour répondre à des besoins spécifiques tels que:*

1. *les «*indices basés sur la dérivée*» [Rou+14; Lüt+21], faisant le lien entre l'Analyse de Sensibilité locale utilisant les informations contenues dans les dérivées directionnelles du modèle, et les indices de Sobol'.*

2. *les indices «*orientés sur une cible*» [RM18; DBM22], que l'on peut positionner à l'interface entre les indices de Sobol' et les indices de Cramér-von-Mises. Nous avons vu que les indices de Sobol' sont une décomposition de la variance du modèle f ; et que les indices de Cramér-von-Mises sont une décomposition de la variance de la fonction de répartition de la sortie. Les indices orientés sur une cible sont un entre-deux puisqu'ils correspondent à la décomposition de la variance d'une ligne de niveau spécifique de la fonction de répartition. Cette ligne de niveau correspond à une «zone de défaillance» en général, c'est-à-dire une région dans laquelle l'algorithme prédit une valeur supérieur à un seuil fixé (e.g. une probabilité de faille d'un système qui ne doit pas excéder un certain pourcentage). Ces indices permettent de répondre à la question «Quelle est la variable qui contribue le plus à ce dépassement de seuil de mon système?».*

3. *Les indices à base de contraste [FKR16], qui sont basés non pas sur une décomposition de la variance (c'est-à-dire de la distance $\mathbb{L}^2$) mais d'une autre fonctionnelle à choisir au préalable afin de quantifier par exemple la sensibilité par rapport à un quantile, à une densité, etc.*

Remarque 9. *De récents travaux concilient l'aspect extrêmement informatif des indices de Sobol' et les indices HSIC [Da 21]. Grâce à la décomposition d'Hoeffding et le fait que l'on puisse voir un indice de Sobol' comme un pourcentage de variance expliquée, ces indicateurs sont explicites et compréhensible de tous, y compris des non-experts. D'un autre côté, l'utilisation d'un «RKHS» (espace reproduisant) offre une grande liberté et adaptabilité aux utilisateurs des indices HSIC, et la possibilité d'insérer des connaissances a priori dans l'analyse par le choix d'un noyau particulier. Il a été montré que certains noyaux permettent de retrouver une décomposition d'Hoeffding et donc de combiner les avantages des deux classes d'indices. En revanche, il n'est pas encore clair*

que cette famille particulière de noyaux soit suffisamment large et englobe suffisamment de noyaux classiques pour que cette technique remplace celles utilisés actuellement.

Remarque 10. *De manière similaire au cadre «Équité Algorithmique», nous avons vu que le vivier d'indices accessibles pour obtenir de l'information est assez conséquent [IL15; Da +21]. Chaque ensemble d'indices permet de répondre à des questions particulières. Il serait donc naïf, dans une finalité d'application ou dans un but normatif (cf Section «Motivations historiques et juridiques») de penser qu'il suffit d'étudier un type d'indice unique pour régler une bonne fois pour toute la question de l'explicabilité, notamment lorsque l'on constate qu'un indice agrégé unique ne permet pas de rendre compte de la totalité d'une situation. De la même manière que pour une métrique d'équité, il est de mise d'être vigilant aux outils utilisés et aux réponses qu'ils apportent. De plus, nous avons vu et nous verrons dans le chapitre suivant que ces indices peuvent être appliqués à différents objets (e.g. sortie de l'algorithme, fonction de répartition de cette sortie, erreur entre la sortie et une valeur de référence...), menant à des nuances dans les réponses qu'ils apportent au praticien.*

1.3.4 Estimation de ces indices

Bien que l'objectif soit de trouver le bon indicateur permettant de répondre à la question posée, il est tout aussi important, en pratique et confronté aux données, d'être capable d'obtenir une bonne estimation de cet objet théorique. Dans cette sous-section, le but est d'explorer les différentes stratégies d'estimation existantes en GSA, ainsi que de fournir les garanties théoriques que l'on peut trouver dans la littérature pour ces estimateurs. Nous mettrons également en exergue la taille des échantillons de données nécessaires pour l'obtention pour l'estimation, montrant que tous les schémas ne sont pas équivalent.

La première stratégie, probablement la plus naïve, consiste à utiliser des méthodes de type Monte-Carlo. Cette estimation n'est pas l'apanage seul des indices basés sur la variance puisqu'on les retrouve notamment, en restant dans un cadre d'Analyse de Sensibilité, lors de l'utilisation des HSIC par exemple [Gre+]. Les indices de Sobol' étant en essence un ratio entre une variance conditionnelle et une variance, il suffit d'estimer chacune de ces quantités par un «plug-in» direct par les techniques classiques de Monte-Carlo. Les résultats de convergence dans ce registre sont similaires aux résultats classiques à cette littérature. On retrouve par exemple la convergence, ainsi que la normalité asymptotique de ces estimateurs. Dans certains contextes, par exemple pour l'analyse de sorties particulières (e.g. pour des indices «target-oriented» [DBM22]), il peut être intéressant de pondérer les distributions étudiées en utilisant de l'échantillonage préférentiel. Se faisant, il est possible d'obtenir plus efficacement l'information souhaitée sur les quantités d'intérêt, au prix d'une connaissance accrue sur les distributions sous-jacentes qui n'est pas toujours accessible. Cette méthode a été utilisée en combinaison avec des indices HSIC pour mener à la GSA de second niveau dont nous parlerons plus en détail dans un chapitre ultérieur, voir

Chapitre 4 et [MML19].

Le second type de schéma d'estimation que l'on peut rencontrer en GSA est appelé le «Pick and Freeze» [Gam+16]. Il part du constat suivant: en permutant les lignes d'un tableau de données contenant mes entrées, sauf pour la colonne correspondant spécifiquement à la variable dont nous voulons connaître l'influence, il est possible de mettre en exergue la dépendance entre cette variable et la sortie que nous avions. Si uniquement cette variable est pertinente pour la sortie, la réponse du modèle à ce nouveau dataset sera exactement la même que précédemment. Dans le cas contraire (aucune influence de la variable sur la sortie), le réponse du modèle à ce nouveau dataset sera une permutation des sorties observées avant. Dans le cadre classique d'indices de Sobol', il suffit donc de s'intéresser à la covariance entre les sorties fournies par le dataset non-modifié et par les sorties fournies par le dataset modifié avec la variable d'intérêt choisie («Pick») et gelée, non-permutée («Freeze»). Cette technique est d'un grand intérêt car elle ne nécessite que de permuter des lignes d'entrées, sans considération de dimension pour l'estimation d'un indice particulier, utilisant pour se faire $2n$ données, où n est le nombre d'individus dans le tableau de données. En revanche, la contrepartie de cette estimation à faible coût est la nécessité d'être capable de (re-)générer la sortie du modèle pour un nouveau vecteur d'entrées, et donc d'avoir accès à ce simulateur. Ce n'est pas toujours le cas et ce sera le principal désavantage de cette méthode face à l'estimation *«given-data»* présentée par la suite Dans ce schéma d'estimation, des performances similaires sont notées (convergence, normalité asymptotique) avec le coût réduit de $2n$ que nous avons mentionné, ce qui a poussé cette estimateur à être largement utilisé. On obtient la totalité des indices de Sobol' quantifiant l'influence directe de chaque variable pour un coût de $2n \times p$ appels au modèle. Enfin, l'estimation pour un groupe de variable se fait de la même manière en gelant toutes les variables incluses dans le groupe et en permutant les autres.

Plus récemment, une nouvelle stratégie d'estimation a vu le jour, toujours basée sur cette réécriture de la variance conditionnelle encodant l'influence comme un covariance. Proposée par Chatterjee [Cha20] comme un nouveau coefficient de corrélation, cette méthode permet une estimation en utilisant uniquement des comparaisons de rangs entre plus proches voisins, voir [Cha20; Gam+20] et Chapitre 4. Pour obtenir cette estimation, il n'est donc pas nécessaire de faire de nouveau appel au modèle, ce qui a mené la communauté à nommer cette estimation comme «*given data*». Cet estimateur est convergent et asymptotiquement normal, pour un coût impossible à battre (n appels). En revanche, si ce schéma fonctionne bien pour l'influence directe d'une variable particulière, des biais inévitables et difficilement corrigibles apparaissent en plus haute dimension, lorsque l'on cherche à quantifier l'influence d'un groupe de variables. C'est le principal défaut qu'à cette méthode face au «Pick and Freeze».

Remarque 11 (Estimation spectrale)**.** *Un dernier type d'approche pour l'estimation des indices basés sur la décomposition de la variance est l'approche spectrale [Da +21]. Le but est de décomposer le modèle f dans une base orthonormée dans laquelle les indices recherchés ont une forme close. L'exemple naturel*

est l'utilisation de la base de Fourier qui, combinée avec l'égalité de Parseval, offre une forme close et aisément estimable des indices voulus. Néanmoins, afin d'avoir une convergence rapide dans cette situation, il convient que les coefficients de Fourier associés aux hautes fréquences soient les plus faibles possibles, ce qui implique une régularité du modèle parfois irréaliste.

Chapter 2

Fairness seen as Global Sensitivity Analysis

2.1 Introduction

Quantifying the influence of a variable on the outcome of an algorithm is an issue of high importance in order to explain and understand decisions taken by machine learning models. In particular, it enables to detect unwanted biases in the decisions that lead to unfair predictions. This problem has received a growing attention over the last few years in the literature on fair learning for Artificial Intelligence. One of the main difficulty lies in the definition of what is (un)fair and the choices to quantify it. Numerous measures have been designed to assess algorithmic fairness, detecting whether a model depends on variables, called sensitive variables, that convey an information that is irrelevant for the model, from a legal or a moral point of view. We refer for instance to [Dwo+11; Cho17;] and [BGL20] and references therein for a presentation of different fairness criteria. Most of these definitions stem back to ensuring the independence between a function of an algorithm output and some sensitive feature that may lead to biased treatment. Hence, understanding and measuring the relationships between a sensitive feature S, which is typically included in $\mathbf{X}$ or highly correlated to it, and the output of the algorithm f using those features to predict a target Y, enables to detect unfair algorithmic treatments. Note that it is not enough to simply remove the sensitive feature from the data – "fairness through unawereness", [Bar+18] – as the algorithm can still "guess" the sensitive feature through its entanglement with the other inputs $\mathbf{X}$. Then, ensuring that predictors are fair is achieved by controlling previous measures, as done in [MCK; WM19; Gra+19; Bar+18; BGL20; Chi+20]. If this notion has been extensively studied for classification, recent work tackle the regression case as in [Gra+19; MCK; Chz+20] or [LLR20].

Global Sensitivity Analysis (GSA) is used in numerous contexts for quantifying the influence of a set of features on the outcome of a black-box algorithm.

Various indicators, usually taking the form of indices between 0 and 1, allow the understanding of how much a feature is important. Multiple set of indices have been proposed over the years such as Sobol' indices, Cramér-von-Mises indices, HSIC – see [JLD06; Da 15; IL15; Gra15; Gam+20] and references therein. The flexibility in the choice allows for deep understanding in the relationship between a feature and the outcome of an algorithm. While the usual assumption in this field is to suppose the inputs to be independent, some works [JLD06; MT12; Gra15] remove this assumption to go further in the understanding of the possible ways for a feature to be influential.
Hence, GSA appears to provide a natural framework to understand the impact of sensitive features. This point of view has been considered when using Shapley values in the context of fairness [FRF] and thus provide local fairness by explainability. Hereafter we provide a full probabilistic framework to use GSA for fairness quantification in machine learning.

Our contribution is two-fold. First, while GSA is usually concerned with independent inputs, we recall extensions of Sobol' indices to non-independent inputs introduced in [MT12] that offer ways to account for joint contribution and correlations between variables while quantifying the influence of a feature. We propose an extension of Cramér-von-Mises indices based on similar ideas. We also prove the asymptotic normality for these extended Sobol' indices to estimate them with a confidence interval, a novelty as far as we know. Then, we provide a consistent probabilistic framework to apply GSA's indices to quantify fairness. We illustrate the strength of this approach by showing that it can model classical fairness criteria, causal-based fairness and new notions such as intersectionality and provide insight for mitigating biases. This provides new conceptual and practical perspectives to fairness in Machine Learning.

The paper is organized as follows. We begin by reviewing existing works on Global Sensitivity Analysis (Section 2.2). We give estimates for the extended Sobol' and Cramér-von-Mises indices, along with a theorem proving asymptotic normality (Theorem 2.2.1). Then, we present a probabilistic framework for Fairness in which we draw the link between fairness measures and GSA indices, along with applications to causal fairness and intersectional fairness (Section 2.3).

2.2 Global Sensitivity Analysis

The use of complex computer models for the analysis of applications from science or real-life experiments is by now the routine. The models are often expensive to run and it is important to know with as few runs as possible the global influence of one or several inputs on the outcome of the system under study. When the inputs or features are regarded as random elements, and the algorithm or computer code is seen as a black-box, this problem is referred to as Global Sensitivity Analysis (GSA). Note that since we consider the algorithm to be a black-box, we only need the association of an input and its output. This make it easy to derive the influence of a feature for an algorithm for which we do not

have access to new runs. We refer the interested reader to [Da 15] or [IL15] and references therein for a more complete overview of GSA.

The main objective of GSA is to monitor the influence of variables $X_1, \cdots, X_p$ on an output variable, or variable of interest, $f(X)$. For this, we compare, for a feature X_i and the output $f(X)$, the probability distribution $\mathbb{P}_{X_i,f(X)}$ and the product probability distribution $\mathbb{P}_{X_i}\mathbb{P}_{f(X)}$ by using a measure of dissimilarity. If these two probabilities are equal, the feature X_i has no influence on the output of the algorithm. Otherwise, the influence should be quantifiable. For this, we have access to a wide range of indices, generally tailored to be valued in $[0, 1]$ and sharing a similar property: the greater the index, the greater the influence of the feature over the outcome. Historically, a variance-decomposition – or Hoeffding decomposition – is used of the output of the black-box algorithm to have access to a second-order moment metric in the so-called Sobol' method. However, these methods were originally developed for independent features. For obvious reasons, this framework is not adapted and has limitations in real-life cases. Additionally, Sobol' methods are intrinsically restrained by the variance-decomposition and others methods have been proposed. We will present two alternatives for Sobol' indices. The first one solves the issue of non-independent features. The second one circumvents the limitations of working with variance-decomposition. We finish this section by merging these two alternatives, inspired by the works of [AC21; Gam+20; Cha20].

Note that the use of other metrics is common in the GSA literature. Each metric has its own intrinsic advantages and disadvantages which have been extensively studied. Moreover, independence tests based on these GSA metrics exist, as shown in [Mey+; Gam+20] and techniques such as bootstrap or Monte-Carlo estimates can be used to obtain confidence intervals for such tests. We restrain ourselves to the Sobol' and Cramér-von-Mises indices because they are historically the basis of GSA literature, computationally tractable and allow for better understanding of usual fairness proxies, as we will show in Section 2.3. We also prove asymptotic normality for extended Sobol' indices, which is a first to the best of our knowledge.

2.2.1 Sobol' indices

A popular and useful tool to quantify the influence of a feature on the output of an algorithm are the Sobol' indices. Initially introduced in [Sob], these indices compare, thanks to the Hoeffding decomposition [Vaa98], the conditional variance of the output knowing some of the input variables with respect to the overall total variance of the output. Such indices have been extensively studied for computer code experiments.

Suppose that we have the relation $f(\mathbf{X}) = f(X_1, \cdots, X_p)$ where f is a square-integrable algorithm considered as a black-box and $X_1, \cdots, X_p$ inputs, with p the number of features. We denote by $p_{\mathbf{X}}$ the distribution of $\mathbf{X}$. For now, we suppose the different inputs to be independent, meaning that $p_{\mathbf{X}} = \otimes_{i=1}^{p} p_{X_k}$. Then, we can use the Hoeffding decomposition [Vaa98] on $f(\mathbf{X})$ – sometimes

also called ANOVA-decomposition – so that we may write

$$f(\mathbf{X}) = \sum_{s \subseteq [\![1,p]\!]} f_s(X_s), \tag{2.1}$$

where f_s are square-integrable functions and X_s the set $\{X_i, i \in s\}$. We can either assume that f is centered or that s can be the null set in this sum: it does not change anything since we are interested in the variance afterwards. We will consider $V := \mathrm{Var}(f(\mathbf{X}))$ and $V_s := \mathrm{Var}(f_s(\mathbf{X}_s))$. Note that the elements of the previous sum are orthogonal in the $L^2(p_{\mathbf{X}})$ sense. So, to compute the variance, we can compute it term by term, and obtain

$$V = \sum_{k=1}^{p} V_k + \sum_{k_2 > k_1}^{p} V_{k_1,k_2} + \cdots + V_{1,\cdots,p}. \tag{2.2}$$

This equation means that the total variance of the output, which is denoted by V, can be split into various components that can be readily interpreted. For instance, V_1 represents the variance of the output $f(\mathbf{X})$ that is only due to the variable X_1 – that is, how much $f(\mathbf{X})$ will change if we take different values for X_1. Similarly, $V_{1,2}$ represents the variance of the output Y that is only due to the combined effect of the variables X_1 and X_2 once the main effects of each variable has been removed – that is, how much $f(\mathbf{X})$ will change if we take different values simultaneously for X_1 and X_2 and remove the changes due to main effects from X_1 only or X_2 only.

By dividing the $V_{(m)}$ by V, with $(m) \subset [\![1,p]\!]$, we obtain:

$$S_{(m)} := \frac{V_{(m)}}{V}, \tag{2.3}$$

which is the expression of the so-called Sobol' sensitivity indices. When (m) is equal to a singleton k, the Sobol' index S_k quantifies the proportion of the output's variance caused by the input X_k on its own. The sum of all indices $S_{(m)}, k \in (m)$ quantifies the proportion of the output's variance caused by the input X_k conjointly with other inputs, and is usually called the Total Sobol' index of X_k and denoted ST_k.

Note that the law of total variance can be written for the random variable $f(\mathbf{X})$ as

$$\mathrm{Var}(f(\mathbf{X})) = Var(\mathbb{E}[f(\mathbf{X})|X_{\sim k}]) + \mathbb{E}[Var(f(\mathbf{X})|X_{\sim k})]. \tag{2.4}$$

In this equation, the left-hand side is the total variance, while the right-hand side is decomposed as two terms: the variance explained by all the variables different of X_k, and all the rest which include any part of variance explained by X_k. After normalization, we have

$$1 = S_{\sim k} + ST_k. \tag{2.5}$$

The alternate definition $ST_k = \frac{\mathbb{E}[Var(f(\mathbf{X})|X_{\sim k})]}{Var(f(\mathbf{X}))}$ is of interest for two reasons. First, we will see in the next section that this formulation can come back in

various contexts, including in Fairness. Additionally, for estimation, this formula is quite interesting since it allows estimation of the importance of a variable without using it directly, which may be in practice unfeasible for various reasons.

2.2.2 Sobol' indices for non-independent inputs

In the classic Sobol' analysis, for an input $f(\mathbf{X})$, two indices, namely the first order and total indices, quantify the influence of the considered feature on the output of the algorithm. When the inputs are not independent, we need to duplicate each index in order to distinguish whether influences caused by correlations between inputs are taken into account or not. Introduced in this framework by [MT12], we use the Lévy-Rosemblatt theorem to create two mappings of interest. We denote by $\sim i$ every index other than i. We create $2p$ mappings between p independent uniform random variables U and the variables $\mathbf{X}$ either by mapping $p_{U_1}p_{U_{\sim 1}}$ to $p_{X_i}p_{X_{\sim i}|X_i}$ – in this case U_1 is denoted by U_1^i – or by mapping $p_{U_{\sim p}}p_{U_p}$ to $p_{X_{\sim i}}p_{X_i|X_{\sim i}}$ – in this case, $U_{\sim p}$ is denoted $U_{\sim p}^{i+1}$. In the Appendix A.1, more in-depth details are given. In the analysis of the influence of an input X_i, the first mapping captures the intrinsic influence of other inputs while the second mapping excludes these influences and shows the variations induced by X_i on its own. Each of these two mappings leads to two indices corresponding to classical Sobol' and Total Sobol' indices. The influence of every input X_i is therefore represented by four indices, see Table 2.1.

Hence, the four Sobol' indices for each variable $X_i, i \in [\![1, p]\!]$ are defined as followed:

$$Sob_{X_i} := \frac{\mathrm{Var}[\mathbb{E}[f(\mathbf{X})|X_i]]}{\mathrm{Var}[f(\mathbf{X})]} \tag{2.6}$$

$$SobT_{X_i} := \frac{\mathbb{E}[\mathrm{Var}[f(\mathbf{X})|Z_i]]}{\mathrm{Var}[f(\mathbf{X})]} \tag{2.7}$$

$$Sob_{X_i}^{ind} := \frac{\mathrm{Var}[\mathbb{E}[f(\mathbf{X})|Z_i]]}{\mathrm{Var}[f(\mathbf{X})]} \tag{2.8}$$

$$SobT_{X_i}^{ind} := \frac{\mathbb{E}[\mathrm{Var}[f(\mathbf{X})|X_{\sim i}]]}{\mathrm{Var}[f(\mathbf{X})]}, \tag{2.9}$$

where the random variable Z_i has the distribution $p_{X_i|X_{\sim i}}$ and is equal to $F^{-1}_{X_i|X_{\sim i}}(U_p^{i+1})$. Note that we denote the Sobol' indices for X_i by the quantities S_i and ST_i under the assumption of independence, and by the quantities $Sob_{X_i}, SobT_{X_i}, Sob_{X_i}^{ind}, SobT_{X_i}^{ind}$ when this assumption is not fulfilled, for more clarity.

Note that these definitions can be extended to multidimensional variables and thus enabling to consider groups of inputs by replacing the subset $\{i\}$ by a subset $s \subset \{1, \cdots, p\}$ in the formulas. More insight on the transformations that allow these definitions can be found in Annex A.1 or in [MT12; MTA15].

Remark 1. *If the features are independent, then for all* $i \in [\![1, \cdots, p]\!]$, $Sob^{ind}_{X_i} = Sob_{X_i}$ *and* $SobT^{ind}_{X_i} = SobT_{X_i}$. *The proof comes from the fact that in the independent case, we have* $U^i_1 = U^{i+1}_p$.

Remark 2. *All previous indices satisfy the following bounds. For all* $i \in \{1, \cdots, p\}$,

$$0 \leq Sob^{ind}_{X_i} \leq Sob_{X_i} \leq SobT_{X_i} \leq 1 \quad \text{and} \quad 0 \leq Sob^{ind}_{X_i} \leq SobT^{ind}_{X_i} \leq SobT_{X_i} \leq 1.$$

We refer to [MT12] and to the law of total variance for the proof. Note that, in general, there are no inequalities between Sob_{X_i} *and* $SobT^{ind}_{X_i}$.

Sobol indices enable to quantify three typical ways for a feature to modify the output of an algorithm.

1. Direct contribution. Firstly, a variable can be of interest, all by itself, without any correlation or joint contribution with the other variables. Consider for example the case where $f(\mathbf{x}) = x_1 + x_2$ and x_1 independent to the rest of the variables. In this example, we would have $Sob_{X_1} = SobT_{X_1} = Sob^{ind}_{X_1} = SobT^{ind}_{X_1} = 0.5$, which means that 50% of the variability of the algorithm is caused by the first variable. In this case, the first variable has a non-null impact on its own on the outcome of the algorithm f.

2. Bouncing contribution. A variable can interact with other variables and influence the output only by its impact on the law of the other variables. For example, consider (x_1, x_2) where $x_2 = \alpha x_1 + \varepsilon$ – where ε is a centered white noise of variance σ^2 – and $f(\mathbf{x}) = x_2$. Then we get $Sob_{X_1} = SobT_{X_1} = (\alpha^2 V(x_1))/(\alpha^2 V(x_1) + \sigma^2)$ while $Sob^{ind}_{X_1} = SobT^{ind}_{X_1} = 0$. The first variable can be highly influent on the outcome of the algorithm f, even if it is not directly responsible for these variations. We call this type of interaction a "bouncing effect" since the variable will need to use another input to reach the outcome of the algorithm.

3. Joint contribution. Lastly, a variable can contribute to an output jointly with other variables. Take for instance the case where (x_1, x_2) are independent and $f(\mathbf{x}) = x_1 \times x_2$. In this case, $Sob_{X_1} = Sob^{ind}_{X_1} = 0 = Sob_{X_2} = Sob^{ind}_{X_2}$ while $SobT_{X_1} = SobT^{ind}_{X_1} = 1 = SobT_{X_2} = SobT^{ind}_{X_2}$ This effect is different of the previous one as the distributions of the input variables are independent but their impact is intertwined. In such a case, the effect is visible and measurable by a variation between first-order and total indices.

These main differences point out why we need four indices in order to assess the sensitivity of a system to a feature. Table 2.1 sums up which index takes correlations or joint contributions into account. The difference between these different indices can be very informative. For example, if the gap between Sob_{X_i} and $SobT_{X_i}$ or between $Sob^{ind}_{X_i}$ and $SobT^{ind}_{X_i}$ is big, then the feature X_i is mainly influential because of its joint contributions with the other features on the output. Conversely, if the gap between $Sob^{ind}_{X_i}$ and Sob_{X_i} or between

Table 2.1: Sobol' indices: what is taken into account and what is not.

Sobol' indices		
	Correlation between variables	Joint contributions
Sob_{X_i}	✓	✗
$SobT_{X_i}$	✓	✓
$Sob^{ind}_{X_i}$	✗	✗
$SobT^{ind}_{X_i}$	✗	✓

$SobT^{ind}_{X_i}$ and $SobT_{X_i}$ is big, a large part of the influence of the feature X_i will be through its intrinsic influence on other features.

Monte-Carlo estimation of the extended Sobol' indices can be computed by using this definitions. These estimators are consistent and converge to the quantities defined as the Sobol' and independent Sobol' indices earlier. Additionally, if we write each of these estimates as A_n/B_n, we can use the Delta-method theorem to prove a central limit theorem.

Theorem 2.2.1. *Each index $\mathcal{S}$ in the equations* (2.6) *to* (2.9) *can be estimated by its empirical counterpart $\mathcal{S}_n$ such that:*

(i) $\mathcal{S}_n \xrightarrow{a.s} \mathcal{S}$.

(ii) $\sqrt{n}(\mathcal{S}_n - \mathcal{S}) \xrightarrow{D} \mathcal{N}(0, \sigma^2)$, *with σ^2 depending on which index we study, see Appendix A.2.*

2.2.3 Cramér-von-Mises indices

Sobol' indices are based on a decomposition of the variance, and therefore only quantify influence of the inputs on the second-order moment of the outcome. Many other criteria to compare the conditional distribution of the output knowing some of the inputs to the distribution of the output have been proposed – by means of divergences, or measures of dissimilarity between distributions for example. We recall here the definition of Cramér-von-Mises indices [Gam+20], an answer to this lack of distributional information that will be of use later in a fairness framework – see Section 2.3.

Classical Cramér-von-Mises indices

The Cramér-von-Mises indices are based on the whole distribution of $f(\mathbf{X})$. They are defined (see [Gam+20]), for every input i, as follows:

$$CVM_i := \frac{\int_{\mathbb{R}} \mathbb{E}\left[(\mu(t) - \mu^i(t))^2\right] d\mu(t)}{\int_{\mathbb{R}} \mu(t)(1-\mu(t)) d\mu(t)} \tag{2.10}$$

where $\mu(t) := \mathbb{E}\left[\mathbb{1}_{f(\mathbf{X})\leq t}\right]$ is the cumulative distribution function of Y and μ^i its conditional version $\mu^i(t) := \mathbb{E}\left[\mathbb{1}_{f(\mathbf{X})\leq t}|X_i\right]$.

This equation can be rewritten as

$$CVM_i = \frac{\int \mathrm{Var}(\mathbb{E}\left[\mathbb{1}_{f(\mathbf{X})\leq t}|X_i\right])d\mu(t)}{\int \mathrm{Var}(\mathbb{1}_{f(\mathbf{X})\leq t})d\mu(t)}. \tag{2.11}$$

As before, these indices extend to the multivariate case. Simple estimators have been proposed [Cha20; Gam+20], and are based on permutations and rankings.

Remark 3. *As mentioned earlier, Sobol' indices quantify correlations and second-order moments but do not take into account information about the distribution of the outcome. However, note the similarity between the definition of the Cramér-von-Mises index and the classical Sobol' index, especially if we rewrite Equation* (2.11) *as:*

$$CVM_i = \int Sob_{X_i}(\mathbb{1}_{f(\mathbf{X})\leq t})\frac{Var(\mathbb{1}_{f(\mathbf{X})\leq t})}{\int Var(\mathbb{1}_{f(\mathbf{X})\leq t})d\mu(t)}d\mu(t). \tag{2.12}$$

Cramér-von-Mises can be seen as an adaptive Sobol' index that emphasizes the regions where the cumulative distribution of the outcome is highly changing, as more information can be obtained in these areas. This enables to capture information about the distribution of the outcome instead of moment-related information.

Extension of the Cramér-von-Mises indices

Classical Cramér-von-Mises indices suffer from the same limitation as Sobol' indices as they are tailored for independent inputs. A natural extension is to create new indices to handle the case of dependent inputs. We propose an extension of the Cramér-von-Mises indices, inspired by the ideas of the extended Sobol' indices and by the works of [AC21]. This new set of indices will capture the influence of a feature independently of the rest of the features.

Definition 2.2.1. *For every input i, we define the independent Cramér-von-Mises indices as:*

$$CVM_i^{ind} := \frac{\int \mathbb{E}(\mathrm{Var}(\mathbb{1}_{f(\mathbf{X})\leq t}|X_{\sim i}))d\mu(t)}{\int \mathrm{Var}(\mathbb{1}_{f(\mathbf{X})\leq t})d\mu(t)} \tag{2.13}$$

This extension enables to compare the influence of a feature on the output of an algorithm without its dependencies with other features.

Remark 4. *This independent Cramér-von-Mises index can be seen as an extension of the $SobT^{ind}$ index.*

This remark is similar to Remark 3. From the independent Total Sobol index shown in (2.9), by changing the output function as a threshold of the

real algorithm and taking the mean along all the possible thresholds, we obtain the independent Cramér-von-Mises index. This index can also be seen as an adaptive form of the $SobT^{ind}$ index.

Estimation of these indices is given in Appendix A.4 by the mean of estimates $\widehat{CVM}_i$. Similarly to Theorem 2.2.1, we have the following theorem.

Theorem 2.2.2. *If we denote by N the number of observations used to compute $\widehat{CVM}_i$, then the sequence $\sqrt{N}\left(CVM_i - \widehat{CVM}_i\right)$ converges towards the centered Gaussian law with a limiting variance ξ^2 whose explicit expression can be found in the proof.*

The proof of this theorem can be found in [GKL18]. Note that new estimation procedures can be efficient with little data, as mentioned in [Gam+20], which will be helpful for measuring intersectional fairness in the following Section.

2.3 Fairness

2.3.1 Sensitivity Indices as Fairness measures

In this section, we provide a probabilistic framework to unify various definitions of Fairness for Group of individual as Global Sensitivity Indices. Fairness amounts to quantify the dependencies between a sensitive feature S and functions of the outcome $f(X)$ and of the values of the variable of interest Y. Several measures of fairness corresponding to different definitions of fairness have been proposed in the machine learning literature. However, all these definitions boil back to a quantification of the mathematical propositions "$f(X) \perp\!\!\!\perp S$" or "$f(X) \perp\!\!\!\perp S|Y$".

For instance, the main common definitions of fairness are the following

- *Statistical Parity*, see for instance in [Dwo+11], requires that the algorithm f, predicting a target Y, has similar outputs for all the values of S in the sense that the distribution of the output is independent of the sensitive variable S, namely $f(\mathbf{X}) \perp\!\!\!\perp S$. In the binary classification case, it is defined as $\mathbb{P}(f(\mathbf{X}) = 1|S) = \mathbb{P}(f(\mathbf{X}) = 1)$ for general S, continuous or discrete.

- *Equality of odds* looks for the independence between the error of the algorithm and the protected variable, i.e implying here conditional independence, i.e $f(\mathbf{X}) \perp\!\!\!\perp S|Y$. This condition is equivalent in the binary case to $\mathbb{P}(f(\mathbf{X}) = 1|Y = i, S) = \mathbb{P}(f(\mathbf{X}) = 1|Y = i)$, for $i = 0, 1$.

- *Avoiding Disparate Treatment* correspond to the fact that similar individuals should have similar outputs. This condition, in the binary case, is written as $\mathbb{P}(f(\mathbf{X}) = 1|\mathbf{X} = \mathbf{x}, S = 0) = \mathbb{P}(f(\mathbf{X}) = 1|\mathbf{X} = \mathbf{x}, S = 1)$. Various refinements of this metric appears, including for instance the situation when similar individuals may not be sharing the same attributes $\mathbf{X} = \mathbf{x}$, *e.g* [Lar+21a].

- Finaly, *Avoiding Disparate Mistreatment* correspond to the equality of misclassification rates across subpopulations. This condition, in the binary case, is written as $\mathbb{P}(f(\mathbf{X}) \neq Y|S = 0) = \mathbb{P}(f(\mathbf{X}) \neq Y|S = 1)$.

Previous notions of fairness are quantified using a *Fairness measure* Λ and a function $\Phi(Y, \mathbf{X})$ such that $\Lambda(\Phi(Y, \mathbf{X}), S) = 0$ in the case of perfect fairness while the constraint is relaxed into $\Lambda(\Phi(Y, \mathbf{X}), S) \leq \varepsilon$, for a small ε, leading to the notion of approximate fairness. The following definition provides a general framework to define fairness measures. GSA measures as defined in 2.2 or described in [Da 15; IL15] are suitable indicators to quantify fairness as follows and these definitions can be extended to continuous predictors and continuous Y.

Definition 2.3.1. *Let Φ be a function of the features $\mathbf{X}$ and of Y. We define a GSA measure for a function Φ and a random variable Z as a $\Gamma(.,.)$ such that $\Gamma(\Phi(Y, \mathbf{X}), Z)$ is equal to 0 if $\Phi(Y, \mathbf{X})$ is independent of Z and is equal to 1 if $\Phi(Y, \mathbf{X})$ is a function of Z. Then, Γ induces a GSA-Fairness measure defined as $\Lambda(\Phi(Y, \mathbf{X}), S) = \Gamma(\Phi(Y, \mathbf{X}), S)$.*

The following examples provide a GSA formulation for most of classical fairness definitions using Sobol' and Cramér-von-Mises indices.

Example 1 (*Statistical Parity*)**.** *The so-called Statistical Parity fairness is achieved by taking $\Lambda(\Phi(Y, \mathbf{X}), S)) = Var(\mathbb{E}[f(\mathbf{X})|S])$. This corresponds to the GSA measure $Sob_S(f(\mathbf{X}))$. If f is a classifier with value in $\{0, 1\}$, we recover for a binary S the classical definition of Disparate Impact,$\mathbb{P}(f(X) = 1|S = 1) = \mathbb{P}(f(X) = 1|S = 0)$, see [Bar+18].*

Example 2 (*Avoiding Disparate Treatment*)**.** *The so-called Avoiding Disparate Treatment fairness is achieved by taking $\Lambda(\Phi(Y, \mathbf{X}), S)) = \mathbb{E}[Var(f(\mathbf{X})|X)]$. This corresponds to the GSA measure $SobT_S(f(\mathbf{X}))$. Note that it is normal for the algorithm not to be conditioned by the sensitive attribute for this GSA measure, cf Equation 2.4 Similarly, for a binary classifier, we recover the classical definition.*

Example 3 (*Equality of Odds*)**.** *The so-called Equality of Odds fairness is achieved by taking $\Lambda(\Phi(Y, \mathbf{X}), S)) = \mathbb{E}[Var(\mathbb{E}[f(\mathbf{X})|S, Y]|Y)]$. This corresponds to the GSA measure $CVM^{ind}(f(\mathbf{X}), S|Y)$. Similarly, for a binary classifier, we recover the classical definition.*

Example 4 (*Avoiding Disparate Mistreatment*)**.** *The so-called Avoiding Disparate Mistreatment fairness is achieved by taking $\Lambda(\Phi(Y, \mathbf{X}), S)) = Var(\mathbb{E}[\ell(f(\mathbf{X}), Y)|S])$ with ℓ a loss function. This corresponds to the GSA measure $Sob_S(\ell(f(\mathbf{X}), Y))$. Similarly, for a binary classifier, we recover the classical definition.*

Among well known fairness measures, we point out that we immediately recover two main fairness measures used in the fair learning literature – namely *Statistical Parity* and *Equality of Odds*. GSA measures can be computed for different function Φ and highlight either the behaviour of the algorithm, $\Phi(Y, \mathbf{X}) = f(\mathbf{X})$, or its performance, $\Phi(Y, \mathbf{X}) = \ell(Y, f(\mathbf{X}))$ for a given loss ℓ. This can lead to

different GSA-Fairness definitions from a same GSA measure, see Examples 1 and 4.

Example 5. *Recent work in Fairness literature exposed various definitions and measures to quantify influence of a sensitive feature, beyond classical notions. For instance, [FRF] uses Shapley values, [Li+19] uses HSIC measures, [GKK18] uses Mutual Information, so on and so forth. All these measures have been extensively studied in GSA literature, as mentioned in previous Section, and these frameworks are included in ours.*

For an additional example, consider the Hirschfeld-Gebelein-Rényi Maximum Correlation Coefficient [Rén59] – denoted HGR – which is defined for two random variables U and V as

$$HGR(U,V) = \sup_{f\in\mathbb{L}^2(\mathbb{P}_U),g\in\mathbb{L}^2(\mathbb{P}_V)} Corr(f(U),g(V)). \tag{2.14}$$

This index is used in [MCK] to quantify fairness and is linked to the Sobol' indices presented earlier as the alternate definition of this quantity given in [Rén59] can be written with Sobol' indices:

$$HGR(U,V) = \sup_{g\in\mathbb{L}^2(\mathbb{P}_V),\mathbb{E}[g]=0,\mathbb{E}[g^2]=1} \mathbb{E}[g(V)|U], \tag{2.15}$$

and therefore,

$$HGR(X_i,f(\mathbf{X})) = \sup_{g\in\mathbb{L}^2(\mathbb{P}_{f(\mathbf{X})})} Sob_{X_i}(g(f(\mathbf{X}))). \tag{2.16}$$

However, we restrain our study here to Sobol' indices mainly for two reasons. First, Sobol' indices are directly equivalent to very classical Fairness metrics, as we will see in the next Section. As such, using HGR is a valid choice as a proxy for Fairness but being fair with respect to HGR will be more difficult to obtain as a result that being fair with respect to Sobol'. Secondly, to compute the HGR index, it is necessary to compute a supremum of Sobol' indices over all the square-integrable functions. This additional operation leads to harder computation. A classical work-around is to approximate this quantity by restraining ourselves to some class by using Reproducing Kernel Hilbert spaces. The interested reader can find more information in [MCK].

In Table 2.2, we summarize the different indices associated to classical studied fairness definitions shown in previous Examples. By considering these fairness definitions as GSA measures, we can explain fairness in terms of simple effects presented in previous section, along with limitations of those definitions. For instance, *Statistical Parity* corresponds to the classical Sobol' index. The nullity of this index implies no direct influence of sensitive variables on the outcome, but can be limited as sensitive variables may have joint effects with other variables not captured by this metric. Therefore, *Statistical Parity* will lack in this regard. On the contrary, since *Avoiding Disparate Treatment* corresponds to Total Sobol' indices, this definition of fairness captures every possible influence of the sensitive feature on the outcome.

Table 2.2: Common fairness definitions and associated GSA measures

Fairness definition	GSA measure associated
Statistical Parity	$\text{Var}(\mathbb{E}[f(\mathbf{X})\|S]) \to Sob_S(f(\mathbf{X}))$
Avoiding Disparate Treatment	$\mathbb{E}[\text{Var}(f(\mathbf{X})\|X)] \to SobT_S(f(\mathbf{X}))$
Equality of odds	$\mathbb{E}[\text{Var}(\mathbb{E}[f(\mathbf{X})\|S,Y]\|Y)] \to CVM^{ind}(f(\mathbf{X}),S\|Y)$
Avoiding Disparate Mistreatment	$\text{Var}(\mathbb{E}[\ell(f(\mathbf{X}),Y)\|S]) \to Sob_S(\ell(f(\mathbf{X}),Y))$

Remark 5. *Note that many fairness measures are defined using discrete or binary sensitive variable. The GSA framework enables to handle continuous variables without additional difficulties. Moreover using kernel methods, GSA indices can be defined for a larger and more "exotic" variety of variables such as graphs or trees, for instance. In particular HSIC (see in [Da 15; BT04; Gre+; Smo+; Mey+]) is a kernel-based GSA measure that has been used in fairness.*

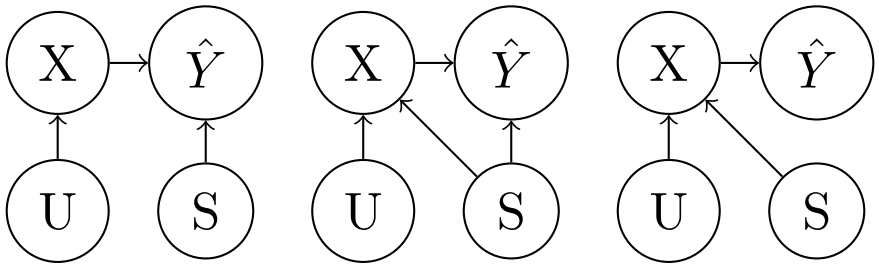
2.3.2 Consequences of seeing Fairness with Global Sensitivity Analysis optics

In this subsection, we enumerate various consequences of studying Fairness with this probabilistic framework coming from the GSA literature.

(i) **Modularity of fairness indicators** Numerous metrics have been proposed in GSA literature to quantify the influence of a feature on the outcome of an algorithm. We already mentioned several of them so far. This diversity enables choices in the quantified fairness since every choice of GSA measure induces a Fairness definition. We presented in previous subsection a concrete example with Sobol' indices, namely between *Disparate Impact* and *Avoiding Disparate Treatment*. Another example would be the use of kernels in HSIC-based indices, as exposed for instance in [Li+19]. By selecting various kernels, specific characteristics associated with fairness can be targeted.

(ii) **Perfect and Approximate fairness** GSA has been especially created to quantify *quasi* independence between variables. Merging GSA and Fairness gives a formal framework to the notion of approximate fairness and computationally justify the use of GSA codes to measure and quantify fairness. Additionally, as mentioned in previous section, GSA literature includes statistical tests for independence between input variables and outcomes, along with confidence intervals. Therefore, it is possible to compute them in order to test whether perfect fairness or approximate fairness is obtained. Moreover, this enables the possibility of auditing algorithms.

(iii) **Choice of the target** The framework presented earlier works for quantifying the influence of a sensitive feature on the outcome of a predictor



(a) Causal generative model in which X and S do not interact.

(b) Causal generative model in which a protected variable S influence both another variable X and the outcome $\hat{Y}$.

(c) Causal generative model in which a protected variable S influence the outcome $\hat{Y}$ only through another variable X.

Figure 2.1: Examples of representation of causal models with directed acyclic graphs.

but also any function of the predictor and of the input variables. This includes the loss of a predictor against a target. The ambivalence of this framework allows links to be made between various fairness definitions. For example, *Disparate Impact* and *Avoiding Disparate Mistreatment* are the same fairness but applied either to the predictor or to the loss of the predictor against a real target. In the first case, we want the algorithm to be independent of the sensitive feature; while in the second case, we want the errors of the predictor to be independent of the sensitive feature. Moreover, it allows for extension of fairness definitions to cases where an algorithm can be biased, as long as it does not make a mistake.

(iv) **Second-level Global Sensitivity Analysis** Recent works in GSA take into account the uncertainty of the distribution of the inputs of an algorithm, see [Mey+]. These tools can help in a fairness framework, especially when the distribution of sensitive features is unknown and unreachable. This will be more deeply studied in future papers.

2.3.3 Applications to Causal Models

Quantifying fairness using measures is a first step to understand bias in Machine Learning. Yet, causality enables to understand the true reasons of discrimination, as it is often related to the causal effect of a variable. The relations between variables describing causality are often modeled using a Directed Acyclic Graph (DAG). We refer to [Pea09; Bon+21].

In this subsection, we show how to address causal notions of fairness using the GSA framework, illustrated by a synthetic and a social example. We show that information gained thanks to Sobol' indices allow to learn some characteristic about the causal model.

We tackle the problem of predicting Y by $\hat{Y}$ knowing (X, S) while the non-sensitive variables are influenced by a non-observed exogeneous variable U. This is modeled by the following equations:

$$X = \phi(U, S) \quad \hat{Y} = \psi(X, S),$$

where ϕ and ψ are some unknown functions. These equations are a consequence of the unique solvability of acyclic models [Bon+21] and are illustrated in the various DAGs of Figure 2.1.

In many practical cases, the causal graph is unknown and we need indices to quantify causality. In the following, we are not interested in the complete knowledge of the graph – which is a NP-hard problem – but only in the existence of paths from S to Y.

Actually, GSA can quantify causal influence following DAG structure, and different GSA indices will correspond to different paths from S to Y. Different type of relationships can be measured in particular with the Total Sobol and the Total Independent Sobol indices to quantify either the presence of a path from S directly to Y or a path from S to another variable X that influences itself the predictor Y. We call this latter effect a "bouncing effect" since Y is influential only through a mediator.
The following proposition explains how specific Sobol indices can be used to detect the presence of causal links between the sensitive variable and the outcome of the algorithm.

Proposition 1 (Quantifying Causality with Sobol Index)**.**

- *The condition $SobT_S = 0$ implies that every path from S to Y is non-existent, that is S and Y belong to two different connected component of the causal graph.*
- *The condition $SobT_S^{ind} = 0$ implies that the direct path from S to Y is non-existent, that is the absence of direct edge between S and Y in the causal graph.*

Hence, using GSA, we can infer the absence of causal link between sensitive features and outcomes of algorithm without knowing the structure of the DAG. Note that, while Sobol' indices are correlation-based, this is not an issue in quantifying causality for fairness, as the sensitive features are usually supposed to be roots of the DAG [Bon+21; Lar+21a].

Example 6 (Causal graphs [Rot+20])**.** *In this example, we specify three causal models and illustrate the previous proposition.*

In Graph 2.1a, S is directly influent on the outcome $\hat{Y}$. There is no interaction between S and X. This happens when S and X are independent for instance. In such a case, Sobol' indices and independent Sobol' indices are the same, as mentioned in Remark 1. The equality $SobT_S = SobT_S^{ind}$ ensures the absence of "bouncing effect" for the sensitive variable S.

In Graph 2.1b, we have no information about the influence of S on the outcome.

In Graph 2.1c, S has no direct influence on the outcome, therefore $SobT_S^{ind} = 0$. This variable can still be influent on the outcome since it may modify other variables of interest. In this case, X is a mediator variable through which the sensitive feature will influence the outcome with a "bouncing effect". A model describing this kind of DAG in a fairness framework is the "College admissions" case, explained below.

Example 7 (College admissions)**.** *This example focus on college admissions process. Consider S to be the gender, X the choice of department, U the test score and $\hat{Y}$ the admission decision. The gender should not directly influence any admission decision $\hat{Y}$, but different genders may apply to departments represented by the variable X at different rates, and some departments may be more competitive than others. Gender may influence the admission outcome through the choice of department but not directly. In a fair world, the causal model for the admission can be modeled by a DAG without direct edge from S to $\hat{Y}$. Conversely, in an unfair world, decisions can be influenced directly by the sensitive feature S – hence the existence of a direct edge between S and $\hat{Y}$. This issue on unresolved discrimination is tackled in [Kil+17; FRF].*

It has been remarked in the literature that it is not easy to calculate causal-based fairness, especially when the joint distribution of mixed input conditional on continuous variables is hard to calculate from the observed data. When access to this joint distribution is not possible, recent works in GSA have proposed new estimation procedures [Gam+20] based on works by Chatterjee [Cha20]. These procedures makes no assumption on the distribution and provide a normally asymptotic estimation of GSA indices (and therefore associated Fairness metrics) at a low cost since it only require sorting of the data, along with the capacity to find closest neighbor of a data point.

2.3.4 Quantifying intersectional (un)fairness with GSA index

Most of fairness results are stated in the case where there is only one sensitive variable. Yet in many cases, the bias and the resulting possible discrimination are the result of multiple sensitive variables. This situation is known as intersectionality, when the level of discrimination of an intersection of several minority groups is worse than the discrimination present in each group as presented in [Cre]. Some recent works provide extensions of fairness measures to take into account the bias amplification due to intersectionality. We refer for instance to [Mor+20] or [Fou+19]. However, quantifying this worst case scenario cannot be achieved using standard fairness measures. The GSA framework allows for controlling the influence of a set of variables and as such can naturally address intersectional notions of fairness.

Intersectional fairness is obtained when multiple sensitive variables (for instance S_1 and S_2 in the most simple case) do not have any joint influence on the output of the algorithm. We propose a definition of intersectional fairness using GSA indices.

Definition 2.3.2. *Let $S_1, S_2, \cdots, S_m$ be sensitive features. It is said that an algorithm output is intersectionaly fair if $\Gamma(\Phi(X, S_1, \cdots, S_m); (S_1, \cdots, S_m)) = 0$. This constraint can be relaxed to $\Gamma(\Phi(X, S_1, \cdots, S_m); (S_1, \cdots, S_m)) \leq \varepsilon$ with ε small for approximate intersectionality fairness.*

Consider two independent protected features S_1 and S_2 (i.e gender and ethnicity). Depending on the chosen definition of fairness, there are situation where fairness is obtained with respect to S_1, with respect to S_2 but where the combined effect of (S_1, S_2) is not taken into account. For instance, let $Y = S_1 \times S_2$. In this toy-case, the Disparate Impact of S_1, as well as the Disparate Impact of S_2, is equal to 1 while the Disparate Impact of (S_1, S_2) is equal to 0. This can be readily seen thanks to the link between fairness and GSA as the Sobol' indices for S_1 and for S_2 are null while the Sobol' index for the couple (S_1, S_2) is maximal.

Proposition 2. *Let $(S_1, S_2, \cdots, S_m)$ be sensitive features. To be fair in the sense of Disparate Impact for S_1 and to be fair in the sense of Disparate Impact for S_2 does not quantify any intersectional fairness in the sense of the Disparate Impact.*

However, if we take again the same toy-case but look at the Total Sobol' indices, we see that $SobT_{S_1} = 0$ implies that $SobT_{(S_1,S_2)} = 0$.

Proposition 3. *Let $(S_1, S_2, \cdots, S_m)$ be sensitive features. To be fair in the sense of Avoiding Disparate Treatment for S_1 implies intersectional fairness for any intersection where S_1 appears.*

Remark 6. *Intersectional fairness is different than classical fairness. Classical fairness only pays attention to the influence of a single sensitive feature on the outcome while intersectional fairness is quantifying only the influence due to interactions between sensitive features. In applications, the goal is usually to have both classical and intersectional fairness. A single fairness definition that covers these two characteristics can be hard to find or too restrictive to readily use. For instance, among Sobol' indices, only the Total Sobol' index induces both a classical and intersectional fairness.*

2.4 Experiments

2.4.1 Synthetic experiments

In this subsection, we focus on the computation of complete Sobol' indices in a synthetic framework. We design three experiments, modeled after the causal generative models shown in Figure 2.1. For simplicity, we consider a Gaussian

model. In each experiment $j, j \in \{1, 2, 3\}$, (X, S, U) are random variables drawn from a Gaussian distribution with covariance matrix C_j, where

$$C_1 = C_2 = \begin{pmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0 \\ 0.5 & 0 & 1 \end{pmatrix}, C_3 = \begin{pmatrix} 1 & 0 & 0.5 \\ 0 & 1 & 0 \\ 0.5 & 0 & 1 \end{pmatrix}.$$

The random variable U is unobserved in this case and therefore does not have Sobol' indices. Its purpose is to simulate exogenous variables that modify the features in X. The target Y_j, described in the Table 2.3 for each of the experiments, is equal to

$$\begin{aligned} Y_1 &= 2 \times X, \\ Y_2 = Y_3 &= 0.7 \times X + 0.3 \times S. \end{aligned}$$

The first experiment shows the difference between independent and non-independent Sobol' indices. The outcome is entirely determined by a single variable X and therefore, $Sob_X = 1$. However, X is intrinsically linked with a sensitive feature because of the covariance matrix, so that $Sob_X^{ind} \neq 0$. This is a concrete example where *Statistical parity* is not obtained for S but *unresolved discrimination* mentioned in Example 7 is obtained, since S is influential only through X.

The second experiment adds a direct path from the variable S to the outcome Y. Since Y can be factorized as an effect from X and an effect of S, we still have $Sob_X = SobT_X$ and $Sob_X^{ind} = SobT_X^{ind}$. However, in this case, X is no longer enough to fully explain the outcome, so that $Sob_X \neq 1$. Sob_S^{ind} quantify the influence of this direct path from S to Y. Note that the difference between Sob_S and Sob_S^{ind} quantify the influence of the path from S to Y through the intermediary variable X.

In the third experiment, S and X are independent and S can only influence the outcome directly. This is the framework of classical Global Sensitivity Analysis. In this case, non-independent and independent Sobol' indices are equal, as mentioned in Remark 1

Note that for these synthetic examples, we have complete access to the joint law of the input variables. In such a case, we can apply the estimation schemes described in Appendix A.2 directly. The code can be found in the following repository GIT.

2.4.2 Real data sets

In this section, we focus on the implementation of Cramér-von-Mises indices on two real-life datasets: the Adult dataset [DG17] and the COMPAS dataset.

For real data sets, we first need to preprocess our data. For the Adult dataset, we applied to the data the same preprocessing as the one described in [Bes+20]. As for the Compas dataset, we used the same preprocessing as [Zaf+17]. Additionally, since access to the joint law of the distribution is not

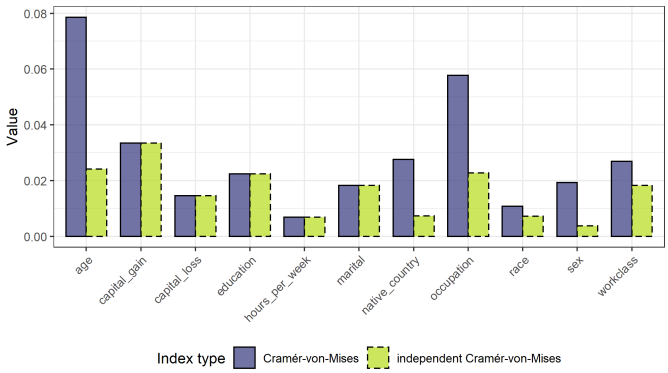
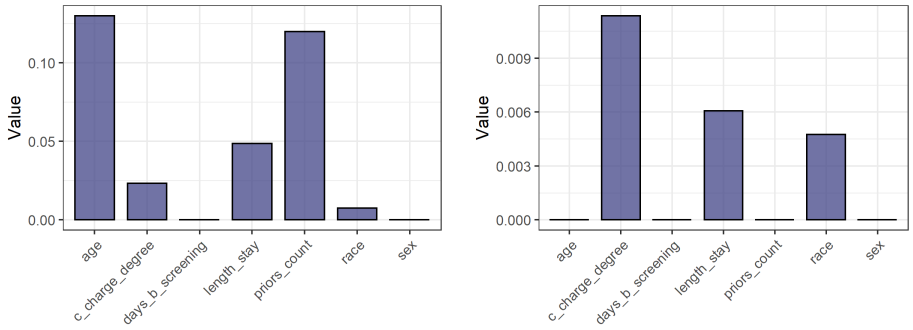


Figure 2.2: Cramér-von-Mises and independent Cramér-von-Mises indices for the Adult dataset.

accessible, we added noise to the binary data to make them continuous and used a Gaussian approximation for the copula, as described in [MTA15], in order to have tractable estimates.

Adult dataset

The Adult dataset consists in 14 attributes for 48,842 individuals. The class label corresponds to the annual income (below/above 50.000 k\$). We study the effect of different attributes. The results for a classifier obtained for an algorithm built using an Extreme Gradient Boosting Procedure are shown in Figure 2.2. We used the same pre-process as [Bes+20] for the choice of variables.
If we look at the independent Cramér-von-Mises, we quantify the direct influence of a variable . We recover the influent indicators – "capital gain", "education-number", "age", "occupation"... – given by other studies [FRF; Bes+20].
The joint influences on the outcome of other variables is also measured using GSA indices. Variables for which independent and classical Cramér-von-Mises indices are the same have no "bouncing" influence. Otherwise, the gap between these two indices quantify this specific effect. For example, the variable "age" correlates with most of the other variables such as "education-number" or "marital-status" for instance. Because of this, most of its influence is through "bouncing effects" and the gap between its two indices (i.e "CVM" and "CVM_indep") is larger than for any other feature. The variable "sex" also plays an important role through its "bouncing" effect. We can see this through the difference between the classical and the independent index associated with this feature. This explains why removing the variable "sex" is not enough to obtain a fair predictor since it influences other variables that affect the prediction. We recover the results obtained by several studies that point out the bias created by the "sex" variable. Note that race may have led to unbalanced decisions as well. Yet, the Cramér-von-Mises index is lower than the one for the "sex" variable, which explains



(a) Cramér-von-Mises indices computed for the COMPAS decile score.

(b) Cramér-von-Mises indices computed on the loss between COMPAS output and real case of recidivism after two years.

Figure 2.3: Cramér-von-Mises indices for the COMPAS dataset.

why the discrimination is lower than the one created by the sex, as emphasized by the study of the Disparate Impact which is in a 95% confidence interval of $[0.34, 0.37]$ for sex and $[0.54, 0.63]$ for ethnic origin in [Bes+20].

COMPAS dataset

The so-called COMPAS dataset, gathered by ProPublica described for instance in in [Was19] , contains information about the recidivism risk predicted by the COMPAS tool, as well as the ground truth recidivism rates, for 7214 defendants. The COMPAS risk score, between 1 and 10 (1 being a low chance of recidivism and 10 a high chance of recidivism), is obtained by an algorithm using all other variables used to compute it, and is used to forecast whether the defendant will reoffend or not. We analysed this dataset with Cramér-von-Mises indices in order to quantify fairness exhibited by the COMPAS algorithm. The preprocessing we used is the same as the one described by [Zaf+17]. The results are shown in Figure 2.3.

First, every independent index is null, which means that the COMPAS algorithm does not rely on a single variable to predict recidivism. Also, gender and ethnicity are virtually not used by the algorithm, opposed to the variables "age" or "$priors\_count$" (the number of previous crimes). Hence as expected, the algorithm appears to be fair. However, when comparing the accuracy of the predictions of the algorithm with real-life two-year recidivism, the "race" variable is found to be influential. Hence we show that the indices we propose recover the bias denounced by Propublica with an algorithm that, despite fair predictions, shows a behavior that favors a part of the population based on the race variable.

2.5 Conclusion

We recalled classical notions both for the Global Sensitivity Analysis and the Fairness literature. We presented new Global Sensitivity Analysis tools by the mean of extended Cramér-von-Mises indices, as well as proved asymptotic normality for the extended Sobol' indices. These sets of indices allow for uncertainty analysis for non-independent inputs, which is a classical situation in real-life data but not often studied in the literature. Concurrently, we link Global Sensitivity Analysis to Fairness in an unified probabilistic framework in which a choice of fairness is equivalent to a choice of GSA measure. We showed that GSA measures are natural tools for both the definition and comprehension of Fairness. Such a link between these two fields offers practitioners customized techniques for solving a wide array of fairness modeling problems.

Table 2.3: Synthetic experiments based on causal DAGs – Figure 2.1

.

	Sob	$SobT$	Sob^{ind}	$SobT^{ind}$
$Y = 2 \times X$				
X	**1.00** (0.99 - **1.00** - 1.00)	**1.00** (0.99 - **1.00** - 1.00)	**0.75** (0.74 - **0.75** - 0.76)	**0.75** (0.74 - **0.75** - 0.76)
S	**0.24** (0.24 - **0.25** - 0.26)	**0.25** (0.24 - **0.25** - 0.26)	**0.00** (0.00 - **0.00** - 0.01)	**0.00** (0.00 - **0.00** - 0.01)
$Y = 0.7 \times X + 0.3 \times S$				
X	**0.91** (0.89 - **0.91** - 0.93)	**0.92** (0.89 - **0.91** - 0.94)	**0.51** (0.46 - **0.48** - 0.52)	**0.52** (0.46 - **0.47** - 0.54)
S	**0.52** (0.48 - **0.53** - 0.55)	**0.54** (0.48 - **0.53** - 0.55)	**0.07** (0.05 - **0.09** - 0.11)	**0.09** (0.06 - **0.09** - 0.12)
$Y = 0.7 \times X + 0.3 \times S$				
X	**0.78** (0.78 - **0.84** - 0.85)	**0.84** (0.80 - **0.84** - 0.86)	**0.81** (0.78 - **0.84** - 0.85)	**0.82** (0.80 - **0.84** - 0.86)
S	**0.13** (0.12 - **0.16** - 0.17)	**0.17** (0.15 - **0.16** - 0.18)	**0.14** (0.12 - **0.16** - 0.17)	**0.15** (0.13 - **0.16** - 0.18)

Legend: Values format is "**experimental value** (lower bound of 95% confidence interval - **theoretical value** - upper bound of 95% confidence interval)".

Chapter 3

Of the use of metamodels in GSA and Fairness

3.1 Introduction

As mentioned in Chapter 1 and Chapter 2, GSA indices are a way to quantify how much influence a given input has on an outcome when no particular information is available about the architecture of the used algorithm. The last decade has seen numerous breakthrough in this field. While, at first, only Sobol' indices were available [Sob; Sal+10], various additions have been proposed such as, but not limited to, HSIC indices through the use of kernels [Gre+; Da 21], Cramér-von-Mises indices [GKL18] or even Shapley indices [OP17], leveraging results from the game theory community. Conjointly, Group Fairness metrics [Dwo+11; MCK; BGL20] are used in order to assess whether a model presents discriminatory biases towards parts of the population, by quantifying the influence of sensitive variables. We showed that it is possible to quantify Group Fairness through the choice of a fairness metric ; and that several of these metrics are in fact deeply linked to tools used in the Global Sensitivity Analysis literature. However, both in the Explainability and the Fairness framework, computation of GSA indices and fairness metrics are done using direct access to the model. Indeed, for estimation of these indicators, we need to be able to sample using the algorithm f.

In practice, models are often too costly to run intensively due to the use of expensive techniques, or outright impossible to directly access. In this case, we must be able to infer properties exhibited by the model (e.g. fairness, explainability) without additional calls to the algorithm. In order to do this, we use metamodels [JNP14]. Metamodels are more and more used by the scientific community as these objects aim at depicting a truthful representation or approximation of a true algorithm, at a lower cost. Usually, the metamodel is the result of an optimization, minimizing an empirical loss involving data coming from the true algorithm. While various architectures exist for this problem – for

instance based on wavelets [DC17], Gaussian Processes [LMS17; Bac13; BBG19] ... – it is important to quantify how truthful this metamodel is. This is usually done by the quantification of the mean square error made by the metamodel, that is the quantification of the $\mathbb{L}_2$ norm of the difference between true and approximating algorithms. The statistical community have been interested in this subject for a long time and results can be found in the literature, see [Bar02; WJ21].

In these premises, the natural question we aim at answering by the end of this chapter is the following: *"Can we infer properties related to an inaccessible model only by looking at a corresponding metamodel?"* We provide hints and results in this direction both for the Explainability and the Fairness frameworks.

To the best of our knowledge, only a handful of papers merge these subjects to try and answer the question above. In a context of GSA, [JNP14] provides an answer for Sobol' indices in the classical GSA framework, that is to say when inputs are considered independent. The authors prove a Central Limit Theorem for estimators of these indices, under very mild conditions. More recently, [Pan21] proves a tight bound in the same framework, linking the differences of Sobol' indices to the mean square error previously mentioned. We build upon these two works to provide a similar bound for multiple variants of Sobol' indices in a non-independent framework, and derive asymptotic rates for wavelet-based metamodels [DC17] and Gaussian Process metamodels [LMS17; Bac13; BBG19].

Finally, when translating this question in a Fairness framework, we are able to assess whether a model is fair with respect to a chosen metric, without having access to this algorithm. This problematic is known as audit and is now a prevalent issue in the Algorithmic Fairness literature, partly because of recent judicial advances, see [LWM21]. Using the results of this chapter, we are able to offer a road-map to audit fairness on a model, only through information gained by using a metamodel. This proves to be useful for auditors that need to certify the fairness of an auditee's model, and alleviate potential trust issues and reluctance from the auditee to provide direct access to a potentially proprietary model. This is done however at a higher cost, due to the intrinsic loss of information incurred while auditing an algorithm through a proxy.

3.2 Usual Sobol'-based GSA indices

The main goal here is to introduce the GSA indices we use after: the Sobol'-based indices. In this nomenclature, we include Sobol' indices, Extended Sobol' indices, Extended Cramér-von-Mises indices and Shapley indices. The first two set of indicators, Sobol' and Extended Sobol' indices, have been defined in Section 2.2, along with useful properties and insights. However, Extended Cramér-von-Mises indices were only partially defined because of the link with some fairness metrics. Moreover, Shapley indices were not defined and its link with Sobol' indices was not made explicit. Therefore, in this section, we provide the definitions and some properties for these missing quantities that will be useful later.

3.2.1 Definition of Extended Cramér-von-Mises indices

Sobol' indices are variance-based. This proves to be a limitation and has been the main motivation behind the creation of Cramér-von-Mises indices [GKL18]. However, as far as the authors know, extensions of the Sobol' indices have not yet been transposed for Cramér-von-Mises indices. Further generalisations to generic metric spaces, along the lines of [Gam+21], are possible but will be investigated in a later work.

Definition 3.2.1. *Classical Cramér-von-Mises indices are defined as*

$$CvM_i(f) = \frac{\int \mathbb{E}\left[(\mathbb{E}[\mathbb{1}_{f(\mathbf{X})\leq t}] - \mathbb{E}[\mathbb{1}_{f(\mathbf{X})\leq t}|X_i])^2\right] dt}{\int \mathrm{Var}(\mathbb{1}_{f(\mathbf{X})\leq t})dt}. \tag{3.1}$$

Note that Cramér-von-Mises indices can be rewritten as

$$CvM_i(f) = \int Sob_i(\mathbb{1}_{f(\mathbf{X}\leq t)}) \times \frac{\mathrm{Var}(\mathbb{1}_{f(X)\leq t})}{\int \mathrm{Var}(\mathbb{1}_{f(X)\leq t})dt} dt. \tag{3.2}$$

By rewriting the definition of Cramér-von-Mises indices in such a fashion, we can define extensions of Cramér-von-Mises by replacing the Sobol' part in the previous equation by one of the extended Sobol' index.

Definition 3.2.2. *Extended Cramér-von-Mises indices are defined as*

$$CvM_i(f) = \int Sob_i(\mathbb{1}_{f(\mathbf{X}\leq t)}) \times \frac{\mathrm{Var}(\mathbb{1}_{f(X)\leq t})}{\int \mathrm{Var}(\mathbb{1}_{f(X)\leq t})dt} dt, \tag{3.3}$$

$$CvMT_i(f) = \int SobT_i(\mathbb{1}_{f(\mathbf{X}\leq t)}) \times \frac{\mathrm{Var}(\mathbb{1}_{f(X)\leq t})}{\int \mathrm{Var}(\mathbb{1}_{f(X)\leq t})dt} dt, \tag{3.4}$$

$$CvM_i^{ind}(f) = \int Sob_i^{ind}(\mathbb{1}_{f(\mathbf{X}\leq t)}) \times \frac{\mathrm{Var}(\mathbb{1}_{f(X)\leq t})}{\int \mathrm{Var}(\mathbb{1}_{f(X)\leq t})dt} dt, \tag{3.5}$$

$$CvMT_i^{ind}(f) = \int SobT_i^{ind}(\mathbb{1}_{f(\mathbf{X}\leq t)}) \times \frac{\mathrm{Var}(\mathbb{1}_{f(X)\leq t})}{\int \mathrm{Var}(\mathbb{1}_{f(X)\leq t})dt} dt. \tag{3.6}$$

Remark 7. *Cramér-von-Mises indices can be seen as an adaptive mean of the Sobol' indices on level-lines of the cumulative distribution of $f(X)$, or "failure domains" defined by $\mathbb{1}_{f(X)\leq t}$. The use of cumulative distributions allows to capture information on the complete distribution of the output and not simply a second-moment information.*

This average can be understood as following. When $Var(\mathbb{1}_{f(X)\leq t})$ is high, then this area is highly informative and the influence of all the inputs should be monitored with attention. Conversely, when $Var(\mathbb{1}_{f(X)\leq t})$ is low, the algorithm response is almost constant. Influence from a feature in this area will not be as informative and as effective and therefore count less.

While the definitions of these indices are straightforward when the link between Cramér-von-Mises and Sobol' is exposed, this is, to the best of our knowledge, the first time they are formalized. The use of the classical Cramér-von-Mises indices is by now common practice but use of the three other indices are yet to be found, with an exception for CvM^{ind} in a special variant in the fairness framework, see [Bén+21].

3.2.2 Definition of Shapley indices

Shapley indices, introduced in Global Sensitivity Analysis by [Owe14; OP17], are a widely used mean to quantify influence with non-independent inputs, especially for industrial cases. This notion comes from the Game Theory and proposes to compute the contribution of an input variable in various "coalitions". When the inputs are independent, its signification is well-understood as it is a weighting of Sobol' indices based on the number of elements in the coalition.

In order to define the Shapley indices, we first need to introduce the *closed Sobol' indices*. In the literature [Da +21; IL15], Sobol' indices are sometimes defined, for a general set A of inputs as

$$S_A^{clos} = \frac{\mathrm{Var}(\mathbb{E}[f(\mathbf{X})|X_A])}{\mathrm{Var}(f(\mathbf{X}))}. \tag{3.7}$$

In fact, this definition S_A^{clos} coincides with S_A for singletons but differ for groups of inputs. To reconcile these two notions, we need to make use of the inclusion-exclusion principle, a special case of the Möbius inversion formula, to obtain

$$S_A = \sum_{B \subset A} (-1)^{|A|-|B|} S_B^{clos} \tag{3.8}$$

$$S_A^{clos} = \sum_{B \subset A} S_B \tag{3.9}$$

In general, Sobol' indices used in the literature correspond to S_A and S_A^{clos} is used for the definition of Shapley indices, as we will see later.

Definition 3.2.3. *Let $\mathbf{X} = (X_1, \ldots, X_p)$ be inputs and f be a square-integrable function. We define the Shapley index as*

$$Sh_i(f) = \frac{1}{p} \sum_{A \subset \sim i} \binom{|A|}{p-1}^{-1} \left(S_{A \cup \{j\}}^{clos} - S_A^{clos} \right). \tag{3.10}$$

Remark 8. *If the inputs are independent, then we have the equality*

$$Sh_i(f) = \sum_{j \in A} \frac{S_A(f)}{|A|}. \tag{3.11}$$

This equality has been one of the main motivation for the use of Shapley indices as an extension of Sobol' indices in the case of dependent inputs.

While previous definition is the most used, we introduce an equivalent definition [Deh17] that will allow for easier proofs of later results.

Definition 3.2.4. *Let Π_p be the set of all permutations of $\{1,\dots,p\}$. For $\pi = (i_1,\dots,i_p) \in \Pi_p$, we define π^i as $(i_1,\dots,i_k = i)$. We can then define the Shapley index as*

$$Sh_i(f) = \frac{1}{n!}\sum_{\pi\in\Pi_p}\left(S^{clos}_{\pi^i} - S^{clos}_{\pi^i\setminus\{i\}}\right). \tag{3.12}$$

This definition provides with an other understanding of Shapley values: the Shapley index of the variable X_i is the mean additional influence of a coalition coming from the input X_i when taking the uniform distribution among all manners of choosing a coalition. However, as mentionned earlier, Shapley alone cannot give the full picture and while its estimation has been well documented, its use should be complemented by other GSA indices [Da +21; FRF; ICI21].

3.3 Upper-bounds for GSA indices between two models

The use of metamodels – or surrogate models – are by now the routine when computer codes are too expensive or inaccessible. They allow to have an approximate result, trading accuracy for a lower computational cost. Often, it is also easier to compute GSA indices on these metamodels than it is on the original function or algorithm f, as access to new runs can be virtually free. However, a natural question is the following: *"If $\hat{f}$ is an approximation of f, how close are the GSA indices of $\hat{f}$ to the GSA indices of f?"*

This question has partially been answered in the special case of Sobol' indices for independent inputs in [JNP14] with asymptotic results and in [Pan21], with this theorem that will be the cornerstone of our extensions and proofs:

Theorem 3.3.1 ([Pan21])**.** *Let f and $\hat{f}$ be two square-integrable functions. Then, for any subset $A \subset \{1,\cdots,d\}$,*

$$|S_A(f) - S_A(\hat{f})| \leq \left(\sqrt{S_A(f)(1-S_A(\hat{f}))} + \sqrt{(1-S_A(f))S_A(\hat{f})}\right)\frac{\|f-\hat{f}\|_{2,\mathbf{X}}}{\mathrm{Var}_{\mathbb{P}_\mathbf{X}}(f)^{1/2}}. \tag{3.13}$$

We extend these results to the wider class of GSA indices that we call "Sobol'-based indices" and that we have defined in previous section. Besides generalisation to other types of indices, our results allow to circumvent the restrictive assumption of independence between inputs.

First, we prove results for fixed functions f and $\hat{f}$. Then, we complete this answer with the case of metamodels chosen from a random design, or simply put, when we use data to choose a $\hat{f}_n$. The main take-away message from these results is that the key component of the upper-bounds is either the $\mathbb{L}^2(\mathbb{P}_\mathbf{X})$-norm of $f-\hat{f}$ in the deterministic framework or the approximation risk for the random design.

3.3.1 Upper-bounds when working with deterministic models

In this subsection, we provide extensions of Theorem 3.3.1 for all indices defined in the previous section. While we suppose for now f and $\hat{f}$ to be given, we make no assumptions on the input distribution. We show that the gap between the GSA indices of two models is upper-bounded by a term involving the $\mathbb{L}^2(\mathbb{P}_{\mathbf{X}})$-norm of $f - \hat{f}$. We format these results in Table 3.1 for the sake of clarity.

Theorem 3.3.2. *Let f and $\hat{f}$ be two functions of $\mathbb{L}^2(\mathbb{P}_{\mathbf{X}})$. Let X_i be an input and let $GSA_i(f)$ be one of the extended Sobol' index for the input variable X_i, that is $GSA_i(f) \in \{Sob_i(f), Sob_i^{ind}(f), SobT_i(f), SobT_i^{ind}(f)\}$. Then*

$$|GSA_i(f) - GSA_i(\hat{f})| \leq \frac{\|f - \hat{f}\|_{2,\mathbf{X}}}{\mathrm{Var}_{\mathbb{P}_{\mathbf{X}}}(f)^{1/2}}. \tag{3.14}$$

Remark 9. *While in practice, we use Equation 3.14, we can in fact obtain a tighter bound. Indeed, we have the following inequality:*

$$|GSA_i(f) - GSA_i(\hat{f})| \leq \left(\sqrt{GSA_A(f)(1 - GSA_A(\hat{f}))} + \sqrt{(1 - GSA_A(f))GSA_A(\hat{f})}\right) \frac{\|f - \hat{f}\|_{2,\mathbf{X}}}{\mathrm{Var}_{\mathbb{P}_{\mathbf{X}}}(f)^{1/2}}. \tag{3.15}$$

However, this additional multiplicative constant is always upper-bounded by 1 and involves the term $GSA(f)$, precisely what we aim to approach by using $GSA(\hat{f})$.

This theorem proves that we can bound the error made when comparing the Sobol' indices of the real algorithm and the Sobol' indices of a metamodel by the approximation error of the metamodel. As mentioned in previous section, this can allow some new strategies for estimating Extended Sobol' indices of f, if these indices can be easily and analytically computed for some approximation $\hat{f}$ of f.

Corollary 1. *Let f and $\hat{f}$ be two continuous functions of $\mathbb{L}^2(\mathbb{P}_{\mathbf{X}})$. Let X_i be an input and let $GSA_i(f)$ be one of the extended Cramér-von-Mises index for the input variable X_i, that is $GSA_i(f) \in \{CvM_i(f), CvM_i^{ind}(f), CvMT_i(f), CvMT_i^{ind}(f)\}$. Then*

$$|GSA_i(f) - GSA_i(\hat{f})| \leq \|f - \hat{f}\|_{2,\mathbf{X}}^{1/2}. \tag{3.16}$$

This theorem proves a similar result for the Cramér-von-Mises indices.

Corollary 2. *Let f and $\hat{f}$ be two functions of $\mathbb{L}^2(\mathbb{P}_{\mathbf{X}})$. Let X_i be an input and let $GSA_i(f)$ be equal to the Shapley index for the input variable X_i, $GSA_i(f) = Sh_i(f)$. Then*

$$|GSA_i(f) - GSA_i(\hat{f})| \leq 2\frac{\|f - \hat{f}\|_{2,\mathbf{X}}}{\mathrm{Var}_{\mathbb{P}_{\mathbf{X}}}^{1/2}(f)}. \tag{3.17}$$

Table 3.1: Upper-bounds for the various used GSA indices.

GSA index	Associated upper-bound (Deterministic)	Associated upper-bound (Data-conditioned)
Extended Sobol' indices	$\frac{\|f-\hat{f}\|_{2,\mathbf{X}}}{\mathrm{Var}^{1/2}(f)}$	$\frac{\mathbb{E}\|f-\hat{f}\|^2}{\mathrm{Var}(f)}$
Extended Cramér-von-Mises indices	$\|f-\hat{f}\|_{2,\mathbf{X}}^{1/2}$	$\mathbb{E}\|f-\hat{f}\|$
Shapley indices	$2 \times \frac{\|f-\hat{f}\|_{2,\mathbf{X}}}{\mathrm{Var}^{1/2}(f)}$	$2 \times \frac{\mathbb{E}\|f-\hat{f}\|^2}{\mathrm{Var}(f)}$

One may remark that the bound obtained for the Cramér-von-Mises indices are not of the same order than the other bounds. This occurs because of the use of the cumulative distributions of $f(\mathbf{X})$ and $\hat{f}(\mathbf{X})$ instead of directly working with the functions. While this is a good thing for the definition of Cramér-von-Mises indices, it leads to a degradation of this upper-bound. We are not sure yet that a better bound is achievable.

3.3.2 Risk bounds for a metamodel chosen from a random design.

So far, we worked with deterministic functions. However, in practice, it is common to choose the approximation $\hat{f}$ of the true model f according to some rules. The most classical framework would be to choose $\hat{f}$ as the minimizer of some empirical quadratic loss, possibly penalised. We show here that we keep upper-bounds on the quadratic risk for the GSA indices defined earlier when working with metamodels, in a very general framework. The main take-away from these theorems is that the main component of the risk bounds is the quadratic approximation loss.

We now suppose that we have a random design $\mathcal{D}$, along with noisy output $Y = f(\mathcal{D}) + \varepsilon$ from which we build a function $\hat{f}$ to estimate the true function f. Let $\mathcal{D} = \{\mathbf{x}_i\}_{i=1,\dots,n}$ be an i.i.d. sample from the distribution $\mathbb{P}_{\mathbf{X}}$ and $\mathcal{L}$ be a deterministic learning procedure $\mathcal{L} : (\mathcal{D}, Y = f(\mathcal{D}) + \varepsilon) \mapsto \hat{f}$ that build from data a metamodel approximating f. Common choices can be a projection method or ordinary least square for instance, but such metamodels can also come from the minimisation of a penalized loss on a more complicated collection of functions, as we will see in next section. However, we show here that we can still bound the risk by the risk of approximation or quadratic loss.

Theorem 3.3.3. *Let f be a function of $\mathbb{L}^2(\mathbb{P}_{\mathbf{X}})$. Let $\hat{f}$ be the approximation of f obtained through a learning procedure $\mathcal{L}$.*

1. *If $GSA_i(f) \in \{Sob_i(f), Sob_i^{ind}(f), SobT_i(f), SobT_i^{ind}(f)\}$, then*

$$\mathbb{E}(GSA(f) - GSA(\hat{f}))^2 \leq \frac{\mathbb{E}\|f - \hat{f}\|^2}{\mathrm{Var}(f)}. \tag{3.18}$$

2. *If $GSA_i(f) \in \{CvM_i(f), CvM_i^{ind}(f), CvMT_i(f), CvMT_i^{ind}(f)\}$, then*

$$\mathbb{E}(GSA(f) - GSA(\hat{f}))^2 \leq \mathbb{E}\|f - \hat{f}\|. \tag{3.19}$$

3. *If $GSA_i(f) = Sh_i(f)$, then*

$$\mathbb{E}(GSA(f) - GSA(\hat{f}))^2 \leq 2 \times \frac{\mathbb{E}\|f - \hat{f}\|^2}{\mathrm{Var}(f)}. \tag{3.20}$$

We sum up this result in the Table 3.1 for the sake of clarity.

Therefore, to quantify the asymptotic rate of these risk bounds, we must analyze the asymptotics of the risk of approximation $\mathbb{E}\|f - \hat{f}\|^2$. In the next section, we derive some rates of convergence for specific learning procedures.

3.4 Rates of convergence of theoretical indices for metamodels from model selection

The results shown in previous section allow to upper-bound the error made by using a metamodel instead of the true model when computing GSA indices. The key element of this bound is the quadratic norm between both models. Therefore, the aim is to quantify how good the metamodel is at approximating f. The goal is classical in the regression literature and different rates of convergence can be found depending on the architecture chosen for metamodeling. We recall here some results and directly apply them to our framework in order to obtain rates of convergence for the quantity $GSA(\hat{f})$ toward the quantity $GSA(f)$. This convergence is with respect to the size of the sample used for the construction of the metamodel. We focus the results on two main families of metamodel: wavelet-based metamodels and Gaussian Proccesses.

3.4.1 Wavelet-based metamodeling:

For wavelet-based metamodeling, we need to assume the true model to be in a Besov space. Besov space have been widely used in recent years as for particular choice of parameters, these sets contain Sobolev and Hölder balls, among others spaces commonly found for instance in the partial differential equation literature [DP88; Che12]. This has led to recent studies of universal estimators and their possible rates of convergence toward a given model. For the analysis in this

subsection, we suppose without loss of generality the inputs to follow the uniform distribution in the unit cube of $\mathbb{R}^d$, $\mathbf{X} \sim \mathcal{U}([0,1]^d)$. We consider here the case where the true model f is in the Besov space $\mathcal{B}^s_{p,q}$, for $p/d < s < \infty$ and $p, q \in [1, \infty)$.

In [Che+20], the authors describe how to approximate any function of the Besov space $\mathcal{B}^s_{p,q}$ by using its decomposition on wavelet bases, with explicit formulas. We refer the interested reader to this work, along with [Här+12; Dau95] for complementary information on this decomposition. This allows us to leverage their results on approximation risk of such metamodels in order to state the following result for GSA indices. While the original result is wider than our application, it requires only one more additional assumption to be used: the noise must have a finite moment of order 4. Suppose that we have $\hat{f}_n$ a wavelet-based metamodel approximating f. We then have the following theorem.

Theorem 3.4.1. *Suppose that $f \in \mathcal{B}^s_{p,q}$ and that the noise ε has a finite moment of order 4. Suppose that we have $\hat{f}_n$ is learned through the procedure explained in [Che+20]. Then, for $GSA \in \{Sob, Sob^{ind}, SobT, SobT^{ind}, Sh\}$*

$$\mathbb{E}[(GSA(f) - GSA(\hat{f}_n))^2] \leq C(s,p)\log(n)n^{-2s/(d+2s)}. \tag{3.21}$$

If $GSA \in \{CvM, CvM^{ind}, CvMT, CvMT^{ind}\}$, then,

$$\mathbb{E}[(GSA(f) - GSA(\hat{f}_n))^2] \leq C(s,p)\log(n)n^{-s/(d+2s)}. \tag{3.22}$$

However, a drawback of this rate is the presence of the term $\log(n)$, as optimal rates of convergence in these spaces seems to be of order $n^{-2s/(d+2s)}$, see [Här+12]. In [Che+20], such a rate is obtained for a linear wavelet estimator. Hence, if we know that the model is linear – and therefore, in our case, that Total Sobol' indices and Sobol' indices are equal – it is possible to leverage this information in order to obtain faster rates. Note that the linear case is also studied in [Bar02] in order to obtain similar rates in $\mathcal{B}^s_{2,\infty}$. Nonetheless, this assumption is usually too unrealistic to be safely made in practice.

3.4.2 Using Gaussian Processes for metamodeling

The principle of GP metamodeling is to assume that the model f is a realisation of a Gaussian Process (GP), usually with null mean function. Then, by using data from evaluations of the model, we can create another Gaussian Process from which realizations will be approximations of the true model. This is widely used since Gaussian Process is adaptable and numerous theorems can be used to gain information on the structure of the process (e.g. Mercer theorem or Karhunen-Loève theorem). We first recall the definition of a Gaussian Process and then state the assumptions needed and the rates of convergence we have here.

Definition 3.4.1. *Let $Z = \{Z(x), x \in \mathbb{R}^d\}$ be a collection of random variables. Then Z is a Gaussian Process if for any $n \in \mathbb{N}, (x_1, \cdots, x_n) \in \mathbb{R}^{d\times n}$, the random vector $(Z(x_1), \cdots, Z(x_n))^T$ is multivariate normal.*

Alternatively, a Gaussian Process is a stochastic process that can be entirely characterized by its mean function $\mu(x) = \mathbb{E}[Z(x)]$ and its covariance function $k(x, x') = \mathrm{Cov}(Z(x), Z(x'))$ for all $x, x' \in \mathbb{R}^d$. We denote $Z \sim GP(\mu, k)$. For simplicity, we assume the underlying space is $\mathbb{R}^d$ but extensions for other spaces can be found in the literature. Additionally, we assume that the covariance function is radial, that is of the form $k(x, x') = \tilde{k}(x - x')$.

For GP metamodeling, we usually suppose that f is a realization of a centered GP $Z \sim GP(0, k)$. We then use the kriging mean and kriging covariance formulas

$$\mu_n(x) = \mu(x) - k(x, \mathcal{D})[k(\mathcal{D}, \mathcal{D}) + \sigma^2 I_n]^{-1} Y \tag{3.23}$$

$$k_n(x, y) = k(x, y) - k(x, \mathcal{D})[k(\mathcal{D}, \mathcal{D}) + \sigma^2 I_n]^{-1} k(\mathcal{D}, y) \tag{3.24}$$

$$\tag{3.25}$$

to create $Z_n \sim GP(\mu_n, k_n)$ an approximation of Z. As n increases, realisations of Z_n tends to the true model f.

In our context, we need only one assumption on the decay of the correlation function. We suppose that there exists positive constants c_1 and c_2, and $m > d/2$ such that

$$c_1(1 + \|\omega\|_2^2)^{-m} \leq \mathcal{F}(\tilde{k})(\omega) \leq c_2(1 + \|\omega\|_2^2)^{-m}. \tag{3.26}$$

Usual covariance functions such as Matérn kernels satisfy such an assumption. We can then state the theorem from [WJ21] that will be of interest later.

Theorem 3.4.2. *Suppose the assumption stated above to hold, and suppose $m < \infty$. Then for all $t \geq C_0$ and n, with probability at least $1 - \exp -t$, we have*

$$\|Z_n - \hat{f}\|^2 = C(1 + t)n^{-\frac{2m-d}{2m}}, \tag{3.27}$$

where C_0 and C are constants not depending on n and t.

Remark 10. *For now, we supposed that the covariance function k was known and that Z_n is a GP with same covariance function as Z. However, this might not always be true. Nonetheless, in the misspecified case (i.e. when another covariance function is used for Z_n), similar results hold – albeit with change in the rate of convergence. We refer to [WJ21] for more details.*

If we use this result in our framework, then we have the following corollary.

Corollary 3. *Let f_n be a realisation of Z_n. Then for all $t \geq C_0$ and n, with probability at least $1 - \exp -t$, we have,*

$$\mathbb{E}\left[(GSA(f) - GSA(f_n))^2\right] \leq C_0(1 + t)n^{-(m+d/2)/2m}. \tag{3.28}$$

3.5 Auditing Algorithmic Fairness

Thanks to the results from the previous sections, we can provide a road-map for auditing algorithms when access to the algorithm f is not possible. For this, we

compute fairness metrics on a metamodel as an alternative to direct computation on the audited algorithm, and various bounds enable us to assess whether the true but inaccessible model is fair with respect to the chosen metric, or not.

While auditing fairness, the goal is to assert whether the quantity $\mathcal{F}(f)$ is above or below a given threshold ε, with $\mathcal{F}$ being a fairness metric provided by a legislator and f an audited model provided by an auditee. Classical (un)fairness metrics include for instance *Disparate Impact*, *Equality of Odds* or *Shapley values* and we refer the reader to the previous chapters for more insights on these metrics. It has been proven [Bén+21; GBM22] that these group fairness metric are equivalent to indices coming from the Global Sensitivity Analysis literature and defined in Section 3.2. However, the true model might not be accessible, with only access to an approximating metamodel $\hat{f}_n$ previously obtained using a sample of size n. This means we cannot have additional runs of $f(.)$ and therefore cannot have convergent estimations of the fairness metric. But we can easily use $\hat{f}_n$, and therefore can use a large sample of size N on the metamodel to estimate the fairness metric. Typically, n is small and used only to create the metamodel, while N can be much larger. In such a case, there is a need for bounds allowing to assess whether f is fair using only information from $\hat{f}_n$.

In order to do this, we use the following inequality:

$$\mathcal{F}(f) \leq \widehat{\mathcal{F}^N}\left(\hat{f}_n\right) + \underbrace{\left|\widehat{\mathcal{F}^N}\left(\hat{f}_n\right) - \mathcal{F}\left(\hat{f}_n\right)\right|}_{estimation\ risk} + \underbrace{\left|\mathcal{F}\left(\hat{f}_n\right) - \mathcal{F}(f)\right|}_{approximation\ risk}, \tag{3.29}$$

where the quantity $\widehat{\mathcal{F}^N}(.)$ is the estimator of the fairness metric $\mathcal{F}$. If the right-hand side of this inequality is less than the threshold ε, then so is the left-hand side and the audited algorithm is fair. However, this is a harder condition to fulfill. We will see later in this section that the additional cost for this type of audit can be reduced if we make further assumptions. However, in general, this gap is the price to pay in order to certify fairness on an inaccessible model.

We stress that, in this inequality, there are two different sample size, N and n. The quantity N is the size of the sample used for estimation of the fairness metric, and can be typically as large as desired since it involves only calls to the metamodel, which is supposed readily available and cheap. The quantity n is the size of the sample used for creating the metamodel, and is representative of how good $\hat{f}_n$ can approximate f. This number may not be as large as N and can be the limiting factor since it is the number of calls to the true but expensive model needed.

Remark 11. *This decomposition is the best we can do under the assumption of metamodel usage. As such, this inequality is tight. To reduce this additional cost, it is necessary to either suppose the metamodel to be in fact equal to the model (to remove the approximation risk) or that the metamodel allow for perfect computation of the fairness metric (to remove the estimation risk).*

3.5.1 (Un)fairness computed for the metamodel

The first quantity of interest in this decomposition is the estimate of the (un)fairness metric, computed for the metamodel. This is usually a straightforward and widely studied subject since it is the direct estimation of the metric when using a dataset. Nonetheless, different scheme of estimation may exist for a given metric. For instance, fairness metric can usually be estimated using Monte-Carlo techniques, or Importance Sampling if necessary. However, for various metrics linked with Sensitivity Analysis for instance, specific schemes can be used such as "Pick and Freeze", Chatterjee-based estimates or even spectral approaches such as FAST. These methods allow for faster, cheaper and/or easier estimations by leveraging additional information on the used metric. More information on these estimation procedures are included in Chapter 1

3.5.2 Estimation risk

The estimation risk $\left|\mathcal{F}\left(\hat{f}_n\right) - \hat{\mathcal{F}}\left(\hat{f}_n\right)\right|$ is the error made by estimation a fairness metric using data and not having direct access to the theoretical value. There are several tools that allow practitioners to upper-bound this error:

- **Well-chosen metamodel:** If $\hat{f}_n$ is of a certain form, it may be possible to directly compute the quantity $\mathcal{F}\left(\hat{f}_n\right)$. In such a case, the estimation risk $\left|\mathcal{F}\left(\hat{f}_n\right) - \hat{\mathcal{F}}\left(\hat{f}_n\right)\right|$ is null. Note however that this is usually not the case as this property is a combination of using the right metamodel and the right metric. An example of such a situation is obtained when using linear models conjointly with metrics based on a variance decomposition (i.e. any fairness metric linked to Sobol' indices). This is also the case when using Causal theory, under the condition of knowing all the causal equations contained in the DAG.

- **Asymptotics:** The asymptotics of the estimation risk can be studied as well. Usually, $\hat{\mathcal{F}}\left(\hat{f}_n\right)$ is chosen so as to be a convergent estimator of $\mathcal{F}\left(\hat{f}_n\right)$. In this case, tools such as asymptotic normality or rates of convergence allow to infer the behaviour of the estimation risk as the number of calls to the metamodel increases. Note that using asymptotics may appear as sub-optimal compared to fixed-size bounds. However, this estimation risk is computed using only calls to the metamodel. This algorithm is supposed to be both accessible and cheaper to evaluate. As such, it is not unrealistic to suppose additional calls possible in order to obtain tighter bounds. An example for these results are for instance Central Limit Theorems that can be found in the literature for various estimators of Sobol' indices.

- **Fixed-sample bounds:** If the estimation risk needs to be studied but a large sample size is not obtainable, then it is necessary to study the

behaviour of $\left|\mathcal{F}\left(\hat{f}_n\right) - \hat{\mathcal{F}}\left(\hat{f}_n\right)\right|$ at fixed sample size. Tools for this are typically concentration inequalities. It is not possible to describe exactly these bounds without information as they highly depend on the form of the estimator. Nonetheless, typical tools found for instance when considering Sobol' indices and therefore fairness metrics derived from them are Bennett inequalities, decay of Fourier coefficient when using FAST estimators. Test of independence can also be of use when considering HSIC-based metrics.

As an example, we show here the various bounds one may obtain for the estimation risk when considering a specific fairness metric: *Disparate Impact*. For a binary classifier φ taking as inputs $\mathbf{X}$ and a binary sensitive variable S, the Disparate Impact is defined as follow:

$$DI(\varphi, \mathbf{X}, S) = \frac{\mathbb{P}\left(\varphi(\mathbf{X}, S) = 1 | S = 0\right)}{\mathbb{P}\left(\varphi(\mathbf{X}, S) = 1 | S = 1\right)}. \tag{3.30}$$

The natural estimator of the Disparate Impact is to plug-in Monte-Carlo estimates for the probabilities in Equation 3.30. This leads to the formula

$$\widehat{DI}(\varphi, \mathbf{X}^{(n)}, S^{(n)}) = \frac{n^{-1}\mathbb{1}_{\varphi(\mathbf{X}^{(n)}, S^{(n)})=1|S=0}}{n^{-1}\mathbb{1}_{\varphi(\mathbf{X}^{(n)}, S^{(n)})=1|S=1}}. \tag{3.31}$$

As for asymptotics, we have the following Central Limit Theorem.

Theorem 3.5.1. *Let DI and $\widehat{DI}$ be the Disparate Impact and its estimator. Then*

$$\sqrt{n}\left(\widehat{DI}(\varphi, \mathbf{X}^{(n)}, S^{(n)}) - DI(\varphi, \mathbf{X}, S)\right) \to \mathcal{N}(0, \sigma^2), \tag{3.32}$$

where σ^2 is equal to $\frac{\mathbb{P}(\varphi(\mathbf{X},S)=1|S=0)}{\mathbb{P}(\varphi(\mathbf{X},S)=1|S=1)^2} \times (1 + \frac{\mathbb{P}(\varphi(\mathbf{X},S)=1|S=0)}{\mathbb{P}(\varphi(\mathbf{X},S)=1|S=1)})$.

Lastly, if we are interested in a concentration inequality, one can use the following theorem.

Theorem 3.5.2. *Let $A_i, B_i, i \in [\![1, n]\!]$ be independent random variables almost surely bounded by C, and let a, b be the mean of respectively A_1 and B_1. Then, we have the following concentration inequality:*

$$\mathbb{P}\left(\left|\frac{n^{-1}\sum_{i=1}^n A_i}{n^{-1}\sum_{i=1}^n B_i} - \frac{a}{b}\right| > t\right) < 2\exp\left(\frac{-nt^2b^2}{2C^2}\right) \tag{3.33}$$

Its corollary is a direct application to the Disparate Impact estimator.

Corollary 4. *Let $\widehat{DI}(\varphi, \mathbf{X}^{(n)}, S^n)$ be the estimator of the Disparate Impact DI. Then the following concentration inequality holds*

$$\mathbb{P}\left(\left|\widehat{DI}(\varphi, \mathbf{X}^{(n)}, S^{(n)}) - DI(\varphi, \mathbf{X}, S)\right| > t\right) < 2\exp\left(\frac{-nt^2}{2}\right). \tag{3.34}$$

3.5.3 Approximation risk

The approximation risk $\left|\mathcal{F}(f) - \mathcal{F}\left(\hat{f}_n\right)\right|$ is the error made by considering and auditing a metamodel instead of the inaccessible true model. This is the additional cost paid by the auditee for choosing not to provide the auditor with the true model. As for the previous item – the estimation risk – several tools are available to derive upper-bounds for this quantity:

- **No metamodeling:** If no metamodel is used – that is to say, if the model is in fact accessible and $f = \hat{f}_n$– then the approximation risk disappear. While the goal of this work is to provide with a general framework for auditing through the use of metamodels, it is important to note that this same framework holds when audit is made easier and auditors have access to the true algorithm.

If this is not the case – if a metamodel is truly needed – then the main goal is to be able to compare the approximation risk $\|\mathcal{F}(f) - \mathcal{F}\left(\hat{f}_n\right)\|_2$ and the distance between the unaccessible model and its approximating metamodel $\|f - \hat{f}_n\|_2$. Such comparisons depend on the form of the fairness metric. Immediate results may be found by using Lipschitz metrics or by computation of the continuity modulus of the metric. Other results comparing this risk and the distance between model and metamodel can be found for Sobol'-based indices in [JNP14], [Pan21] or results from the previous section.

- **Asymptotics:** As soon as the approximation risk is bounded by the distance between a model and its metamodel, it remains to study the behaviour of $\|f - \hat{f}_n\|_2$. This is usually done in the regression literature. As such, rates of convergences can be found in previous section for wavelet-based metamodeling, Gaussian Process metamodeling, etc. However, note that a limitation of such asymptotics is that a large number of calls to the true but inaccessible model are needed, in opposition to asymptotics for the estimation risk. For this reason, asymptotics can be costly or outright impossible to obtain. Depending on the possible budget and accessibility of the model, it might be easier to choose a metamodel so that $\|f - \hat{f}_n\|_2$ is easier to compute, even if it deteriorates the estimation of the fairness metric, since additional calls to the metamodel are more readily available.

- **Fixed-sample bounds:** Because of the point raised above, it might be necessary to consider bounds for the quantity $\|f - \hat{f}_n\|_2$, at fixed sample size. Note that, in this case, the size considered is the size of the sample needed to obtain a metamodel, and not linked with the estimation of the fairness metric. As before, these bounds are usually obtained through use of concentration inequalities.

Remark 12. *It is not unusual for analysis of the metamodel to be more complicated than other parts of the audit. This is mainly due to the lack of access to the true model. As such, if the metamodel is provided to the auditors directly by the*

auditee, there is a shift of trust from considerations related to the fairness measure ("How fair is the algorithm?") to considerations on the approximation level of the metamodel ("Is this metamodel really as close or as well-approximating as the auditee is claiming?"). This can be an issue to keep in mind, especially if the metamodel is chosen for "fairwashing" [Aïv+19], that is to say voluntarily chosen so as to minimize $\left|\mathcal{F}\left(\hat{f}_n\right) - \hat{\mathcal{F}}\left(\hat{f}_n\right)\right| + \hat{\mathcal{F}}\left(\hat{f}_n\right)$.

Chapter 4

Further developments at the interface of GSA & Fairness

In this chapter, we study complementary results of interest at the interface between explainability and algorithmic fairness. These results answer specific questions or provide additional insights into these two frameworks and their current limitations.

Firstly, we study the Chatterjee-based estimation of Sobol' indices in Section 4.1. This recent scheme of estimation is important for several practical reasons. It does not require additional calls to the model – with a *given-data* assumption – and involves only rankings and permutations, leading to a fast numerical estimation. We provide a concentration inequality for this estimator, complementary to the central limit theorem already found in the literature. Additionally, we extend aforementioned theorems by proving a multivariate version that can be of particular interest when working with more than one input.

Secondly, we tackle an other subject which is known in the literature as *second-level GSA*. The premise of this topic is that input distributions may not be perfectly known in practice. This uncertainty can have repercussions on the value of GSA indices, as influence of a variable may come from its underlying distribution. While the impact of distributional perturbations is not a new topic, recent works adopt a "meta" point of view by applying GSA indices *to* GSA indices. This allows to quantify how impactful the distribution of an input variable is on its influence on the outcome of an algorithm. These ideas can find direct application in algorithmic fairness as they allow the practitioner to assess fairness distributional robustness.

4.1 Around Chatterjee estimation

This estimation scheme comes from recent works from Chatterjee & al [Cha20]. The most expensive part of estimation of Sobol' indices is the computation of

the numerator of the index, that is the variance of the conditional expectation. Indeed, this is where the dependency between the outcome and a specific variable, say for instance the first coordinate X^1, is. However, due to the conditioning, naive estimation schemes can quickly become expensive. In [Cha20; AC21], the authors propose a new way to quantify this conditional expectation and its variance, through the use of ranks and nearest neighbors. In essence, the idea behind this method is the following. If X^1 is independent of Y, two points whose values along the coordinate X_1 are close have no reason to have close outputs, since the information on the first coordinate is worthless for the outcome. Therefore, if these points are close to each other with respect to the first coordinate, the ranking of their outputs has no reason to exhibit particular structure. On the contrary, if X^1 is the main driver of the outcome of the model, input vectors with similar values for the feature X^1 should have similar outputs, and therefore the ranks of their outputs should be close one to another.

We mentioned earlier that this estimation is quite interesting, especially for industrial applications, because it does not need additional calls to the model. The conditional variance, which usually requires evaluations in specific points (see for instance the "Pick-and-Freeze" method) can be estimated directly with the data at hand, making use of sortings and rankings. This fact has led this estimator to be coined as a *given-data* estimator.

In this section, we provide several new results such as concentration inequalities, a multivariate central theorem and hints for Berry-Esseen type results.

For this chapter, we denote when needed with a superindex the coordinate along – $\mathbf{X} = (X^1, \cdots, X^d)$ and with a subindex the element in a sample. We can, without loss of generality, assume that we are interested in the Sobol' index of the first variable X^1, and that f is bounded. We assume for simplicity that there cannot be ties in the values taken by the first coordinate X^1 nor in the values of $f(\mathbf{X})$ – that is implicitly to assume that the density of the distributions $\mathbb{P}_{X^1}$ and $\mathbb{P}_{f(\mathbf{X})}$ are continuous. Then, we denote by $\sigma_n(i)$ the permutation that sorts the inputs $(X_i^1)_{i\in[\![1,n]\!]}$, so that

$$X^1_{\sigma_n(1)} < X^1_{\sigma_n(2)} < \cdots < X^1_{\sigma_n(n)},$$

with the convention $X^1_{\sigma_n(n+1)} = X^1_{\sigma_n(1)}$. Chatterjee-based estimator for Sobol' indices is then defined as following:

$$\xi_n(\mathbf{X}_1, \cdots, \mathbf{X}_n) = \frac{n^{-1}\sum_{i=1}^n f(\mathbf{X}_{\sigma_n(i)}) f(\mathbf{X}_{(\sigma_n(i+1))}) - \left(n^{-1}\sum_{i=1}^n f(\mathbf{X}_{(i)})\right)^2}{n^{-1}\sum_{i=1}^n f(\mathbf{X}_{(i)})^2 - \left(n^{-1}\sum_{i=1}^n f(\mathbf{X}_{(i)})\right)^2}. \tag{4.1}$$

Remark 13. *While the assumption of continuous random variables might seem too restrictive as it forbids the use of Chatterjee-based estimation when in presence of categorical inputs – as it is often the case in Fairness – this assumption is done for simplicity. Indeed, the estimation is possible – albeit a bit more technical – when ties are possible, by simply choosing at random the nearest neighbor among the tied points. However, for the results shown later, we restrict ourselves to the case of continuous random variables.*

Remark 14. *Note that the only difference between Chatterjee-based estimation and Pick-and-Freeze estimation is the first term of the numerator, that is to say the term* $n^{-1}\sum f(\mathbf{X}_{\sigma_n(i)})f(\mathbf{X}_{\sigma_n(i+1)})$*. This term – and its equivalent in other methods – is the core of the Sobol' estimation since it is quantifying the impact of* X^1 *on the outcome. The other terms are regular plug-in estimators of the quantities* $\mathbb{E}[Y^2]$ *and* $\mathbb{E}[Y]^2$*.*

Remark 15. *A similar formula involving the ranks of the sequence* $(f(\mathbf{X}_{\sigma_n(i)}))$ *instead of the numerical values can be found in the literature. This alternate formula is an estimator of Cramér-von-Mises indices. We focus here on the Sobol' estimator but similar results can be derived from this analysis.*

This quantity converges as n increases towards the Sobol' index of the input variable X^1, by construction of the permutationσ, see [Cha20]. For quantification of Sobol' indices of higher order, we refer to [Cha20; Gam+20] where a formula is given, using a permutation chosen so as to solve a traveling salesman problem.

If the outcome of the model is multidimensional – that is if $\mathbf{f} = (f_1(\mathbf{X}), \cdots, f_l(\mathbf{X}))^T$ and for all $k \in [\![1, l]\!]$, $f_k(\mathbf{X})$ is square integrable – we can define an extension of Sobol' indices – see [Gam+14] – as

$$\bar{S}_u(\mathbf{f}) = \bar{S}_u((f_1(\mathbf{X}), \cdots, f_l(\mathbf{X}))^T) = \frac{\sum_{k=1}^{l} \operatorname{Var} \mathbb{E}[f_k(\mathbf{X})|X_u]}{\sum_{k=1}^{l} \operatorname{Var} f_k(\mathbf{X})}. \tag{4.2}$$

Note that this index is the only extension to benefit from several "good" properties such as lack of dependency on the model or invariance to left-composition of $\mathbf{f}$ by any isometry.

Because of the expression of Equation 4.2, the equivalent Chatterjee-based estimator of this quantity is

$$\bar{\xi_n}(\mathbf{f}, \mathbf{X}_1, \cdots, \mathbf{X}_n) = \frac{\sum_{k=1}^{l} \left(n^{-1} \sum_{i=1}^{n} f_k(\mathbf{X}_{\sigma_n(i)}) f_k(\mathbf{X}_{\sigma_n(i+1)}) - \left(n^{-1} \sum_{i=1}^{n} f_k(\mathbf{X}_{(i)}) \right)^2 \right)}{\sum_{k=1}^{l} \left(n^{-1} \sum_{i=1}^{n} f_k(\mathbf{X}_{(i)})^2 - \left(n^{-1} \sum_{i=1}^{n} f_k(\mathbf{X}_{(i)}) \right)^2 \right)}. \tag{4.3}$$

Note that the permutation is the same for all the functions $f_k, k \in [\![1, k]\!]$.

In the literature, one can find – see [Gam+20] – a Central Limit Theorem for the estimator ξ_n – i.e. for unidimensional output – under some technical assumptions. In the following, we complete this result with concentration inequalities for both ξ_n and $\bar{\xi_n}$ – tackling the framework of multidimensional outputs. We also extend the existing result by proving a Central Limit Theorem for $\bar{\xi_n}$ – multidimensional output – and by weakening the needed assumption on X. In order to complete the statistical analysis of this estimator, we then provide hints for a Berry-Esseen type result.

4.1.1 Concentration inequality

We provide here a concentration inequality for the estimator ξ_n. While new, this inequality is similar to what can be derived for Pick-and-Freeze estimators, see

[Gam+16]. In order to obtain this result, we need to assume that the outcome of the model is bounded. We recall here the Mc-Diarmid's inequality that will be used later in the proof.

Theorem 4.1.1 (Mc-Diarmid's inequality)**.** *We denote by $E_{\mathbf{X}}$ the set in which $\mathbf{X}$ takes value. Let $\varphi : E_{\mathbf{X}}^n \to \mathbb{R}$ be a function such that there are constants $c_1, \cdots, c_n$ for which for all $i \in [\![1, n]\!]$ and all $(\mathbf{x}_1, \cdots, \mathbf{x}_n) \in E_{\mathbf{X}}^n$,*

$$\sup_{\mathbf{x}_i' \in E_{\mathbf{X}}} |\varphi(\mathbf{x}_1, \cdots, \mathbf{x}_i', \cdots \mathbf{x}_n) - \varphi(\mathbf{x}_1, \cdots, \mathbf{x}_n)| \leq c_i. \tag{4.4}$$

Consider independent random variables $\mathbf{X}_1, \cdots, \mathbf{X}_n$ where $\mathbf{X}_i \in E_{\mathbf{X}}$ for all i. Then for all $t > 0$,

$$\mathbb{P}\left(|\varphi(\mathbf{X}_1, \cdots, \mathbf{X}_n) - \mathbb{E}[\varphi(\mathbf{X}_1, \cdots, \mathbf{X}_n)]| \geq t\right) \leq 2\exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right). \tag{4.5}$$

We can now state the theorem.

Theorem 4.1.2. *Let $b > 0$ and $t > 0$. Assume that the random variable $f(\mathbf{X})$ belongs to $[-b/2, b/2]$. Then, if S is the Sobol' index of f with respect to X_1, we have*

$$\mathbb{P}\left(|\xi_n(\mathbf{X}_1, \cdots, \mathbf{X}_n) - S(f)| \geq t\right) \leq 2\exp\left(-\frac{2t^2 n}{b^4(3(S+t) - 1 + (S+t-1)/n)^2}\right). \tag{4.6}$$

Proof. This proof combines two part. The first part aims at rewriting the ratio of sums as only one sum, so that classical results can be used. This idea has been used for obtaining concentration inequalities for the Pick-and-Freeze estimator, see [Gam+14]. The second part is to circumvent the issue of the random permutation, by using the Mc-Diarmid inequality. Note that this was used by [Cha20], albeit for ranks. We recall that since Sobol' indices are invariant by homothety, we can assume that $f(\mathbf{X})$ is centered.

We need to compute a quantity of the form

$$\mathbb{P}\left(\frac{a_n - b_n^2}{c_n - d_n^2} - \frac{a}{c} > t\right) \tag{4.7}$$

where a_n, b_n, c_n and d_n are statistical series of mean respectively $a, b = 0, c$ and $d = 0$. However, because of the ratio, we cannot use classical results from the literature that are stated for sums of random variables. For this reason, we need to change this expression.

One can show that this is in fact equivalent to the quantification of the probability

$$\mathbb{P}\left(a_n - b_n^2 - \left(\frac{a}{c} + t\right)(c_n - d_n^2) - \mathbb{E}\left[a_n - b_n^2 - \left(\frac{a}{c} + t\right)(c_n - d_n^2)\right] > tc\right) \tag{4.8}$$

since we have the equality $\mathbb{E}[a_n - b_n^2 - (\frac{a}{c} + t)(c_n - d_n^2)] = tc$.

In our case, we have

$$a_n = \frac{1}{N} \sum f(\mathbf{X}_{\sigma_n(i)}) f(\mathbf{X}_{(\sigma_n(i+1))}) \tag{4.9}$$

$$b_n^2 = d_n^2 = \left(\frac{1}{N} \sum f(\mathbf{X}_{(i)}) \right)^2 \tag{4.10}$$

$$c_n = \frac{1}{N} \sum f(\mathbf{X}_{(i)})^2 \tag{4.11}$$

$$S = \frac{a}{c} \tag{4.12}$$

$$\tag{4.13}$$

We can then define a function that maps the inputs $(\mathbf{X}_1, \cdots, \mathbf{X}_n)$ to the quantity of interest in the probability. Let G be the real-valued function so that

$$G(\mathbf{X}_1, \cdots, \mathbf{X}_n) = \frac{1}{n} \sum_{i=1}^{n} f(\mathbf{X}_{\sigma_n(i)}) f(\mathbf{X}_{\sigma_n(i+1)}) + (S + t - 1) \left(\frac{1}{n} \sum_{i=1}^{n} f(\mathbf{X}_{\sigma_n(i)}) \right)^2 - (S + t) \left(\frac{1}{n} \sum_{i=1}^{n} f(\mathbf{X}_{\sigma_n(i)})^2 \right). \tag{4.14}$$

In order to use the Mc-Diarmid's inequality, we need to control the following quantity:

$$\sup_{\mathbf{x}_i' \in E_{\mathbf{X}}} \left| G(x_1, \cdots, x_i', \cdots x_n) - G(x_1, \cdots, x_n) \right|. \tag{4.15}$$

One can insert the expression of G to obtain

$$\begin{aligned} \sup_{\mathbf{x}_i' \in E_{\mathbf{X}}} |G(\mathbf{x}_1, \cdots, \mathbf{x}_i', \cdots \mathbf{x}_n) - & G(\mathbf{x}_1, \cdots, \mathbf{x}_n)| \leq \\ & \left| \frac{1}{n} \left(f(\mathbf{x}_i') - f(\mathbf{x}_i) \right) \left(f(\mathbf{x}_{\sigma_n(i)}) + f(\mathbf{x}_{\sigma_n^{-1}(i)}) \right) \right. \\ & + \frac{(S + t - 1)}{n^2} \left(f(\mathbf{x}_i') - f(\mathbf{x}_i) \right) \left(f(\mathbf{x}_i) + f(\mathbf{x}_i') + \sum_{k \neq i} f(\mathbf{x}_k) \right) \\ & \left. - \frac{(S + t)}{n} \left(f(\mathbf{x}_i') - f(\mathbf{x}_i) \right) \left(f(\mathbf{x}_i) + f(\mathbf{x}_i') \right) \right|. \end{aligned}$$

Using the fact that f is bounded by a constant b, we therefore obtain

$$\sup_{\mathbf{x}_i' \in E_{\mathbf{X}}} |G(\mathbf{x}_1, \cdots, \mathbf{x}_i', \cdots \mathbf{x}_n) - G(\mathbf{x}_1, \cdots, \mathbf{x}_n)| \leq \frac{b^2}{n} \left(1 + 3(S + t) + \frac{1}{N}(S + t - 1) \right).$$

Note that this bound is the same for all coordinates.

We then use the Mc-Diarmid's inequality stated above with the constants $c_i = \frac{b^2}{n}\left(1 + 3(S+t) + \frac{1}{N}(S+t-1)\right)$. This allows us to conclude that

$$\mathbb{P}\left(|\xi_n(\mathbf{X}_1, \cdots, \mathbf{X}_n) - S(f)| \geq t\right) \leq 2\exp\left(-\frac{2t^2 n}{b^4(3(S+t) - 1 + (S+t-1)/n)^2}\right),$$

which concludes. □

We can now state the concentration inequality for the estimator $\bar{\xi_n}$, which is a direct corollary from Theorem 4.1.2.

Corollary 5. *Let $b_k > 0$ for every $k \in [\![1, l]\!]$ and $t > 0$. Assume that the random variable $f_k(\mathbf{X})$ belongs to $[-b_k/2, b_k/2]$, and let $B^2 = \sum_{k=1}^{l} b_k^2$ Then, if $\bar{S}$ is the Sobol' index of $\mathbf{f}$ with respect to X_1, we have*

$$\mathbb{P}\left(\bar{\xi_n} - \bar{S} > t\right) \leq\leq 2\exp\left(-\frac{2t^2 n}{B^4(3(\bar{S}+t) - 1 + (\bar{S}+t-1)/n)^2}\right). \tag{4.16}$$

Proof. The proof is a direct adaptation of the previous one. We define the function

$$F(\mathbf{X}_1, \cdots, \mathbf{X}_n) = \sum_{k=1}^{l} G_k(\mathbf{X}_1 \cdots, \mathbf{X}_n), \tag{4.17}$$

where G_k is the real-valued function

$$G_k(\mathbf{X}_1, \cdots, \mathbf{X}_n) = \frac{1}{n}\sum_{i=1}^{n} f_k(\mathbf{X}_{\sigma_n(i)}) f_k(\mathbf{X}_{\sigma_n(i+1)}) + (S+t-1)\left(\frac{1}{n}\sum_{i=1}^{n} f_k(\mathbf{X}_{\sigma_n(i)})\right)^2 - (S+t)\left(\frac{1}{n}\sum_{i=1}^{n} f_k(\mathbf{X}_{\sigma_n(i)})^2\right). \tag{4.18}$$

and use the results shown in the univariate proof to derive the result using straight-forward manipulations.

□

Remark 16 (A negative result)**.** *In the literature, one can find a similar concentration inequality for the Pick-and-Freeze estimator, see [Gam+14]. However, instead of using the Mc-Diarmid's inequality in order to conclude the proof, the authors leverage a result stated for Lipschitz functionals found in the works of Ledoux ([Led01], Corollary 1.17), stated as follows.*

Corollary 6 (Ledoux)**.** *Let $P = \mu_1 \otimes \cdots \otimes \mu_n$ be a product probability measure on the cartesian product $X = X_1 \times \cdots \times X_n$ of metric spaces (X_i, d_i) with finite diameters $D_i, i = 1, \cdots, n$, endowed with the l^1 metric $d = \sum d_i$. Let F be a 1-Lipschitz function on (X, d). Then, for every $r \geq 0$,*

$$P\left(F - \mathbb{E}_P[F] \geq r\right) \leq \exp\left(-\frac{r^2}{2D^2}\right), \tag{4.19}$$

where $D^2 = \sum D_i^2$.

While this result seems promising at first, it does not apply in our case. Indeed, the Pick-and-Freeze estimator takes two samples of outputs $(f(\mathbf{X}_1), \cdots, f(\mathbf{X}_n))$ *and* $\left(f(X_1^1\,', X_1^{\sim 1}), \cdots, f(X_n^1\,', X_n^{\sim 1})\right)$ *and then apply a similar functional to the function* G *in our proof, which can be proven to be Lipschitz by considering the partial derivatives, to obtain their results.*

In our case, we need first to apply a random perturbation and therefore start directly with the inputs $(\mathbf{X}_1, \cdots, \mathbf{X}_n)$*. In the end, we consider the functional* $\varphi \circ f^{\otimes n} \circ \bar{\sigma}$ *where*

$$\begin{aligned}
\varphi(y_1, \cdots, y_n) &= \frac{1}{n}\sum y_i y_{i+1} + (S+t-1)\left(\frac{1}{n}y_i\right)^2 - (S+t)\frac{1}{n}\sum y_i^2 \\
f^{\otimes n}(\mathbf{x}_1, \cdots, \mathbf{x}_n) &= (f(x_1), \cdots, f(\mathbf{x}_n)) \\
\bar{\sigma}(\mathbf{x}_1, \cdots, \mathbf{x}_n) &= (\mathbf{x}_{\sigma(1)}, \cdots, \mathbf{x}_{\sigma_n}).
\end{aligned}$$

It is straightforward to verify that φ *is Lipschitz. However, even if we assume* f *(and therefore* $f^{\otimes n}$*) to be Lipschitz, it would remain to prove that* $\bar{\sigma}$ *is Lipschitz, which is false... To be precise,* $\bar{\sigma}$ *is Lipschitz only in the coordinate along which the sort is applied – the first for simplicity. Indeed, one can write*

$$X^1_{\sigma(i)} = \max_{u \subset [\![1,n]\!], |u| = n+1-i} \min_{j \in u}(X_j)$$

and as the max *and the* min *functions are* $1-$*Lipschitz, so is the sort operation. However, as soon as we add one coordinate, this is no longer the case: take for instance the two vectors* $((0,0),(1,1))$ *and* $((1,0),(0,1))$ *as a minimal counter-example.*

Remark 17 (A second negative result)**.** *Another idea for this inequality would be to take advantage of the literature on concentration for randomly permuted sums, see [Alb19; ACW16]. Indeed, since the* $\mathbf{X}_i$ *are i.i.d., the permutation* σ *is chosen randomly and uniformly along the set* $\mathcal{S}_n$ *of all permutations of the set* $[\![1,n]\!]$*. Using this fact, one can be tempted to use the theorem found in [ACW16]. This result is an equivalent of Theorem 6 stated in the remark above for a randomly permuted vector, with the only additional assumption that* F *is both Lipschitz and convex. However, in our case, while* φ *is Lipschitz, it is not convex and the result cannot be applied.*

4.1.2 Central Limit Theorem

For unidimensional output, the Central Limit Theorem for the estimator ξ_n is given in [Gam+20]. We can therefore state our extension for the estimator $\bar{\xi_n}$.

Theorem 4.1.3. *Assume that* X^1 *is a random variable defined on a interval of* $\mathbb{R}$ *whose distribution has a continuous and below-bounded density. Let* $\mathbf{f}$ *be a square integrable model with multidimensional output. Let* $\bar{\xi_n}$ *be the estimator of the Sobol' index* $\bar{S}_1(f)$*, both defined earlier. Then*

$$\sqrt{n}\left(\bar{\xi_n} - \bar{S}_1(\mathbf{f})\right) \to \mathcal{N}(0, s^2), \tag{4.20}$$

where s^2 *has an explicit formula.*

The proof is given in Appendix A.5.5.

Remark 18. *For technical reasons, X^1 is assumed in the original proof to be distributed with uniform law between 0 and 1. The reason is the need of asymptotic concentration of order statistics around precise, fixed points. This is easily done when X^1 is a standard uniform random variable, as its order statistics follow a Beta distribution. However, we can obtain a similar result for any random variable X^1 which have a below-bounded density $f_{X^1}(x) \geq M > 0$ for all x on which X^1 is defined. This is done by doing a Taylor expansion of the quantile function around well-chosen points. This allows us to weaken the needed assumptions of this theorem.*

The next step in statistical analysis of this Chatterjee estimator is to prove a Berry-Esseen theorem. A natural idea in this direction is to follow the proof of the Berry-Esseen theorem for the Pick-and-Freeze estimator, see [Gam+16]. This proof is based on Pinelis' theorem [PM09], which states:

Theorem 4.1.4 (Pinelis)**.** *Let $(V_i)_{i\geq 1}$ be a sequence of i.i.d. centered random variables in $\mathbb{R}^d$ for some $d \in \mathbb{N}^\star$. Let $\varphi : \mathbb{R}^d \to \mathbb{R}$ be some measurable function with $f(0) = 0$ such that*

$$\exists \varepsilon > 0, \exists M_\varepsilon > 0, s.t. |\varphi(x) - D\varphi(0)(x)| \leq \frac{M_\varepsilon}{2}\|x\|^2, \tag{4.21}$$

where $D\varphi(0)$ is the Fréchet derivative of φ at point 0. Take any $p \in (2, 3]$ and assume $\sigma = \left(\mathbb{E}[(D\varphi(0)(V))^2]\right)^{1/2} > 0$ and $\left(\mathbb{E}\|V\|^p\right)^{1/p} < \infty$. Then for all $z \in \mathbb{R}$,

$$\left| \mathbb{P}\left(\frac{\varphi(\bar{V_n})}{\sigma/\sqrt{n}} \leq z \right) - \Phi(z) \right| \leq \frac{\kappa}{n^{p/2-1}} \tag{4.22}$$

where Φ is the cumulative distribution function of a standard gaussian and κ is a generic constant that depends only upon p.

By applying this theorem to the sequence of random variables

$$(V_i) = (f(\mathbf{X}_{\sigma_n(i)})f(\mathbf{X}_{(\sigma_n(i+1))}) - \mathrm{Var}(\mathbb{E}[f(\mathbf{X})|X^1]), f(\mathbf{X}_{(i)}), (f(\mathbf{X}_{(i)}))^2 - \mathrm{Var}(f(\mathbf{X}))) \tag{4.23}$$

– we assume that $f(\mathbf{X})$ is centered for convenience – and with

$$\varphi(x, y, z) = \frac{x - y^2 - \mathrm{Var}(\mathbb{E}[f(\mathbf{X})|X^1])}{z - y^2 + \mathrm{Var}(f(\mathbf{X}))} - S, \tag{4.24}$$

we would obtain by direct application of the theorem the wanted result. However, this is not possible because of the lack of independency between the random variables V_i. Indeed, the permutation $\sigma(.)$ is introducing dependency and the assumptions are no longer verified.

In fact, the Chatterjee estimator is based on a ratio of U-statistics. By using a similar decomposition to the one used for concentration inequalities above, we would obtain a sum of U-statistics, which remains a U-statistic. For this type of

estimator, one can find in the literature various results providing Berry-Esseen bounds at different rates, see for instance [CJ78]. However, we will not provide the detail here as this result is left for future works, including its extension for estimators of Sobol' indices for multivariate output.

4.2 Second-Level GSA & Fairness robustness

4.2.1 Second-level GSA

Second-level Global Sensitivity Analysis is a recent preoccupation in Uncertainty Quantification. While the indices defined in the rest of this manuscript are useful tools to quantify the influence of inputs on the outcome of a model – and by extension provide some explainability – they are usually starting from the assumption that the input distribution is perfectly known. However, while this can be true in controlled environments such as experimentation plans or computer codes, this assumption does not always hold. In such a case, this uncertainty in the input distribution can tarnish the information provided by sensitivity indices and should be taken into account. In clear terms, we are interested in the following question: *"If the input distribution changes, what happens to the GSA indices?"*

Distributional perturbations

A first approach to answer this question is through local distributional modifications. In [HG19], the authors propose to study the operator which takes into input the probability distribution of the random vector $\mathbb{P}_X$ and returns a Sobol' index of interest. By computing the Fréchet derivative of this operator at the presumed original distribution, they manage to analyze the robustness of this operator and therefore of the Sobol' index to local distributional perturbations. By extending this point of view to any GSA measure, this provides a simple tool to assert local robustness to distributional changes. Assume we have a distribution $\mathbb{P}_X$ generating the inputs for which we compute a GSA metric GSA. We can define a distributional perturbation as a $\nu_{\mathbf{X}}$ such that $\mathbb{P}_X + \nu_{\mathbf{X}}$ still is a distribution. Technical details and constraints imposed on the perturbation are given in [HG19].

We can then reformulate our earlier question as the quantification, for small perturbations $\nu_{\mathbf{X}}$, of

$$GSA_{\mathbb{P}_X}(f) - GSA_{\mathbb{P}_X+\nu_{\mathbf{X}}}(f). \tag{4.25}$$

This quantity provides information on how much of a change in the influence of a feature one can expect if the distribution is changed from a nominal one to a perturbed one. In the case of Sobol' indices, this gap is explicitly computed as we have the following theorem:

Theorem 4.2.1 (Theorem 3.1 of [HG19])**.** *Let $\mathcal{S}_u$ be the operator that associates to a distribution $\mathbb{P}_X$ a Sobol' index $S_{X^u}(f)$ for an input variable X_u and model*

f. Then the operator $\mathcal{S}_u$ is Fréchet differentiable at $\mathbb{P}_X$ with Fréchet derivative the bounded linear operator $D\mathcal{S}_u$ whose expression is given explicitly in [HG19].

In addition of explicit computation of this quantity, an important element of this theorem that can be used in other contexts than Sobol' indices is the use of the Fréchet derivative of the GSA measure – here Sobol' indices – in order to upper-bound the gap defined in Equation 4.25. This proves to be a prospective element for local robustness, including in a Fairness framework. We explore this in next subsection 4.2.2.

However, while this result provides an almost complete understanding of the behavior of Sobol' indices in the neighborhood of a specific distribution, and hints for extensions to any GSA or Fairness measure, one may be interested in more global results.

Second-level GSA, HSIC2 & Importance Sampling

The second way to obtain a global form of distributional robustness takes its roots in [MML19]. In this work, the authors' interest is in the HSIC methodology. Indeed, the HSIC indices are a mean to quantify the impact of a random variable on the output of a model, through the lens of a kernel. As such, what happens if one chooses the random variable to be an uncertain distribution – using a kernel on distributions – and the output of interest to be the HSIC index itself?

In clear terms, the authors, if we denote by $HSIC(f, Z)$ the HSIC index of the random variable Z applied to the model f, propose to study the quantity

$$HSIC2(f, Z, \mathbb{P}_Z) = HSIC(HSIC(f, Z), \mathbb{P}_Z) \tag{4.26}$$

where $\mathbb{P}_Z$ is seen itself as a random variable – e.g. in a parametric framework, the generative parameters are random.

This quantity is an indicator of how much the distributional uncertainty impacts the computed importance of an input Z on the outcome of f. Intuitively, if the value is small, the impact of Z on f would stay the same under any change of distribution. On the contrary, if the value is high, then special care must be given to the generative distribution since even a small change can have huge consequences on the sensitivity analysis conducted so far.

While interesting and popular, usage of HSIC indices can sometimes be complicated due to the intrinsic choice of kernel, and its analysis can be more obscure than for others GSA indices, as mentioned in previous chapters. As such, for second-level GSA, we want in fact not to restrict ourselves to HSIC indices and therefore compute the quantity

$$GSA2_i(f) = GSA_{\theta_i}\left(GSA_{X_i}(f)\right), \tag{4.27}$$

where $GSA_Z(f)$ is a generic GSA index applied at algorithm f and random variable Z (e.g. Sobol' indices).

We can estimate this quantity by plugging-in estimators $\widehat{GSA}(.)$ to obtain the following expression:

$$\widehat{GSA2}_i(f) = \widehat{GSA}_{\theta_i}\left(\widehat{GSA}_{X_i}(f)\right). \tag{4.28}$$

Proposition 4. *If $\widehat{GSA}$ is consistent estimator of GSA and if GSA(.) is continuous, then the estimator $\widehat{GSA2}$ is a consistent estimator of GSA2.*

However, the main issue with this direct plug-in of estimators into the formula of GSA is the cost associated with such a scheme. Indeed, for each estimation of the inner-loop – that is to say, as soon as we have a target distribution parameterized by θ , in order to compute $\left(\widehat{GSA}_{X^i}(f)\right)$ – we need a sample of realizations of $\mathbf{X}$ of size at least n if we use a given-data scheme or $2n$ if we use a classical Pick-and-Freeze scheme. But in order to estimate the outer-loop – that is to say the $\widehat{GSA}_{\theta_i}(.)$ part of the estimator – we need to choose a new θ, redraw realizations of $\mathbf{X}$ along this new distribution and repeat the estimation of the inner-loop, at least N times. This "rinse and repeat" scheme is quite inefficient as this require in total up to $N \times n$ calls to the model, at the minimum.

Note that this issue was already raised in [MML19] for the use of second-level HSIC. In the work, the authors propose a *single-loop* scheme, leveraging additional information on the distributions in order to circumvent such a high cost. This technique, akin to Importance Sampling, works especially well for HSIC since it only involves expectations, which are quite straightforward to combine with change of distributions. On a similar note, recent works by [DBM22] provide similar results for Shapley values in the various estimation schemes. We complete their results in a straightforward fashion for Sobol' indices.

We want to use Importance Sampling in order to derive formulas that allow the estimation of Sobol' indices for a given distribution – denoted for simplicity by its parameter θ – using data sampled from an other distribution $\bar{theta}$. In other words, once we have one sample $(\mathbf{X}_1, \cdots, \mathbf{X}_n)$ with a well-chosen distribution, we want to be able to compute Sobol' indices for any other distribution. We refer to [MML19; DBM22] for more details on what makes a distribution "well-chosen".

For Sobol' indices, we are interested in a ratio of the form

$$\frac{\mathbb{E}_\theta\left[\mathbb{E}_\theta\left[f(\mathbf{X})|X_1\right]^2\right] - E_\theta\left[f(\mathbf{X})\right]^2}{E_\theta\left[f(\mathbf{X})^2\right] - E_\theta\left[f(\mathbf{X})\right]^2}. \tag{4.29}$$

We can rewrite this using expectations with respect to the distribution of reference with parameter $\bar{\theta}$ as

$$\frac{\mathbb{E}_{\bar{\theta}}\left[\mathbb{E}_{\bar{\theta}}\left[f(\mathbf{X})g^\theta_{\bar{\theta}(\mathbf{X}^{\sim 1})}|X_1\right]^2 g^\theta_{\bar{\theta}}(X^1)\right] - E_{\bar{\theta}}\left[f(\mathbf{X})g^\theta_{\bar{\theta}}(\mathbf{X})\right]^2}{E_{\bar{\theta}}\left[f(\mathbf{X})^2 g^\theta_{\bar{\theta}}(\mathbf{X})\right] - E_{\bar{\theta}}\left[f(\mathbf{X})g^\theta_{\bar{\theta}}(\mathbf{X})\right]^2}, \tag{4.30}$$

which involves ratio of densities respectively $g^\theta_{\bar{\theta}}(X^1)$, $g^\theta_{\bar{\theta}}(\mathbf{X}^{\sim 1})$ and $g^\theta_{\bar{\theta}}(\mathbf{X})$ – the ratio of the densities $d\mathbb{P}_{\theta,X^1}/d\mathbb{P}_{\bar{\theta},X^1}$, $d\mathbb{P}_{\theta,\mathbf{X}^{\sim 1}}/d\mathbb{P}_{\bar{\theta},\mathbf{X}^{\sim 1}}$ and $d\mathbb{P}_{\theta,\mathbf{X}}/d\mathbb{P}_{\bar{\theta},\mathbf{X}}$– and expectations under the distribution $\bar{\theta}$, that we suppose perfectly known and under which is drawn our sample, since this is our distribution of reference. Note that for this, we need the Radon–Nikodym derivative of $\mathbb{P}_{\theta,\mathbf{X}}$ against $\mathbb{P}_{\bar{\theta},\mathbf{X}}$ to be non-null and finite for every θ. Therefore, we can derive the Pick-and-Freeze

based Importance Sampling of Sobol' indices:

$$\hat{S} = \frac{\frac{1}{N}\sum_{i=1}^{N} f(X_i^1, \mathbf{X}_i^{\sim 1}) f(X_i^1, \mathbf{X}_i^{\prime,\sim 1}) g_{\bar{\theta}}^{\theta}(X_i^1) g_{\bar{\theta}}^{\theta}(\mathbf{X}^{\sim 1}) g_{\bar{\theta}}^{\theta}(\mathbf{X}^{\prime,\sim 1}) - \left(\frac{1}{N}\sum_{i=1}^{N} f(\mathbf{X}_i) g_{\bar{\theta}}^{\theta}(\mathbf{X}_i)\right)^2}{\frac{1}{N}\sum_{i=1}^{N} f(\mathbf{X}_i)^2 g_{\bar{\theta}}^{\theta}(\mathbf{X}_i) - \left(\frac{1}{N}\sum_{i=1}^{N} f(\mathbf{X}_i) g_{\bar{\theta}}^{\theta}(\mathbf{X}_i)\right)^2}. \tag{4.31}$$

Thanks to Slutsky lemma we know that this estimator converges towards the theoretical quantity which is the Sobol' index under distribution θ. This can be done in a similar fashion to obtain the Chatterjee-based Importance Sampling estimator, by using the replacing the first product by a product of sorted outputs as was done in previous section.

In order to obtain a GSA2 estimator, we can compose our estimator as follow. First, we sample $\mathbf{X}$ under the distribution $\bar{\theta}$ in order to obtain the realizations $(\mathbf{X}_1, \cdots, \mathbf{X}_n)$, which is done at the cost of n calls to the model. Then, dealing with the parameters θ, we draw a sample $(\theta_1, \cdots, \theta_N)$. For each one of these parameters, we compute $Sob_{\theta_i}(f)$ using Importance-Sampling, which is done without any additional calls to the model. Then, we use Chatterjee-based estimation in order to compute the outer GSA index, by sorting the sample $(\theta_1, \cdots, \theta_N)$ and ranking the associated vector $(Sob_{\theta_1}(f), \cdots, Sob_{\theta_N}(f))$. Therefore, instead of paying $N \times n$, we only require n calls to the function, albeit at the cost of higher knowledge of the distributions at stake since this requires knowledge of the Radon-Nikodym derivatives.

4.2.2 Fairness robustness

In previous subsection, we explored how second-level GSA can be used to leverage more information about the influence of input parameters – or their generative distributions – on an output. Translated into Algorithmic Fairness, these techniques allow deep scrutiny of sensitive variables in order to determine not only if an algorithm is unfair but also if its generative distribution has any impact on this unfairness characteristic. This can be assimilated to robustness. To the best of our knowledge, this approach is yet to be presented in the Fairness literature. The closest related work we found is [Fer+22]. In this work, the authors are interested in changes of fairness properties under *dataset modification*. The approach of this work is from the data to the wanted property – that is fairness – while ours is the other way around, from intrinsic properties of the model to (un)fairness detection and robustness. As such, these works are complementary.

If we recall the methods mentioned above, we described two types of distributional robustness in GSA2, that we can convert to Fairness robustness.

Local robustness

The first way to ensure a form of robustness is robustness under local distributional perturbations. Let $\mathcal{F}(f, \mathbb{P}_\theta)$ be a Fairness measure on the model f and with $\mathbb{P}_\theta$ the distribution of the inputs $(\mathbf{X}, S)$ – we recall that in a Fairness

framework, we have two types of inputs, non-sensitive inputs denoted by $\mathbf{X}$ and a sensitive feature denoted by S. We denote in the following by ν a small perturbation with respect to a distributional distance $d(.,.)$ – i.e. the Total Variation distance, a Wasserstein distance... – so that $\mathbb{P}_{\theta+\nu}$ still is a distribution and $d(\mathbb{P}_\theta, \mathbb{P}_{\theta+\nu})$ is smaller that a given ε. With such notations, we are interested in the quantity

$$|\mathcal{F}(f, \mathbb{P}_\theta) - \mathcal{F}(f, \mathbb{P}_{\theta+\nu})|. \tag{4.32}$$

Indeed, if this quantity is smaller in a neighborhood of $\mathbb{P}_\theta$ that a given threshold, we can guarantee that local changes of distribution will not impact our Fairness property by much. In clear words, if we certified a model as fair for a given distribution, the model can keep this certificate as long as the new distribution does not differ too greatly.

We have several options to bound this quantity. A first direct result can be given for the Disparate Impact, see the following theorem.

Proposition 5. *Let $DI(f, \mathbb{P}_\theta)$ be the variant of the Disparate-Impact measure defined as*

$$DI(f, \mathbb{P}_\theta) = \mathbb{P}_\theta\left(f(\mathbf{X}, S) = 1 | S = 0\right) - \mathbb{P}_\theta\left(f(\mathbf{X}, S) = 1 | S = 1\right). \tag{4.33}$$

Then we have

$$|\mathcal{F}(f, \mathbb{P}_\theta) - \mathcal{F}(f, \mathbb{P}_{\theta+\nu})| \leq d_{TV}(\mathbb{P}_\theta, \mathbb{P}_{\theta+\nu}), \tag{4.34}$$

where $d_{TV}(\mathbb{P}, \mathbb{Q})$ is the Total Variation distance between the distribution $\mathbb{P}$ and the distribution $\mathbb{Q}$.

Proof is a direct consequence of the definition of the Total Variation distance and the linearity of this fairness measure. However, in order to derive sharper bounds, an idea would be to use Fréchet derivatives. Indeed, if we denote by $D\mathcal{F}(f, \theta)(.)$ the Fréchet derivative of the fairness measure applied at point $(f, \mathbb{P}_\theta)$, then we obtain

$$\mathcal{F}(f, \mathbb{P}_\theta) - \mathcal{F}(f, \mathbb{P}_{\theta+\nu}) = D\mathcal{F}(f, \theta)(\mathbb{P}_\theta - \mathbb{P}_{\theta+\nu}) + \tau(\|\mathbb{P}_\theta - \mathbb{P}_{\theta+\nu}\|), \tag{4.35}$$

where we use the metric on the distribution space. As long as the Fréchet derivative is bounded at the point $(f, \mathbb{P}_\theta)$, then we can obtain the wanted result for a neighborhood of the distribution $\mathbb{P}_\theta$. Similar computations have been done previously for Sobol' indices – see [HG19].

Remark 19. *Note that having an explicit formula for the Fréchet derivative of the fairness measure is an important step towards fair learning. Indeed, one way to obtain fair learning is to add the fairness metric as a penalization in the loss function. By doing so, this ensures that the algorithm chosen by the learning procedure exhibits a Fairness property, see for instance [Ris+21]. However, in order to do so for instance with neural networks, it is necessary to know both the derivative of the constraint with respect to the weights in the network (forward pass) and the Fréchet derivative with respect to perturbations of the law (backward*

pass). While the results above provide tools to tackle this backward propagation, we provide for Sobol' indices a result concerning the forward propagation.

If we suppose that f is parametric, chosen according to a parameter w – for instance if f is a neural network, then we can derive the following result:

$$\frac{\partial Sob(f_w)}{\partial w} = \frac{2Cov(\mathbb{E}[\frac{\partial f}{\partial w}|X_i], \mathbb{E}[f_w(X)|X_i])Var(f_w(X)) - Var(\mathbb{E}[f_w(X)|X_i]) \times 2Cov(\frac{\partial f}{\partial w}, f_w(X))}{Var(f_w(X))^2}. \tag{4.36}$$

Proof. We start with the functional equality:

$$\frac{\partial \frac{F}{G}}{\partial w}(x,w) = \frac{\frac{\partial F}{\partial w}(x,w)G(x,w) - F(x,w)\frac{\partial G}{\partial w}(x,w)}{(G(x,w))^2}. \tag{4.37}$$

It remains then to compute $\frac{\partial}{\partial w}Var\mathbb{E}[f_w(X)|X_i]$ and $\frac{\partial}{\partial w}Var(f_w(X))$.

For both of these quantities, we use the fact that the variance is the dot product of the $\mathbb{L}^2$ space. We have, for g and h in $\mathbb{L}^2(\mathbb{P}_X)$,

$$\frac{\partial}{\partial w}\langle g(X,w), h(X,w)\rangle = \langle \frac{\partial g}{\partial w}(X,w), h(X,w) + \langle g(X,w), \frac{\partial h}{\partial w}(X,w)\rangle.$$

Then, we infer that $\frac{\partial}{\partial w}Var\mathbb{E}[f_w(X)|X_i] = 2Cov(\mathbb{E}[\frac{\partial f}{\partial w}|X_i], \mathbb{E}[f_w(X)|X_i])$ and that $\frac{\partial}{\partial w}Varf_w(X) = 2Cov(\frac{\partial f}{\partial w}, f_w(X))$. By plugging these results in the previous equality, we obtain Equation 4.36 □

The complete framework around fair learning and complete results for classical fairness metrics are left for future works.

Global robustness

In sensitivity analysis, additional information on the model is obtained on a global scale through second-level GSA. As such, it seems straightforward to deploy the same tools in a Fairness framework to assert global impact of generative parameters on the Fairness property.

Note that, in Algorithmic Fairness, it is usual to consider binary variables. This may seem contradictory with the GSA framework which often consider inputs to be real-valued and continuous. Moreover, it has led the Fairness literature to develop tools that easily leverage this dichotomy "$S = 0$ vs. $S = 1$", as mentioned in previous chapters. However, it is common for the generative distribution of these binary inputs to be parameterized by continuous parameters – e.g. the proportion of a minority in a given population, the mean and variance of a given physical characteristic that can be above a threshold, so on and so forth. As such, the GSA2 framework proposed in the previous section is a good way to work with these parameters, as it uses natural extensions of classical Fairness tools.

Nonetheless, when considering classical fairness metrics such as Disparate Impact, Equality of Odds, etc – namely, indicators based on the confusion matrix

of the classifier f – one can see that these metrics are already robust in the sense that they do not depend on parameters of the inputs distribution. Indeed, in the definition of these metrics appear a conditioning on the sensitive categorical feature in which information from the generative parameters is lost.

However, over the years, other metrics have been proposed, and we refer the reader to Chapter 2 on this subject. Additionally, sensitive features can be continuous. In both of these premises, second-level analysis can be done and will yield additional information, as it does in the GSA framework.

Chapter 5

Conclusions et perspectives

Au long de ce manuscrit, nous avons étudié les aspects distributionnels de l'Analyse de Sensibilité et de l'Équité Algorithmique. Le but de cet étude est d'obtenir, *in fine*, des méthodes utilisables sur des algorithmes d'Intelligence Artifitielle permettant à la fois de gagner en interprétabilité pour des modèles de type «boîte noire» mais aussi de détecter d'éventuels comportements discriminants exhibés par ces modèles.

Nous avons vu que l'Analyse de Sensibilité, dans une optique d'explicabilité, est un moyen de récupérer une information globale sur un modèle *a priori* inaccessible par le biais d'indicateurs quantifiants l'importance de chaque variable d'entrée sur la sortie. Plusieurs méthodes existent, chacune étant plus ou moins adaptée pour répondre à une question précise. Ainsi, l'étude du modèle par le biais de plusieurs de ces indices permet une compréhension fine du comportement global de l'algorithme.

De manière similaire, un aspect de l'Équité Algorithmique est de déterminer si des variables sensibles sont influentes sur la sortie d'un modèle, ou sur diverses caractéristiques similaires (e.g. *accuracy* dans le cas d'un classifieur). Ceci se fait par le biais de métriques ou d'indicateurs qui permettent la détection d'un comportement discriminant, et le cas échéant par la quantification de la force de ce biais discriminant.

Ainsi, ces deux domaines, bien que possédant chacun ses spécificités, se rejoignent et des techniques de l'un sont adaptables pour obtenir des outils pertinents dans l'autre. La description de ces différentes techniques est le sujet du Chapitre 1 et les points communs qui les rejoignent dans le cadre distributionnel le sujet du Chapitre 2. Pour aller plus loin, il serait possible de regarder si ce type d'analogie entre Explicabilité et Équité tient également dans un cadre local, c'est-à-dire pour des méthodes apportant de l'information autour d'un individu donné.

Par la suite (Chapitre 3), nous nous sommes intéressés aux difficultés qui peuvent survenir lors de l'utilisation d'un métamodèle (une approximation d'un vrai modèle qui reste inaccessible). En effet, en étudiant un métamodèle, les indicateurs mentionnés au-dessus (que ce soit en Analyse de Sensibilité ou en

Équité Algorithmique) peuvent prendre des valeurs différentes que celles pour le vrai modèles. Nous avons vu que cette différence se quantifie et peut se contrôler par le biais de bornes exactes ou probabilistes suivant le type d'approximation choisie. Bien que se faisant au prix de données additionnelles et d'un cadre un peu plus lourd, ces résultats ouvrent la piste de l'audit. À terme, il serait possible de quantifier la force d'un potentiel biais discriminant d'un algorithme sans même avoir accès à celui-ci mais en utilisant seulement une approximation, avec un coût additionnel. Néanmoins, un certain nombre de problématiques demeurent, notamment le choix de la métrique ou encore la question de la confiance en l'approximation fournie lors de l'audit. Ces questions sont autant mathématiques que juridiques et sociales puisque, et nous l'avons mentionné au début de ce manuscrit, ces sujets sont autant techniques qu'éthiques.

Enfin, dans le Chapitre 4, nous avons considéré deux problèmes différents. Tout d'abord, nous avons considéré un nouvel estimateur d'indices classiques paru récemment dans la littérature. Cet estimateur, par le biais d'une approche novatrice basées sur des rankings, permet de considérablement réduire le coût d'estimation. De plus, elle ne nécessite pas de nouveaux points ni de schéma d'estimation particulier, ce qui en fait un outil crucial pour l'industrie. Nous avons prouvé des propriétés statistiques (inégalités de concentration, normalité asymptotique pour sortie multivariée) et donné des indices et résultats négatifs pour d'autres propriétés de cet estimateur. Une perspective serait de mener à terme ces travaux en prouvant le reste des résultats, notamment des théorèmes de type Berry-Esseen afin d'avoir à disposition une analyse complète de cet outil. Ensuite, nous avons considéré le problème de la robustesse distributionnelle, en rappelant des résultats existants dans la littérature et en expliquant la signification de tels résultats dans le cadre de l'Équité Algorithmique. Une perspective, encore une fois, serait de mener à bien cette analyse afin de fournir des outils permettant de quantifier de manière globale les changements qu'implique une incertitude distributionnelle et de potentiellement corriger ces changements.

Pour terminer, cette thèse se positionnent à l'interface entre deux domaines qui sont le résultat de besoins éthiques et sociaux autant que scientifiques face à une technologie naissante et hautement ancrée dans un formalisme mathématique et informatique. Une perspective future peut être de continuer à répondre à ces besoins et de favoriser les échanges entre ces différents domaines. À l'instar du parallèle tiré ici entre Explicabilité et Équité, il est tout à fait possible que d'autres «vertues» cherchées pour le Machine Learning soient liées de manière similaire (e.g. les problématiques de *confidentialité*).

Bibliography

[] *Recent Trends in Learning From Data*. URL: https://link.springer.com/book/10.1007/978-3-030-43883-8 (visited on 07/04/2022).

[AC21] Mona Azadkia and Sourav Chatterjee. "A simple measure of conditional dependence". In: *arXiv:1910.12327 [cs, math, stat]* (Mar. 28, 2021). arXiv: 1910.12327. URL: http://arxiv.org/abs/1910.12327 (visited on 02/09/2022).

[ACW16] Radosław Adamczak, Djalil Chafai, and Paweł Wolff. "Circular law for random matrices with exchangeable entries". In: *Random Structures & Algorithms* 48.3 (2016), pp. 454–479.

[Aïv+19] Ulrich Aïvodji et al. "Fairwashing: the risk of rationalization". In: *International Conference on Machine Learning*. PMLR. 2019, pp. 161–170.

[Alb19] Mélisande Albert. "Concentration inequalities for randomly permuted sums". In: *High Dimensional Probability VIII*. Springer, 2019, pp. 341–383.

[Bac13] François Bachoc. "Parametric estimation of covariance function in Gaussian-process based Kriging models. Application to uncertainty quantification for computer experiments". PhD thesis. Université Paris-Diderot-Paris VII, 2013.

[Bar+18] Eustasio del Barrio et al. "Obtaining fairness using optimal transport theory". In: *arXiv:1806.03195 [math, stat]* (July 18, 2018). arXiv: 1806.03195. URL: http://arxiv.org/abs/1806.03195 (visited on 10/27/2020).

[Bar02] Yannick Baraud. "Model selection for regression on a random design". In: *ESAIM: Probability and Statistics* 6 (2002), pp. 127–146. ISSN: 1292-8100, 1262-3318. DOI: 10.1051/ps:2002007. URL: http://www.esaim-ps.org/10.1051/ps:2002007 (visited on 02/01/2022).

[BBG19] Julien Bect, François Bachoc, and David Ginsbourger. "A supermartingale approach to Gaussian process based sequential design of experiments". In: *Bernoulli* 25.4A (2019), pp. 2883–2919.

[BC22] Laurence Barry and Arthur Charpentier. "The Fairness of Machine Learning in Insurance: New Rags for an Old Man?" In: *arXiv preprint arXiv:2205.08112* (2022).

[Bén+21] Clément Bénesse et al. "Fairness seen as Global Sensitivity Analysis". In: *arXiv:2103.04613 [math, stat]* (Sept. 20, 2021). arXiv: 2103.04613. URL: http://arxiv.org/abs/2103.04613 (visited on 03/29/2022).

[Bes+20] Philippe Besse et al. *A survey of bias in Machine Learning through the prism of Statistical Parity for the Adult Data Set*. Apr. 6, 2020. DOI: 10.48550/arXiv.2003.14263. arXiv: 2003.14263[cs,stat]. URL: http://arxiv.org/abs/2003.14263 (visited on 07/04/2022).

[BGL20] Eustasio del Barrio, Paula Gordaliza, and Jean-Michel Loubes. "Review of Mathematical frameworks for Fairness in Machine Learning". In: *arXiv:2005.13755 [cs, stat]* (May 26, 2020). arXiv: 2005.13755. URL: http://arxiv.org/abs/2005.13755 (visited on 10/27/2020).

[BGW19] Benjamin R Baer, Daniel E Gilbert, and Martin T Wells. "Fairness criteria through the lens of directed acyclic graphical models". In: *arXiv preprint arXiv:1906.11333* (2019).

[Big+21] Daniele Bigoni et al. "Nonlinear dimension reduction for surrogate modeling using gradient information". In: *arXiv preprint arXiv:2102.10351* (2021).

[Bon+21] Stephan Bongers et al. "Foundations of Structural Causal Models with Cycles and Latent Variables". In: *The Annals of Statistics* 49.5 (Oct. 1, 2021). ISSN: 0090-5364. DOI: 10.1214/21-AOS2064. arXiv: 1611.06221[cs,stat]. URL: http://arxiv.org/abs/1611.06221 (visited on 07/04/2022).

[BT04] Alain Berlinet and Christine Thomas-Agnan. "A Collection of Examples". In: *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Ed. by Alain Berlinet and Christine Thomas-Agnan. Boston, MA: Springer US, 2004, pp. 293–343. ISBN: 978-1-4419-9096-9. DOI: 10.1007/978-1-4419-9096-9_7. URL: https://doi.org/10.1007/978-1-4419-9096-9_7 (visited on 07/04/2022).

[CGP11] Gaëlle Chastaing, Fabrice Gamboa, and Clémentine Prieur. "Generalized Hoeffding-Sobol Decomposition for Dependent Variables -Application to Sensitivity Analysis". In: *arXiv:1112.1788 [math, stat]* (Dec. 8, 2011). arXiv: 1112.1788. URL: http://arxiv.org/abs/1112.1788 (visited on 07/03/2019).

[CGS10] G. Carlier, A. Galichon, and F. Santambrogio. "From Knothe's Transport to Brenier's Map and a Continuation Method for Optimal Transport". In: *SIAM Journal on Mathematical Analysis* 41.6 (Jan. 1, 2010). Publisher: Society for Industrial and Applied Mathematics, pp. 2554–2576. ISSN: 0036-1410. DOI: 10.1137/080740647. URL: https://epubs.siam.org/doi/abs/10.1137/080740647 (visited on 10/27/2020).

[Cha20] Sourav Chatterjee. "A New Coefficient of Correlation". In: *Journal of the American Statistical Association* 0.0 (Apr. 27, 2020). Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01621459.2020.1758115, pp. 1–21. ISSN: 0162-1459. DOI: 10.1080/01621459.2020.1758115. URL: https://doi.org/10.1080/01621459.2020.1758115 (visited on 10/27/2020).

[Che+20] Christophe Chesneau et al. "Nonparametric estimation in a regression model with additive and multiplicative noise". In: *Journal of Computational and Applied Mathematics* 380 (Dec. 2020), p. 112971. ISSN: 03770427. DOI: 10.1016/j.cam.2020.112971. arXiv: 1906.07695. URL: http://arxiv.org/abs/1906.07695 (visited on 02/07/2022).

[Che12] Christophe Chesneau. "On the adaptive wavelet deconvolution of a density for strong mixing sequences". In: *Journal of the Korean Statistical Society* 41.4 (Dec. 1, 2012), pp. 423–436. ISSN: 1226-3192. DOI: 10.1016/j.jkss.2012.01.005. URL: https://www.sciencedirect.com/science/article/pii/S1226319212000191 (visited on 07/04/2022).

[Chi+20] Silvia Chiappa et al. "A General Approach to Fairness with Optimal Transport." In: *AAAI*. 2020, pp. 3633–3640.

[Chi19] Silvia Chiappa. "Path-Specific Counterfactual Fairness". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 33.1 (July 17, 2019). Number: 01, pp. 7801–7808. ISSN: 2374-3468. DOI: 10.1609/aaai.v33i01.33017801. URL: https://ojs.aaai.org/index.php/AAAI/article/view/4777 (visited on 07/04/2022).

[Cho17] Alexandra Chouldechova. "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments". In: *Big Data* 5.2 (June 2017). Publisher: Mary Ann Liebert, Inc., publishers, pp. 153–163. ISSN: 2167-6461. DOI: 10.1089/big.2016.0047. URL: https://www.liebertpub.com/doi/10.1089/big.2016.0047 (visited on 07/04/2022).

[Chz+20] Evgenii Chzhen et al. "Fair Regression via Plug-in Estimator and Recalibration With Statistical Guarantees". In: (2020), p. 35.

[CJ78] Herman Callaert and Paul Janssen. "The Berry-Esseen theorem for U-statistics". In: *The Annals of Statistics* (1978), pp. 417–421.

[Cre] Kimberle Crenshaw. "Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics". In: (), p. 31.

[Csi67] I. Csiszar. "Information-type measures of difference of probability distributions and indirect observation". In: *Studia Scientiarum Mathematicarum Hungarica* (Jan. 1, 1967). ISSN: 0081-6906. URL: https://www.scinapse.io (visited on 07/04/2022).

[Da +21] Sebastien Da Veiga et al. *Basics and Trends in Sensitivity Analysis | SIAM Digital Library*. SIAM, 2021. URL: https://epubs.siam.org/doi/book/10.1137/1.9781611976694 (visited on 07/04/2022).

[Da 15] Sebastien Da Veiga. "Global sensitivity analysis with dependence measures". In: *Journal of Statistical Computation and Simulation* 85.7 (May 3, 2015), pp. 1283–1305. ISSN: 0094-9655, 1563-5163. DOI: 10.1080/00949655.2014.945932. URL: http://www.tandfonline.com/doi/abs/10.1080/00949655.2014.945932 (visited on 06/26/2019).

[Da 21] Sébastien Da Veiga. "Kernel-based ANOVA decomposition and Shapley effects–Application to global sensitivity analysis". In: *arXiv preprint arXiv:2101.05487* (2021).

[Dau95] Ingrid Daubechies. "Ten Lectures on Wavelets, CBMS-NSF regional conference series in applied mathematics, vol. 61, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1992". In: *MR MR1162107 (93e: 42045)* 33 (1995), p. 37.

[DBM22] Julien Demange-Chryst, François Bachoc, and Jérôme Morio. "Shapley effect estimation in reliability-oriented sensitivity analysis with correlated inputs by importance sampling". In: *arXiv preprint arXiv:2202.12679* (2022).

[DC17] Hongzhe Dai and Zhenggang Cao. "A wavelet support vector machine-based neural network metamodel for structural reliability assessment". In: *Computer-Aided Civil and Infrastructure Engineering* 32.4 (2017), pp. 344–357.

[De-+19] Maria De-Arteaga et al. "Bias in bios: A case study of semantic representation bias in a high-stakes setting". In: *proceedings of the Conference on Fairness, Accountability, and Transparency*. 2019, pp. 120–128.

[Deh17] Pierre Dehez. "On Harsanyi Dividends and Asymmetric Values". In: *International Game Theory Review* 19.3 (Sept. 1, 2017). Publisher: World Scientific Publishing Co., p. 1750012. ISSN: 0219-1989. DOI: 10.1142/S0219198917500128. URL: https://www.worldscientific.com/doi/abs/10.1142/S0219198917500128 (visited on 01/27/2022).

[DG17] Dheeru Dua and Casey Graff. *UCI Machine Learning Repository*. 2017. URL: http://archive.ics.uci.edu/ml.

[DP88] Ronald A. Devore and Vasil A. Popov. "Interpolation of Besov Spaces". In: *Transactions of the American Mathematical Society* 305.1 (1988). Publisher: American Mathematical Society, pp. 397–414. ISSN: 0002-9947. DOI: 10.2307/2001060. URL: https://www.jstor.org/stable/2001060 (visited on 07/04/2022).

[Dwo+11] Cynthia Dwork et al. "Fairness Through Awareness". In: *arXiv:1104.3913 [cs]* (Nov. 28, 2011). arXiv: 1104.3913. URL: http://arxiv.org/abs/1104.3913 (visited on 10/27/2020).

[Dwo+20] Cynthia Dwork et al. "Abstracting fairness: Oracles, metrics, and interpretability". In: *arXiv preprint arXiv:2004.01840* (2020).

[FBL16] Anthony W Flores, Kristin Bechtel, and Christopher T Lowenkamp. "False positives, false negatives, and false analyses: A rejoinder to machine bias: There's software used across the country to predict future criminals. and it's biased against blacks". In: *Fed. Probation* 80 (2016), p. 38.

[Fer+22] Julien Ferry et al. "Improving fairness generalization through a sample-robust optimization method". In: *Machine Learning* (2022), pp. 1–62.

[FKL21] Jean-Claude Fort, Thierry Klein, and Agnès Lagnoux. "Global sensitivity analysis and Wasserstein spaces". In: *SIAM/ASA Journal on Uncertainty Quantification* 9.2 (2021), pp. 880–921.

[FKR16] Jean-Claude Fort, Thierry Klein, and Nabil Rachdi. "New sensitivity analysis subordinated to a contrast". In: *Communications in Statistics-Theory and Methods* 45.15 (2016), pp. 4349–4364.

[Fou+19] James Foulds et al. *An Intersectional Definition of Fairness*. Sept. 10, 2019. DOI: 10.48550/arXiv.1807.08362. arXiv: 1807.08362[cs, stat]. URL: http://arxiv.org/abs/1807.08362 (visited on 07/04/2022).

[FRF] Christopher Frye, Colin Rowat, and Ilya Feige. "Asymmetric Shapley values: incorporating causal knowledge into model-agnostic explainability". In: (), p. 11.

[Gam+14] Fabrice Gamboa et al. "Sensitivity analysis for multidimensional and functional outputs". In: *Electronic Journal of Statistics* 8.1 (2014), pp. 575–603.

[Gam+16] Fabrice Gamboa et al. "Statistical inference for Sobol pick-freeze Monte Carlo method". In: *Statistics* 50.4 (2016), pp. 881–902.

[Gam+20] Fabrice Gamboa et al. "Global Sensitivity Analysis: a new generation of mighty estimators based on rank statistics". In: *arXiv:2003.01772 [math, stat]* (Mar. 3, 2020). arXiv: 2003.01772. URL: http://arxiv.org/abs/2003.01772 (visited on 10/27/2020).

[Gam+21] Fabrice Gamboa et al. "Sensitivity analysis in general metric spaces". In: *Reliability Engineering & System Safety* 212 (Aug. 2021), p. 107611. ISSN: 09518320. DOI: 10.1016/j.ress.2021.107611. URL: https://linkinghub.elsevier.com/retrieve/pii/S0951832021001563 (visited on 07/04/2022).

[GBM22] Bishwamittra Ghosh, Debabrota Basu, and Kuldeep S Meel. "How Biased is Your Feature?: Computing Fairness Influence Functions with Global Sensitivity Analysis". In: *arXiv preprint arXiv:2206.00667* (2022).

[GG19] Hila Gonen and Yoav Goldberg. "Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them". In: *arXiv preprint arXiv:1903.03862* (2019).

[GKK18] AmirEmad Ghassami, Sajad Khodadadian, and Negar Kiyavash. *Fairness in Supervised Learning: An Information Theoretic Approach*. July 29, 2018. arXiv: 1801.04378[cs,math,stat]. URL: http://arxiv.org/abs/1801.04378 (visited on 07/04/2022).

[GKL18] Fabrice Gamboa, Thierry Klein, and Agnès Lagnoux. "Sensitivity analysis based on Cramér von Mises distance". In: *SIAM/ASA Journal on Uncertainty Quantification* 6.2 (Apr. 2018). Publisher: ASA, American Statistical Association, pp. 522–548. DOI: 10.1137/15M1025621. URL: https://hal.archives-ouvertes.fr/hal-01163393 (visited on 02/09/2022).

[Gra+19] Vincent Grari et al. "Fairness-Aware Neural R\'eyni Minimization for Continuous Features". In: *arXiv:1911.04929 [cs, stat]* (Nov. 12, 2019). arXiv: 1911.04929. URL: http://arxiv.org/abs/1911.04929 (visited on 10/27/2020).

[Gra15] Mathilde Grandjacques. "Analyse de sensibilité pour des modèles stochastiques à entrées dépendantes : application en énergétique du bâtiment". PhD thesis. Université Grenoble Alpes, Nov. 9, 2015. URL: https://tel.archives-ouvertes.fr/tel-01266397 (visited on 10/27/2020).

[Gre+] Arthur Gretton et al. "Kernel Methods for Measuring Independence". In: (), p. 55.

[GSW96] Paul Gustafson, C Srinivasan, and Larry Wasserman. "Local sensitivity analysis". In: *Bayesian statistics* 5 (1996), pp. 197–210.

[Här+12] Wolfgang Härdle et al. *Wavelets, approximation, and statistical applications*. Vol. 129. Springer Science & Business Media, 2012.

[HDV20] James M. Hickey, Pietro G. Di Stefano, and Vlasios Vasileiou. "Fairness by Explicability and Adversarial SHAP Learning". In: *arXiv:2003.05330 [cs, stat]* (June 26, 2020). arXiv: 2003.05330. URL: http://arxiv.org/abs/2003.05330 (visited on 10/27/2020).

[HG19] Joseph Hart and Pierre A Gremaud. "Robustness of the Sobol'indices to distributional uncertainty". In: *International Journal for Uncertainty Quantification* 9.5 (2019).

[ICI21] Marouane Il Idrissi, Vincent Chabridon, and Bertrand Iooss. "Developments and applications of Shapley effects to reliability-oriented sensitivity analysis with correlated inputs". In: *arXiv:2101.08083 [math, stat]* (May 19, 2021). arXiv: 2101.08083. URL: http://arxiv.org/abs/2101.08083 (visited on 01/27/2022).

[IL15] Bertrand Iooss and Paul Lemaître. "A Review on Global Sensitivity Analysis Methods". In: *Uncertainty Management in Simulation-Optimization of Complex Systems: Algorithms and Applications*. Ed. by Gabriella Dellino and Carlo Meloni. Operations Research/Computer Science Interfaces Series. Boston, MA: Springer US, 2015, pp. 101–122. ISBN: 978-1-4899-7547-8. DOI: 10.1007/978-1-4899-7547-8_5.

URL: https://doi.org/10.1007/978-1-4899-7547-8_5 (visited on 10/27/2020).

[IP17] Bertrand Iooss and Clémentine Prieur. "Shapley effects for sensitivity analysis with dependent inputs: comparisons with Sobol' indices, numerical estimation and applications". July 2017. URL: https://hal.inria.fr/hal-01556303 (visited on 01/27/2022).

[JLD06] Julien Jacques, Christian Lavergne, and Nicolas Devictor. "Sensitivity analysis in presence of model uncertainty and correlated inputs". In: *Reliability Engineering & System Safety*. The Fourth International Conference on Sensitivity Analysis of Model Output (SAMO 2004) 91.10 (Oct. 1, 2006), pp. 1126–1134. ISSN: 0951-8320. DOI: 10.1016/j.ress.2005.11.047. URL: http://www.sciencedirect.com/science/article/pii/S0951832005002231 (visited on 10/27/2020).

[JNP14] Alexandre Janon, Maëlle Nodet, and Clémentine Prieur. "Uncertainties assessment in global sensitivity indices estimation from metamodels". In: *International Journal for Uncertainty Quantification* 4.1 (2014).

[Kil+17] Niki Kilbertus et al. "Avoiding Discrimination through Causal Reasoning". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017, pp. 656–666. URL: https://proceedings.neurips.cc/paper/2017/file/f5f8590cd58a54e94377e6ae2eded4d9-Paper.pdf.

[KMM15] Matthew Kay, Cynthia Matuszek, and Sean A Munson. "Unequal representation and gender stereotypes in image search results for occupations". In: *Proceedings of the 33rd annual acm conference on human factors in computing systems*. 2015, pp. 3819–3828.

[KMR16] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. "Inherent trade-offs in the fair determination of risk scores". In: *arXiv preprint arXiv:1609.05807* (2016).

[Kus+18] Matt J. Kusner et al. *Counterfactual Fairness*. Mar. 8, 2018. DOI: 10.48550/arXiv.1703.06856. arXiv: 1703.06856[cs,stat]. URL: http://arxiv.org/abs/1703.06856 (visited on 07/04/2022).

[Lar+21a] Lucas de Lara et al. *Transport-based Counterfactual Models*. Aug. 30, 2021. DOI: 10.48550/arXiv.2108.13025. arXiv: 2108.13025[cs]. URL: http://arxiv.org/abs/2108.13025 (visited on 07/04/2022).

[Lar+21b] Lucas de Lara et al. *Counterfactual Models: The Mass Transportation Viewpoint*. May 4, 2021. DOI: 10.1145/3351095.3372845. URL: https://hal.archives-ouvertes.fr/hal-03216124 (visited on 07/04/2022).

[Led01] Michel Ledoux. *The concentration of measure phenomenon*. 89. American Mathematical Soc., 2001.

[Lév55] Paul Lévy. "Théorie de l'Addition de Variables Aléatoires." In: *The Mathematical Gazette* 39.330 (Dec. 1955). Publisher: Cambridge University Press, pp. 344–344. ISSN: 0025-5572, 2056-6328. DOI: 10.2307/3608623. URL: https://www.cambridge.org/core/journals/mathematical-gazette/article/abs/theorie-de-laddition-de-variables-aleatoires-by-paul-levy-pp-xx-385-second-edition-1954-1200f-gauthiervillars-paris/73294A1332D0CC9727CF681F859605D1 (visited on 07/04/2022).

[Li+19] Zhu Li et al. *Kernel Dependence Regularizers and Gaussian Processes with Applications to Algorithmic Fairness*. Nov. 11, 2019. DOI: 10.48550/arXiv.1911.04322. arXiv: 1911.04322[cs,stat]. URL: http://arxiv.org/abs/1911.04322 (visited on 07/04/2022).

[LLR20] Thibaut Le Gouic, Jean-Michel Loubes, and Philippe Rigollet. *Projection to Fairness in Statistical Learning*. Publication Title: arXiv e-prints ADS Bibcode: 2020arXiv200511720L Type: article. May 1, 2020. URL: https://ui.adsabs.harvard.edu/abs/2020arXiv200511720L (visited on 07/04/2022).

[LMS17] Loic Le Gratiet, Stefano Marelli, and Bruno Sudret. "Metamodel-based sensitivity analysis: polynomial chaos expansions and Gaussian processes". In: *Handbook of uncertainty quantification*. Springer, 2017, pp. 1289–1325.

[Lüt+21] Nora Lüthen et al. "Global sensitivity analysis using derivative-based sparse Poincar\'e chaos expansions". In: *arXiv preprint arXiv:2107.00394* (2021).

[LWM21] Johann Laux, Sandra Wachter, and Brent Mittelstadt. "Taming the few: Platform regulation, independent audits, and the risks of capture created by the DMA and DSA". In: *Computer Law & Security Review* 43 (2021), p. 105613.

[MCK] Jérémie Mary, Clément Calauzènes, and Noureddine El Karoui. "Fairness-Aware Learning for Continuous Attributes and Treatments". In: (), p. 10.

[Mey+] Anouar Meynaoui et al. "Adaptive test of independence based on HSIC measures". In: (), p. 63.

[MML19] Anouar Meynaoui, Amandine Marrel, and Béatrice Laurent. "New statistical methodology for second level global sensitivity analysis". In: *arXiv preprint arXiv:1902.07030* (2019).

[Mor+20] Giulio Morina et al. *Auditing and Achieving Intersectional Fairness in Classification Problems*. June 8, 2020. arXiv: 1911.01468[cs,stat]. URL: http://arxiv.org/abs/1911.01468 (visited on 07/04/2022).

[MT12] Thierry A. Mara and Stefano Tarantola. "Variance-based sensitivity indices for models with dependent inputs". In: *Reliability Engineering & System Safety* 107 (Nov. 2012), pp. 115–121. ISSN: 09518320. DOI: 10.1016/j.ress.2011.08.008. URL: https://linkinghub.elsevier.com/retrieve/pii/S0951832011001724 (visited on 10/27/2020).

[MTA15] Thierry A. Mara, Stefano Tarantola, and Paola Annoni. "Non-parametric methods for global sensitivity analysis of model output with dependent inputs". In: *Environmental Modelling & Software* 72 (Oct. 2015), pp. 173–183. ISSN: 13648152. DOI: 10.1016/j.envsoft.2015.07.010. URL: https://linkinghub.elsevier.com/retrieve/pii/S1364815215300153 (visited on 10/22/2019).

[ODP19] Luca Oneto, Michele Donini, and Massimiliano Pontil. *General Fair Empirical Risk Minimization*. Dec. 27, 2019. DOI: 10.48550/arXiv.1901.10080. arXiv: 1901.10080[cs,stat]. URL: http://arxiv.org/abs/1901.10080 (visited on 07/04/2022).

[OP17] Art B Owen and Clémentine Prieur. "On Shapley value for measuring importance of dependent inputs". In: *SIAM/ASA Journal on Uncertainty Quantification* (2017). Publisher: ASA, American Statistical Association. URL: https://hal.archives-ouvertes.fr/hal-01379188 (visited on 02/09/2022).

[Owe14] Art B. Owen. "Sobol' Indices and Shapley Value". In: *SIAM/ASA Journal on Uncertainty Quantification* 2.1 (Jan. 2014), pp. 245–251. ISSN: 2166-2525. DOI: 10.1137/130936233. URL: http://epubs.siam.org/doi/10.1137/130936233 (visited on 10/21/2019).

[Pan21] Ivan Panin. "Risk of estimators for Sobol' sensitivity indices based on metamodels". In: *Electronic Journal of Statistics* 15.1 (Jan. 2021). Publisher: Institute of Mathematical Statistics and Bernoulli Society, pp. 235–281. ISSN: 1935-7524, 1935-7524. DOI: 10.1214/20-EJS1793. URL: https://projecteuclid.org/journals/electronic-journal-of-statistics/volume-15/issue-1/Risk-of-estimators-for-Sobol-sensitivity-indices-based-on-metamodels/10.1214/20-EJS1793.full (visited on 11/22/2021).

[Pea09] Judea Pearl. *CAUSALITY, 2nd Edition, 2009*. Cambridge university press, 2009. URL: http://bayes.cs.ucla.edu/BOOK-2K/ (visited on 07/04/2022).

[PM09] Iosif Pinelis and Raymond Molzon. "Berry-Esséen bounds for general nonlinear statistics, with applications to Pearson's and non-central Student's and Hotelling's". In: *arXiv preprint arXiv:0906.0177* (2009).

[Raw04] John Rawls. "A theory of justice". In: *Ethics*. Routledge, 2004, pp. 229–234.

[Rén59] A. Rényi. "On measures of dependence". In: *Acta Mathematica Academiae Scientiarum Hungarica* 10.3 (Sept. 1, 1959), pp. 441–451. ISSN: 1588-2632. DOI: 10.1007/BF02024507. URL: https://doi.org/10.1007/BF02024507 (visited on 07/04/2022).

[Ris+21] Laurent Risser et al. "Tackling Algorithmic Bias in Neural-Network Classifiers using Wasserstein-2 Regularization". In: *arXiv:1908.05783 [cs, stat]* (Nov. 12, 2021). arXiv: 1908.05783. URL: http://arxiv.org/abs/1908.05783 (visited on 05/11/2022).

[RM18] Hugo Raguet and Amandine Marrel. "Target and conditional sensitivity analysis with emphasis on dependence measures". In: *arXiv preprint arXiv:1801.10047* (2018).

[Ros52] Murray Rosenblatt. "Remarks on a Multivariate Transformation". In: *Annals of Mathematical Statistics* 23.3 (Sept. 1952). Publisher: Institute of Mathematical Statistics, pp. 470–472. ISSN: 0003-4851, 2168-8990. DOI: 10.1214/aoms/1177729394. URL: https://projecteuclid.org/euclid.aoms/1177729394 (visited on 10/27/2020).

[Rot+20] Dominik Rothenhäusler et al. *Anchor regression: heterogeneous data meets causality*. May 8, 2020. DOI: 10.48550/arXiv.1801.06229. arXiv: 1801.06229[stat]. URL: http://arxiv.org/abs/1801.06229 (visited on 07/04/2022).

[Rou+14] Olivier Roustant et al. "Crossed-derivative based sensitivity measures for interaction screening". In: *Mathematics and Computers in Simulation* 105 (2014), pp. 105–118.

[RSG16] Marco Tulio Ribeiro, Sameer Singh, and Carlos 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*. 2016, pp. 1135–1144.

[Sal+10] Andrea Saltelli et al. "Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index". In: *Computer physics communications* 181.2 (2010), pp. 259–270.

[Sam+21] Wojciech Samek et al. "Explaining deep neural networks and beyond: A review of methods and applications". In: *Proceedings of the IEEE* 109.3 (2021), pp. 247–278.

[Sch19] Bernhard Schölkopf. *Causality for Machine Learning*. Dec. 23, 2019. DOI: 10.48550/arXiv.1911.10500. arXiv: 1911.10500[cs,stat]. URL: http://arxiv.org/abs/1911.10500 (visited on 07/04/2022).

[Sha16] Lloyd S Shapley. "17. A value for n-person games". In: *Contributions to the Theory of Games (AM-28), Volume II*. Princeton University Press, 2016, pp. 307–318.

[Smo+] Alex Smola et al. "A Hilbert Space Embedding for Distributions". In: (), p. 20.

[Sob] I M Sobol. "On sensitivity estimation for nonlinear mathematical models". In: (), p. 8.

[Vaa98] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press, 1998. ISBN: 978-0-521-78450-4. DOI: 10.1017/CBO9780511802256. URL: https://www.cambridge.org/core/books/asymptotic-statistics/A3C7DAD3F7E66A1FA60E9C8FE132EE1D (visited on 07/04/2022).

[Van00] Aad W Van der Vaart. *Asymptotic statistics*. Vol. 3. Cambridge university press, 2000.

[Was19] Anne L. Washington. *How to Argue with an Algorithm: Lessons from the COMPAS ProPublica Debate*. Rochester, NY, Feb. 4, 2019. URL: https://papers.ssrn.com/abstract=3357874 (visited on 07/04/2022).

[WJ21] Wenjia Wang and Bing-Yi Jing. "Convergence of Gaussian process regression: Optimality, robustness, and relationship with kernel ridge regression". In: *arXiv preprint arXiv:2104.09778* (2021).

[WM19] Robert C. Williamson and Aditya Krishna Menon. "Fairness risk measures". In: *arXiv:1901.08665 [cs, stat]* (Jan. 24, 2019). arXiv: 1901.08665. URL: http://arxiv.org/abs/1901.08665 (visited on 11/20/2019).

[WMR17] Sandra Wachter, Brent Mittelstadt, and Chris Russell. "Counterfactual explanations without opening the black box: Automated decisions and the GDPR". In: *Harv. JL & Tech.* 31 (2017), p. 841.

[Zaf+17] Muhammad Bilal Zafar et al. "Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment". In: *Proceedings of the 26th International Conference on World Wide Web*. Apr. 3, 2017, pp. 1171–1180. DOI: 10.1145/3038912.3052660. arXiv: 1610.08452[cs,stat]. URL: http://arxiv.org/abs/1610.08452 (visited on 07/04/2022).

List of Figures

Appendix A

Appendix

A.1 Lévy-Rosemblatt theorem and associated mappings

The aim of the Lévy-Rosenblatt transform is to find a transport map between the correlated $\mathbf{X}$ and independent uniform variables $\mathbf{U} \in \mathbb{R}^p$. From now, we assume the distribution of $\mathbf{X}$ to be absolutely continuous.

Theorem A.1.1 (Lévy-Rosemblatt theorem,[Lév55; Ros52])**.** *: there is a bijection (denoted "RT" for Rosemblatt transform) between $p(\mathbf{X})$ and p independent uniform random variables*

$$(X_i, (X_{i+1}|X_i), \cdots, (X_{i-1}|X_{\sim(i-1)})) \sim p_{\mathbf{X}} \xrightarrow{RT} (U_1^i, \cdots, U_p^i) \sim \mathcal{U}^p(0,1). \tag{A.1}$$

Example 8. *In the following, we will always be interested in two groups of variables: the sensitive variable X_i and the rest of the variables $X_{\sim i}$. Therefore, it may help to understand the special case where $\mathbf{X} = (X_1, X_2)$ since it encapsulates all the difficulty. In this case, we have two different ways to decompose $p_{\mathbf{X}}$.*

(i) If we decompose $p_{\mathbf{X}}$ as $p_{X_1} \times p_{X_2|X_1}$, then we can map this to (U_1^1, U_2^1). With this mapping, we can draw random variables with distributions p_{X_1} and $p_{X_2|X_1}$. For this, we need only to have access to independent uniform random variables and use the inverse Rosenblatt transform. We denote as F_T the cumulative distribution function of the random variable T. The inverse Rosenblatt transform is then given by

$$z_1 = F_{X_1}^{-1}(u_1^1) \tag{A.2}$$

$$z_2 = F_{X_2|X_1=x_1}^{-1}(u_2^1). \tag{A.3}$$

We first draw a random variable Z_1 with distribution p_{X_1} from an uniform random variable by quantile inversion. Now that we have this realisation

z_1, we have the second distribution $p_{X_2|X_1=z_1}$. We then draw a random variable Z_2 that follows the distribution $p_{X_2|X_1=z_1}$ and such that the couple (Z_1, Z_2) has the same distribution as (X_1, X_2). This random variable is similar to X_2 but does not contain its correlation with X_1.

(ii) Similarly, if we decompose $p_{\mathbf{X}}$ as $p_{X_2} \times p_{X_1|X_2}$, then we can map this to (U_1^2, U_2^2).

Note that the only case where these two mappings are similar is when X_1 and X_2 are independent. In that case, $p_{X_1} = p_{X_1|X_2}$ and $p_{X_2} = p_{X_2|X_1}$.

Several things need to be said about this transform.

Remark 20. *It enables to transform a set of possibly dependent random variables into a set of random variables without any dependencies. Moreover, for one such set of independent variables $\mathbf{U}^i$, there exists a function g_i square integrable such that $f(\mathbf{X}) = g_i(\mathbf{U}^i)$. One way to compute Sobol' indices for the output $f(\mathbf{X})$ is therefore to use the Hoeffding decomposition of $g_i(\mathbf{U}^i)$.*

Remark 21. *In terms of information, U_1^i carries as much information as X_i since $U_1^i = F_{X_i}(X_i)$. Note that this include the eventual dependency with other variables. This means that the Sobol' indices of U_1^i will correspond to the Sobol' indices of X_i as defined in the previous section. Meanwhile, the law of U_n^i is associated with the law of $X_{i-1}|X_{\sim(i-1)}$. This conditional distribution aim to capture all the remaining randomness in X_{i-1} when the intrinsic effects of the others inputs on it has been removed. Therefore, it has all the remaining information in the law of X_{i-1} when the contribution of the other variables are discarded.*

Remark 22. *The previous point is the reason why we do not need to consider all $n!$ possible Rosenblatt Transforms of $\mathbf{X}$. Since we are only interested in the information carried by a variable – with (X_i) – and by the law of this same variable without its dependencies in the other variables – with $(X_i|X_{\sim i})$, we are only interested in U_1^i and U_n^i, for all i. Therefore, we can without loss of generality, consider a cyclic permutation. That being said, if, for numerical reasons, other Rosenblatt transforms are easier to work with, there is no theoretical reasons not to use them.*

In the classic Sobol' analysis, for an input Y, we have two indices that quantify the influence of the considered feature on the output of the algorithm, namely the first order and total indices. Now, thanks to the Lévy-Rosemblatt, we have two different mappings of interest: the mapping from U_1^i to X_i that includes the intrinsic influence of other inputs over this particular input and the mapping from U_p^{i+1} to $X_i|X_{\sim i}$ that excludes these influences and shows the variation induced by this input on its own. These two different mappings will each lead to two indices (the Sobol' and Total Sobol' indices of U_1^i, and the ones of U_p^{i+1}) so every input X_i will be represented by four indices.

A.2 Estimates of extended Sobol' indices

We recall that in the independent Sobol' framework, for every input X_k, we have two different mappings: the mapping from U_1^k to X_k that includes the intrinsic influence of other inputs over this particular input and the mapping from U_p^{k+1} to $X_k|X_{\sim k}$ that excludes these influences and shows the variation of this input on its own. These two different mappings will each lead to two indices (the Sobol indices of U_1^k and the ones of U_p^{k+1}) so every input X_k will be represented by four indices, explained in the following subsection.

As seen previously, the four Sobol' indices for each variable $X_i, i \in [\![1, n]\!]$ are defined as followed:

$$Sob_{X_i} = \frac{V[\mathbb{E}[g_i(\mathbf{U}^i)|U_1^i]]}{V[g_i(\mathbf{U}^i)]} = \frac{V[\mathbb{E}[f(\mathbf{X})|X_i]]}{V[f(\mathbf{X})]} \tag{A.4}$$

$$SobT_{X_i} = \frac{\mathbb{E}[V[g_i(\mathbf{U}^i)|U_{\sim 1}^i]]}{V[g_i(\mathbf{U}^i)]} = \frac{\mathbb{E}[V[f(\mathbf{X})|Z_i]]}{V[f(\mathbf{X})]} \tag{A.5}$$

$$Sob_{X_i}^{ind} = \frac{V[\mathbb{E}[g_{i+1}(\mathbf{U}^{i+1})|U_p^{i+1}]]}{V[g_{i+1}(\mathbf{U}^{i+1})]} = \frac{V[\mathbb{E}[f(\mathbf{X})|Z_i]]}{V[f(\mathbf{X})]} \tag{A.6}$$

$$SobT_{X_i}^{ind} = \frac{\mathbb{E}[V[g_{i+1}(\mathbf{U}^{i+1})|U_{\sim p}^{i+1}]]}{V[g_{i+1}(\mathbf{U}^{i+1})]} = \frac{\mathbb{E}[V[f(\mathbf{X})|X_{\sim i}]]}{V[f(\mathbf{X})]} \tag{A.7}$$

We recall that these indices use the Rosemblatt transform, a bijection between independent uniforms and the distribution of the features. This bijection can be inverted to generate samples from uniforms. We denote the inverse of the Rosemblatt transform as IRT – Inverse Rosemblatt Transform. Thanks to the IRT, we can generate four samples:

$$\begin{aligned}
(u_1^i, \cdots, u_p^i) &\xrightarrow{IRT} \mathbf{x} = (x_i, \cdots, x_{i-1}) \sim p(\mathbf{X}),\\
(u_1^{i\prime}, \cdots, u_p^{i\prime}) &\xrightarrow{IRT} \mathbf{x}' = (x'_i, \cdots, x'_{i-1}) \sim p(\mathbf{X}),\\
(u_1^i, u_2^{i\prime}, \cdots, u_p^{i\prime}) &\xrightarrow{IRT} \mathbf{x}^i = (x_i, x'_{i+1} \cdots, x'_{i-1}) \sim p(X_i)p(X_{\sim i}|X_i),\\
(u_1^{i\prime}, \cdots, u_{p-1}^{i\prime}, u_p^i) &\xrightarrow{IRT} \mathbf{x}^{i-1} = (x'_i, x'_{i+1} \cdots, x_{i-1}) \sim p(X_{\sim i-1})p(X_{i-1}|X_{\sim i-1}).
\end{aligned} \tag{A.8}$$

Once we obtain, for each $i \in \{1, \cdots, p\}$, the four samples defined above, we can compute the estimators of the Sobol' and independent Sobol' indices as follows:

$$\begin{aligned}
\widehat{Sob_{X_i}} &= \frac{\frac{1}{N}\sum_{k=1}^{N} f(\mathbf{x}_k) \times \left(f(\mathbf{x}_k^i) - f(\mathbf{x}'_k)\right)}{\hat{V}} \\
\widehat{SobT^{ind}_{X_i}} &= \frac{\frac{1}{N}\sum_{k=1}^{N} \left(f(\mathbf{x}_k^{i-1}) - f(\mathbf{x}'_k)\right)^2}{2\hat{V}} \\
\widehat{Sob^{ind}_{i-1}} &= \frac{\frac{1}{N}\sum_{k=1}^{N} f(\mathbf{x}_k) \times \left(f(\mathbf{x}_k^{i-1}) - f(\mathbf{x}'_k)\right)}{\hat{V}} \\
\widehat{SobT_{X_i}} &= \frac{\frac{1}{N}\sum_{k=1}^{N} \left(f(\mathbf{x}_k^{i}) - f(\mathbf{x}'_k)\right)^2}{2\hat{V}},
\end{aligned} \tag{A.9}$$

where $\mathbf{x}_k^* = (x_{k,1}^*, \cdots, x_{k,p}^*)$ is the $k-$th Monte-Carlo trial in the sample $\mathbf{x}^*$, $k \in \{1, n\}$ and $\hat{V}$ is the total variance estimate that can be computed as the average of the total variances computed with each sample $\mathbf{x}^*$.

A.3 Central Limit Theorem for Sobol' indices

We recall the theorem 2.2.1 we presented in Section 2.2.

Theorem A.3.1. *Each index $\mathcal{S}$ in the equations* (2.6) *to* (2.9) *can be written as A/B and the corresponding estimate $\mathcal{S}_n$ can be written as A_n/B_n. For each of these indices, we have a central limit theorem:*

$$\sqrt{n}(\mathcal{S}_n - \mathcal{S}) \xrightarrow{D} \mathcal{N}(0, \sigma^2) \tag{A.10}$$

with σ^2 depending on which index we study.

We propose to study the central limit theorem for the estimator of the index Sob_{X_i} proposed in Appendix A.2. Note that the result is the same for other estimators of the Sobol' indices proposed in the same section.

If we denote

$$Z_n = \begin{pmatrix} n^{-1}\sum f(X_{i,k}, X_{\sim i,k}) f(X_{i,k}, X'_{\sim i,k}) \\ n^{-1}\sum f(X_{i,k}, X_{\sim i,k}) f(X'_{i,k}, X'_{\sim i,k}) \\ n^{-1}\sum f(X_{i,k}, X_{\sim i,k}) \\ n^{-1}\sum f^2(X_{i,k}, X_{\sim i,k}) \end{pmatrix} \tag{A.11}$$

then the estimator $\widehat{Sob_{X_i}}$ of the Sobol' index Sob_{X_i} is equal to $h(Z_n)$ where

$$h(\beta_1, \beta_2, \beta_3, \beta_4) = \frac{\beta_1 - \beta_2}{\beta_4 - \beta_3^2}.$$

Applying the delta-method [Van00], we obtain the convergence of $h(Z_n)$ to $h(Z) = Sob_{X_i}$

$$\sqrt{n}\left(\widehat{Sob_{X_i}} - Sob_{X_i}\right) \to \mathcal{N}(0, \nabla h(\beta) \Sigma \nabla h(\beta)^T), \tag{A.12}$$

for which we need to compute the gradient of h

$$\nabla h(\beta_1,\beta_2,\beta_3,\beta_4) = \left(\frac{1}{\beta_4-\beta_3^2}, -\frac{1}{\beta_4-\beta_3^2}, \frac{2(\beta_1-\beta_2)\beta_3}{(\beta_4-\beta_3^2)^2}, \frac{-(\beta_1-\beta_2)}{(\beta_4-\beta_3^2)^2}\right)^T$$

and the correlation matrix Σ for the variable Z_n which is

$$\Sigma = \begin{pmatrix} \sigma_{11}^2 & \sigma_{12}^2 & \sigma_{13}^2 & \sigma_{14}^2 \\ \sigma_{12}^2 & \sigma_{22}^2 & 0 & 0 \\ \sigma_{13}^2 & 0 & \sigma_{33}^2 & \sigma_{34}^2 \\ \sigma_{14}^2 & 0 & \sigma_{34}^2 & \sigma_{44}^2 \end{pmatrix} \tag{A.13}$$

where the values $\sigma_{ij}^2 = Cov(Z_i, Z_j)$ are given as

$$\begin{aligned}
\sigma_{11}^2 =& \mathrm{Var}(f(X,X_{\sim i})f(X,X'_{\sim i})) \\
\sigma_{12}^2 =& \mathbb{E}[f^2(X,X_{\sim i})f(X,X'_{\sim i})f(X',X'_{\sim i})] \\
\sigma_{13}^2 =& \mathbb{E}[f^2(X,X_{\sim i})f(X,X'_{\sim i})] \\
\sigma_{14}^2 =& \mathbb{E}[f^3(X,X_{\sim i})f(X,X'_{\sim i})f(X',X'_{\sim i})] - \mathbb{E}[f^2(X,X_{\sim i})]\mathbb{E}[f(X,X'_{\sim i})f(X,X'_{\sim i})] \\
\sigma_{22}^2 =& \mathrm{Var}(f(X,X_{\sim i}))^2 \\
\sigma_{33}^2 =& \mathrm{Var}(f(X,X_{\sim i})) \\
\sigma_{34}^2 =& \mathbb{E}[f^3(X,X_{\sim i})] \\
\sigma_{44}^2 =& \mathbb{E}[f^4(X,X_{\sim i}) - \mathbb{E}[f^2(X,X_{\sim i})]^2.
\end{aligned} \tag{A.14}$$

A.4 Estimation of Cramér-von-Mises indices

We propose two ways of estimating the extended Cramér-von-Mises indices that we denote by $U(Y,X_i|X_{\sim i})$ defined in (2.13).

The first one is to use the fact that

$$\begin{aligned}
U(Y,X_i|\mathbf{Z}) &= \frac{\int \mathbb{E}(\mathrm{Var}(\mathbb{E}\left[\mathbb{1}_{Y\le t}|X_i,\mathbf{Z}\right]|\mathbf{Z}))d\mu(t))}{\int \mathrm{Var}(\mathbb{1}_{Y\le t})d\mu(t))} \\
&= T(Y,X_i|\mathbf{Z}) \times (1 - T(Y,\mathbf{Z})).
\end{aligned} \tag{A.15}$$

We need to estimate $T(Y,X_i|X_{\sim i})$ and $T(Y,X_{\sim i})$. Estimates for both theses quantities are taken from [AC21].

Consider a triple of random variables (X,Z,Y) and an i.i.d sample $(X_i,Z_i,Y_i)_{1\le i\le n}$. For simplicity, we still suppose the random variables to be diffuse (that is without ties). The random variable Z is used for the conditioning.

For each i, let $N(i)$ be the index j such that Z_j is the nearest neighbor of Z_i with respect to the Euclidean distance and let $M(i)$ be the index j such that (X_j,Z_j) is the nearest neighbor of (X_i,Z_i). Let R_i be the rank of Y_i, that is the number of j such that $Y_j \le Y_i$.

The correlation coefficient defined in [AC21] is defined as:

$$T_n(Y, X|Z) = \frac{\sum_{i=1}^n \left(\min\{R_i, R_{M(i)}\} - \min\{R_i, R_{N(i)}\}\right)}{\sum_{i=1}^n \left(R_i - \min\{R_i, R_{N(i)}\}\right)}. \tag{A.16}$$

The authors of [AC21] prove that this estimator converges almost surely to a deterministic limit $T(Y, X|Z)$ which is equal to the quantity we defined in the first section. In order to estimate the extended Cramér-von-Mises sensitivity index CVM_X^{ind}, we propose the estimator

$$U_n(Y, X_i|X_{\sim i}) = T_n(Y, X_i|X_{\sim_i}) \times (1 - T_n(Y, X_{\sim i}))\,. \tag{A.17}$$

The convergence of the estimator $U_n(Y, X_i|X_{\sim i})$ to the quantity of interest $U(Y, X_i|X_{\sim i})$ is immediate.

We propose an alternative method for the estimation of this index. We take advantage of the estimates given in [AC21] and [Cha20]. We have the two following convergences almost surely:

$$Q_n(Y, X|Z) = n^{-2} \sum_{j=1}^n \left(\min\{R_j, R_{M(j)}\} - \min\{R_j, R_{N(j)}\}\right) \to \int \mathbb{E}(\mathrm{Var}(\mathbb{E}\left[\mathbb{1}_{Y\leq t}|X, Z\right]|Z))d\mu(t)) \tag{A.18}$$

$$S_n(Y) = n^{-3} \sum_{j=1}^n L_j(n - L_j) \to \int \mathrm{Var}(\mathbb{1}_{Y\leq t})d\mu(t)) \tag{A.19}$$

where L_j is the number of k such that $Y_k \geq Y_j$.

Proposition 6 (Estimator of the extended Cramér-von-Mises indices)**.** *The quantity defined as $\tilde{U}_n(Y, X|Z) = Q_n(Y, X|Z)/S_n(Y)$ is a consistent estimator of $U(Y, X_i|X_{\sim i})$.*

The proof is obtained directly using classical probability tools.

A.5 Proofs

A.5.1 Proof of Theorem A.1.1

Proof. Indeed, we can always write

$$p_{\mathbf{X}} = p_{X_i} \times p_{X_{i+1}|X_i} \times \cdots \times p_{X_{i-1}|X_{\sim(i-1)}}. \tag{A.20}$$

Since we are back to a product of marginals, we have a hierarchical independence. We choose the cyclical hierarchy (X_i, followed by $X_{i+1}|X_i$, then $X_{i+2}|X_i, X_{i+1}$, and so on and so forth till $X_{i-1}|X_{\sim(i-1)}$) as we are in fact only interested in the first and the last elements of this hierarchy (X_i and $X_{i-1}|X_{\sim(i-1)}$). We can always map univariate random variables to uniform distributions by matching the quantiles by using the cumulative distribution function – one can view this operation as hierarchical Optimal Transport, see [CGS10] – and by doing so for

each variable defined above, we have the so-called Levy-Rosenblatt transform, denoted here as RT, that is:

$$(X_i, (X_{i+1}|X_i), \cdots, (X_{i-1}|X_{\sim(i-1)})) \sim p_{\mathbf{X}} \xrightarrow{RT} (U_1^i, \cdots, U_p^i) \sim \mathcal{U}^p(0,1). \tag{A.21}$$

□

A.5.2 Proof of Examples following 2.3.1

Proof. We will show here how each definition of fairness and GSA measure presented in Table 2.2 match for binary classification with S binary.

(i) The definition of *Statistical Parity* is given by $|\mathbb{P}(f(\mathbf{X}) = 1|S = 1) - \mathbb{P}(f(\mathbf{X}) = 1|S = 0)|$. For simplicity, we consider $\mathrm{Var}(f(\mathbf{X})) = 1$. If we compute the Sobol' index of the predictor $f(\mathbf{X})$ for the protected variable S, we obtain:

$$\begin{aligned} Sob_S(f(\mathbf{X})) &= \mathrm{Var}_S(\mathbb{E}_{\mathbf{X}\setminus S}[f(\mathbf{X})|S]) \\ &= \mathbb{E}_S\mathbb{E}^2_{\mathbf{X}\setminus S}[f(\mathbf{X})|S] - \mathbb{E}_{\mathbf{X}}[f(\mathbf{X})|S]^2 \\ &= \mathbb{P}(S=1)\mathbb{P}(f(\mathbf{X})=1|S=1)^2 + \mathbb{P}(S=0)\mathbb{P}(f(\mathbf{X})=1|S=0)^2 - \mathbb{P}(f(\mathbf{X})=1)^2 \\ &= \mathbb{P}(S=1)\mathbb{P}(S=0) \times [\mathbb{P}(f(\mathbf{X})=1|S=1) - \mathbb{P}(f(\mathbf{X})=1|S=0)]^2 \\ &= \mathbb{P}(S=1)\mathbb{P}(S=0) \times DI^2. \end{aligned}$$

We see that the quantity of interest in *Statistical Parity* is the same as the Sobol' index, up to a constant depending on the proportion in each class of the protected variable.

(ii) For *avoiding Disparate mistreatment*, the quantity of interest is $|\mathbb{P}(f(\mathbf{X}) \neq Y|S = 1) - \mathbb{P}(f(\mathbf{X}) \neq Y|S = 0)|$. This can be obtained by replacing $f(\mathbf{X})$ by $\mathbb{1}_{f(\mathbf{X})\neq Y}$ in the quantity of interest for *Statistical Parity*. Therefore, by the same computation as previously, we can link *avoiding Disparate mistreatment* to the Sobol' index of the error of the predictor $\mathbb{1}_{f(\mathbf{X})\neq Y}$ for the protected variable S.

(iii) For *Equality of Odds*, we are interest in the difference $|\mathbb{P}(f(\mathbf{X})|Y = i, S = 1) - \mathbb{P}(f(\mathbf{X})|Y = i, S = 0)|$ for $i = 0, 1$. Each of this difference can be expressed as seen before as $\mathrm{Var}_S(E_X[f(\mathbf{X})|Y = i, S])$. Since we want this quantity to be equal to zero for each i, we can compute *Equality of Odds* with $\mathbb{E}_Y\mathrm{Var}_S(E_X[f(\mathbf{X})|Y, S])$, which is the extended Cramèr-von-Mises index of the predictor for the protected variable S.

(iv) For *avoiding Disparate Treatment*, the quantity of interest is very similar to *Statistical Parity* since we are interested in proving $f(\mathbf{X})|\mathbf{X} \setminus S \perp\!\!\!\perp S$. By similar computations as before, this fairness boils back to looking at $\mathbb{E}_{\mathbf{X}\setminus S}\mathrm{Var}\mathbb{E}_{\mathbf{X}\setminus S}[f(\mathbf{X})|\mathbf{X}]$. This can be simplified into $\mathbb{E}_{\mathbf{X}\setminus S}\mathrm{Var}[f(\mathbf{X})|\mathbf{X} \setminus S]$, which is the Total Sobol' index of the predictor for the protected variable S.

□

A.5.3 Proof of Proposition 1

Proof. The proof is a direct consequence of the Hoeffding decomposition of the function $Y = \psi(X, S)$. By factorizing $\mathbb{P}_Y$ as $\mathbb{P}_{Y|X,S}\mathbb{P}_{X|S}\mathbb{P}_S$, we can write

$$Y = \psi_X(X(S)) + \psi_S(S) + \psi_{S,X}(S) \times \psi_{X,S}(X(S))$$

If $SobT_S^{ind} = 0$ then $\text{Var}(\psi_S(S) + \psi_{S,X}(S) \times \psi_{X,S}(X(S))) = 0$. By orthogonality in the Hoeffding decomposition, $\text{Var}(\psi_S(S)) = \text{Var}(\psi_{S,X}(S) \times \psi_{X,S}(X(S))) = 0$, which lead to $\psi_S(S) = \psi_{S,X}(S) \times \psi_{X,S}(X(S)) = 0$. It holds that $Y = \psi_X(X(S))$.

For the second part of the proposition, we apply the same reasoning by factorizing $\mathbb{P}_Y$ as $\mathbb{P}_{Y|X,S}\mathbb{P}_{S|X}\mathbb{P}_X$. We can write

$$Y = \psi'_S(S(X)) + \psi'_X(X) + \psi'_{S,X}(X) \times \psi'_{X,S}(S(X))$$

If $SobT_S = 0$ then $\text{Var}(\psi'_S(S(X)) + \psi'_{S,X}(X) \times \psi'_{X,S}(S(X))) = 0$. By orthogonality in the Hoeffding decomposition, $\text{Var}(\psi'_S(S(X))) = \text{Var}(\psi'_{S,X}(X) \times \psi'_{X,S}(S(X)) = 0$, which lead to $\psi'_S(S(X)) = \psi'_{S,X}(X) \times \psi'_{X,S}(S(X)) = 0$. It holds that $Y = \psi'_X(X)$.

□

A.5.4 Proof of Proposition 2 and Proposition 3

Proof. Without loss of generality, we can consider only two sensitive features S_1 and S_2. Because of the various bounds on Sobol' indices explained in previous Section, we know that $SobT_{S_1,S_2} \leq SobT_{S_1}$. $SobT_{S_1}$ is the GSA measure associated with *Avoiding Disparate Treatment*. This means that to be fair in the sense of *Avoiding Disparate Treatment* implies the nullity of $SobT_{S_1}$ and therefore the nullity of $SobT_{S_1,S_2}$. The second result is a direct consequence of the absence of bounds between Sob_{S_1} and Sobol' indices for (S_1, S_2) and an example has been given in the previous toy-case in introduction of the Subsection. We can find cases where Sob_{S_1} is arbitrary high and Sob_{S_1,S_2} is null, such as $f(X) = S_1$; and cases where Sob_{S_1} is null and Sob_{S_1,S_2} is arbitrary high, such as $f(X) = S_1 \times S_2$. □

A.5.5 Proof of the Central Limit Theorem for the Chatterjee estimator of Sobol' on multivariate outputs

For convenience, we denote $Y = f(\mathbf{X}) = (f_1(\mathbf{X}), \cdots, f_l(\mathbf{X}))^T$ as $f(X, W)$, with $X = X_1$ and $W = X_{\sim 1}$. We assume that f is bounded, along with its first two derivatives. In order to prove a Central Limit Theorem for the estimator $\bar{\xi}_n(X, Y)$, given in 4.3, we follow the univariate proof given in [Gam+20], albeit with two modifications. Firstly, we assume that $\mathbf{f}$ is valued in a compact set of $\mathbb{R}^d$. Secondly, instead of assuming the random variable X to be uniform between 0 and 1, we only assume that its density is below-bounded and continuous on an interval $[a, b]$. We also denote by $q_{j,n}$ the $j/(n+1)$-th quantile of X, so that $q_{j,n} = F_X^{-1}(\frac{j}{n+1})$.

The random variables X and W are defined on a product space $\Omega = \Omega_X \times \Omega_W$; so that for any $\omega \in \Omega$, there exists $\omega_X \in \Omega_X$ and $\omega_W \in \Omega_W$ and we have $(X, W)(\omega) = (X(\omega_X), W(\omega_W))$. Further, we consider π_W the projection on Ω_W and the product measure $\mathbb{P} = \mathbb{P}_X \otimes \mathbb{P}_W = \mathcal{L}_X \otimes \mathcal{L}_W$, where $\mathcal{L}_X$ is the distribution of X and $\mathcal{L}_W$ is the distribution of W. Naturally, $\mathbb{P}_W = \mathbb{P} \circ \pi_W^{-1}$.

We aim to prove a CLT for the estimator $\bar{xi}_n(\mathbf{f}, (X, W))$ of the classical first-order Sobol' index with respect to X given by (4.2), the estimator of which defined in (4.3) is given by

$$\bar{\xi_n}(\mathbf{f}, \mathbf{X}^n) = \frac{\sum_{k=1}^{l} \left(n^{-1} \sum_{j=1}^{n} f_k(\mathbf{X}^{\sigma_n(j)}) f_k(\mathbf{X}^{(\sigma_n(j+1))}) - \left(n^{-1} \sum_{j=1}^{n} f_k(\mathbf{X}^{\sigma_n(j)}) \right)^2 \right)}{\sum_{k=1}^{l} \left(n^{-1} \sum_{j=1}^{n} f_k(\mathbf{X}^{\sigma_n(j)})^2 - \left(n^{-1} \sum_{j=1}^{n} f_k(\mathbf{X}^{\sigma_n(j)}) \right)^2 \right)}.$$

where σ_n is defined at the beginning of the section4.1.

Proof of Theorem 4.1.3

The proof will proceed as follows. First, we prove a CLT for

$$\begin{pmatrix} \frac{1}{n} \sum_{j=1}^{n-1} f_1(X_{\sigma_n(j)}, W_{\sigma_n(j)}) f_1(X_{\sigma_n(j+1)}, W_{\sigma_n(j+1)}) \\ \frac{1}{n} \sum_{j=1}^{n} f_1(X_{\sigma_n(j)}, W_{\sigma_n(j)}) \\ \frac{1}{n} \sum_{j=1}^{n} f_1(X_{\sigma_n(j)}, W_{\sigma_n(j)})^2 \\ \vdots \\ \frac{1}{n} \sum_{j=1}^{n-1} f_l(X_{\sigma_n(j)}, W_{\sigma_n(j)}) f_l(X_{\sigma_n(j+1)}, W_{\sigma_n(j+1)}) \\ \frac{1}{n} \sum_{j=1}^{n} f_l(X_{\sigma_n(j)}, W_{\sigma_n(j)}) \\ \frac{1}{n} \sum_{j=1}^{n} f_l(X_{\sigma_n(j)}, W_{\sigma_n(j)})^2 \end{pmatrix}^T$$

that amounts to prove a CLT for

$$\begin{pmatrix} \frac{1}{n}\sum_{j=1}^{n-1} f_1(X_{\sigma_n(j)}, W_{\sigma_n(j)}) f_1(X_{\sigma_n(j+1)}, W_{\sigma_n(j+1)}) \\ \frac{1}{n}\sum_{j=1}^{n-1} f_1(X_{\sigma_n(j)}, W_{\sigma_n(j)}) \\ \frac{1}{n}\sum_{j=1}^{n-1} f_1(X_{\sigma_n(j)}, W_{\sigma_n(j)})^2 \\ \vdots \\ \frac{1}{n}\sum_{j=1}^{n-1} f_l(X_{\sigma_n(j)}, W_{\sigma_n(j)}) f_l(X_{\sigma_n(j+1)}, W_{\sigma_n(j+1)}) \\ \frac{1}{n}\sum_{j=1}^{n-1} f_l(X_{\sigma_n(j)}, W_{\sigma_n(j)}) \\ \frac{1}{n}\sum_{j=1}^{n-1} f_l(X_{\sigma_n(j)}, W_{\sigma_n(j)})^2 \end{pmatrix}^T$$

since f is bounded. Secondly, we use the so-called delta method [Van00, Theorem 3.1] to conclude to Theorem 4.1.3.

It is worth noticing that the permutation on the W's do not affect the result as seen in the sequel. For $j = 1, \dots n-1, k \in [\![1, l]\!]$ and denoting by $q_{j,n}$ the $j/(n+1)$-th quantile of X, the introduction of

$$\Delta_{n,j,k} = f_k\left(X_{\sigma_n(j)}, W_j\right) - f_k\left(q_{j,n}, W_j\right), \quad W_{n,j} = \left(q_{j,n}, W_j\right) \tag{A.22}$$

leads to $f_k\left(X_{\sigma_n(j)}, W_{\sigma_n(j)}\right) \overset{\mathcal{L}}{=} f\left(X_{\sigma_n(j)}, W_j\right) = \left(\Delta_{n,j,k} + f_k\left(W_{n,j}\right)\right)$ and

$$\begin{aligned} f_k\left(X_{\sigma_n(j)}, W_{\sigma_n(j)}\right) f_k\left(X_{\sigma_n(j+1)}, W_{\sigma_n(j+1)}\right) & \\ &\overset{\mathcal{L}}{=} f_k\left(X_{\sigma_n(j)}, W_j\right) f_k\left(X_{\sigma_n(j+1)}, W_{j+1}\right) \\ &= \Big(f_k\left(W_{n,j}\right) + \Delta_{n,j,k}\Big)\Big(f_k\left(W_{n,j+1}\right) + \Delta_{n,j+1,k}\Big) \\ &= f_k\left(W_{n,j}\right) f_k\left(W_{n,j+1}\right) + \Delta_{n,j,k} f_k\left(W_{n,j+1}\right) \\ &\quad + \Delta_{n,j+1,k} f_k\left(W_{n,j}\right) + \Delta_{n,j,k}\Delta_{n,j+1,k}. \end{aligned}$$

Thus we are led to establish a CLT for

$$Z_n = \frac{1}{n}\sum_{j=1}^{n-1} \begin{pmatrix} f_1(W_{n,j}) f_1(W_{n,j+1}) + \Delta_{n,j,1} f_1\left(W_{n,j+1}\right) + \Delta_{n,j+1,1} f_1\left(W_{n,j}\right) + \Delta_{n,j,1}\Delta_{n,j+1,1} \\ f_1(W_{n,j}) + \Delta_{n,j,1} \\ \left(f_1(W_{n,j}) + \Delta_{n,j,1}\right)^2 \\ \vdots \\ f_l(W_{n,j}) f_l(W_{n,j+1}) + \Delta_{n,j,l} f_l\left(W_{n,j+1}\right) + \Delta_{n,j+1,l} f_l\left(W_{n,j}\right) + \Delta_{n,j,l}\Delta_{n,j+1,l} \\ f_l(W_{n,j}) + \Delta_{n,j,l} \\ \left(f_l(W_{n,j}) + \Delta_{n,j,l}\right)^2 \end{pmatrix}. \tag{A.23}$$

Let us discard the negligible terms in the CLT for Z_n. For this we need to establish that

$$X_{\sigma_n(j)} - q_{j,n} = O_{\mathbb{P}}\left(\frac{1}{\sqrt{n}}\right). \tag{A.24}$$

To obtain this result, we first prove it for $U \sim \mathcal{U}([0,1])$. Classical results for order statistics allow us to assert that, if we denote $U_{(j)}$ for the j-th element in the sorted sequence $(U_1, \ldots, U_n)$, then $U_{(j)} \sim \text{Beta}(j, n+1-j)$. Therefore, we have that

$$U_{(j)} - \frac{j}{n+1} = O_{\mathbb{P}}\left(\frac{1}{\sqrt{n}}\right). \tag{A.25}$$

By using a Taylor expansion of the quantile function of X around the point $j/(n+1)$, we obtain

$$X_{\sigma_n(j)} - F_X^{-1}\left(\frac{j}{n+1}\right) = \left(U_{(j)} - \frac{j}{n+1}\right) \times \varphi_X(F_X^{-1}(D_j))^{-1}, \tag{A.26}$$

where D_j is a random variable belonging to the segment $[U_{(j)}, j/(n+1)]$ and φ_X is the density of the distribution of X. Since, by assumption, the density of X is below-bounded, the asymptotic comportment of $X_{\sigma_n(j)} - q_{j,n}$ is the same one as $U_{(j)} - j/(n+1)$, hence the result.

As explained below, (A.24) will imply that for every $k \in [\![1, l]\!]$

$$\frac{1}{n}\sum_{j=1}^{n-1} \Delta_{n,j,k}^2 = O_{\mathbb{P}}\left(\frac{1}{n}\right) \quad \text{and} \quad \frac{1}{n}\sum_{j=1}^{n-1} \Delta_{n,j,k}\Delta_{n,j+1,k} = O_{\mathbb{P}}\left(\frac{1}{n}\right). \tag{A.27}$$

First of all, we expand $\Delta_{n,j,k}$ (resp. $\Delta_{n,j+1,k}$) using the Taylor-Lagrange formula, for any $j = 1, \ldots n-1$ and $k = 1, \ldots, l$ and we obtain

$$\Delta_{n,j,k} = \left(X_{\sigma_n(j)} - q_{j,n}\right) f_{k,x}\left(W_{n,j}\right) + \frac{1}{2}\left(X_{\sigma_n(j)} - q_{j,n}\right)^2 f_{k,xx}\left(\delta_{n,j,k}, W_{\sigma_n(j)}\right), \tag{A.28}$$

where $\delta_{n,j,k}$ (resp. $\delta_{n,j+1,k}$) lies in the unordered segment $(X_{\sigma_n(j)}, q_{j,n})$ (resp. $(X_{\sigma_n(j+1)}, q_{j+1,n})$) and where $f_{k,x}$ and $f_{k,xx}$ are the first and second derivatives of f_k with respect to the first coordinate. This leads to expansions for $\Delta_{n,j,k}^2$ and $\Delta_{n,j,k}\Delta_{n,j+1,k}$:

$$\begin{aligned}
\Delta_{n,j,k}^2 &= \left(X_{\sigma_n(j)} - q_{j,n}\right)^2 \left(f_{k,x}\left(W_{n,j}\right) + \frac{1}{2}\left(X_{\sigma_n(j)} - q_{j,n}\right) f_{k,xx}\left(\delta_{n,j,k}, W_{\sigma_n(j)}\right)\right)^2 \\
\Delta_{n,j,k}\Delta_{n,j+1,k} &= \left(X_{\sigma_n(j)} - q_{j,n}\right)\left(X_{\sigma_n(j+1)} - q_{j+1,n}\right) \\
&\quad \times \left(f_{k,x}\left(W_{n,j}\right) + \frac{1}{2}\left(X_{\sigma_n(j)} - q_{j,n}\right) f_{k,xx}\left(\delta_{n,j,k}, W_{\sigma_n(j)}\right)\right) \\
&\quad \times \left(f_{k,x}\left(W_{n,j+1}\right) + \frac{1}{2}\left(X_{\sigma_n(j+1)} - q_{j+1,n}\right) f_{k,xx}\left(\delta_{n,j+1,k}, W_{\sigma_n(j+1)}\right)\right).
\end{aligned}$$

Finally, using the boundedness of f_k, $f_{k,x}$, and $f_{k,xx}$, together with (A.24), (A.27) follows.

Remark that the proof of (A.27) yields also

$$\frac{1}{n}\sum_{j=1}^{n-1} \Delta_{n,j,k} = O_{\mathbb{P}}\left(\frac{1}{\sqrt{n}}\right), \tag{A.29}$$

from which it is clear that this term will contribute in the CLT on Z_n. Then (A.27) entails that the asymptotic study reduces to that of the empirical mean of $Z_{n,j} = B_{n,j} + C_{n,j}$ where

$$B_{n,j} = \begin{pmatrix} f_1(W_{n,j})\, f_1(W_{n,j+1}) \\ f_1(W_{n,j}) \\ f_1(W_{n,j})^2 \\ \vdots \\ f_l(W_{n,j})\, f_l(W_{n,j+1}) \\ f_l(W_{n,j}) \\ f_l(W_{n,j})^2 \end{pmatrix} \text{ and } C_{n,j} = \begin{pmatrix} \Delta_{n,j,1} f_1(W_{n,j+1}) + \Delta_{n,j+1,1} f_1(W_{n,j}) \\ \Delta_{n,j,1} \\ 2\Delta_{n,j,1} f_1(W_{n,j}) \\ \vdots \\ \Delta_{n,j,l} f_l(W_{n,j+1}) + \Delta_{n,j+1,l} f_l(W_{n,j}) \\ \Delta_{n,j,l} \\ 2\Delta_{n,j,l} f_l(W_{n,j}) \end{pmatrix}. \tag{A.30}$$

First, we consider $B_{n,j}$ in (A.30) and we establish the following result, the proof of which has been postponed to Appendix A.5.5.

Lemma 1. *As $n \to \infty$, the random vector B_n given by $\frac{1}{n}\sum_{j=1}^{n-1} B_{n,j}$ satisfies a CLT. More precisely, $\sqrt{n}\big(B_n - m_B\big) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}_{3l}(0, \Sigma_B)$, where*

$$m_B = \big(\mathbb{E}[Y_1 Y_1'], \mathbb{E}[Y_1], \mathbb{E}[Y_1^2], \cdots, \mathbb{E}[Y_l Y_l'], \mathbb{E}[Y_l], \mathbb{E}[Y_l^2]\big)^\top, \tag{A.31}$$

and for each k, $Y_k' = f_k(X, W')$, W' is an independent copy of W, and Σ_B has an explicit expression given in Appendix A.5.5.

Remark that Y_k' is the so-called Pick-Freeze version of Y_k with respect to X. Secondly, we establish a conditional CLT for the empirical mean of the $C_{n,j}$'s defined in (A.30). The reader is referred to Appendix A.5.5 for the proof of this result.

Lemma 2. *There exists a measurable set $\Pi \in \Omega_W$ having $\mathbb{P}_W$-probability one such that, for any $\omega_W \in \Pi$, we have*

$$\sqrt{n} C_n(\cdot, \omega_W) \xrightarrow[n\to\infty]{\mathcal{L}_X} \mathcal{N}_{3l}(0, \Sigma_C).$$

Moreover, Σ_C does not depend on ω_W and has an explicit expression given Appendix A.5.5.

Considering the characteristic function of the vector $\sqrt{n}(B_n - \mathbb{E}[B_n], C_n)$, one may write

$$\mathbb{E}\left[e^{i(\sqrt{n}\langle s, (B_n - \mathbb{E}[B_n])\rangle + \sqrt{n}\langle t, C_n\rangle)}\right] = \mathbb{E}\left[e^{i\sqrt{n}\langle s, (B_n - \mathbb{E}[B_n])\rangle} \mathbb{E}\left[e^{i\sqrt{n}\langle t, C_n\rangle} \big| \mathcal{F}_W\right]\right]$$

for any s and $t \in \mathbb{R}^{3l}$. On the one hand, $\mathbb{E}\left[e^{i\sqrt{n}\langle t, C_n\rangle} \big| \mathcal{F}_W\right]$ converges a.s. to $\exp\{-t^\top \Sigma_C t/2\}$ which is not random. On the other hand, $\sqrt{n}\langle s, (B_n - \mathbb{E}[B_n])\rangle$ converges in distribution to a Gaussian random variable denoted by B_s. By Slutsky's lemma,

$$\left(\sqrt{n}\langle s, (B_n - \mathbb{E}[B_n])\rangle, \mathbb{E}\left[e^{i\sqrt{n}\langle t, C_n\rangle} \big| \mathcal{F}_W\right]\right)$$

converges in distribution to $(B_s, \exp\{-t^\top \Sigma_C t/2\})$. We consider the application $h\colon (u,v) \in \mathbb{R} \times D(0,1) \mapsto e^{iu}v \in \mathbb{C}$ where $D(0,1)$ is the unit disc in $\mathbb{C}$. The continuity and the boundedness of h lead to the convergence in distribution of $e^{i\sqrt{n}\langle s,(B_n-\mathbb{E}[B_n])\rangle}\left[e^{i\sqrt{n}\langle t,C_n\rangle}\middle|\mathcal{F}_W\right]$ and we conclude to the asymptotic normality of $\sqrt{n}(B_n - \mathbb{E}[B_n], C_n)$ to a Gaussian random vector of dimension $6l$ with zero mean and variance-covariance matrix $\begin{pmatrix}\Sigma_B & 0\\ 0 & \Sigma_C\end{pmatrix}$. It remains to apply the so-called delta method [Van00, Theorem 3.1] and Slutsky's lemma to get the required result. The details of the computation of the asymptotic variance σ^2 can be found in Appendix A.5.5.

Proof of Lemma 1

The proof of Lemma 1 is the same as its equivalent in [Gam+20], with minimal changes for the asymptotic covariance matrix. Indeed, we obtain a $3l \times 3l$ covariance matrix $(\Sigma_B)_{i,j}$ as follows.

- If $i = 3(k-1)+1, j = 3(k'-1)+1, (k,k') \in [\![1,l]\!]^2$, then we obtain

$$(\Sigma_B)_{i,j} = \mathbb{E}[\mathrm{Cov}(f_k(\mathbf{X})f_k(X,W'); f_{k'}(\mathbf{X})f_{k'}(X,W')|X)] + 2\mathbb{E}[\mathrm{Cov}(f_k(\mathbf{X})f_k(X,W'); f_{k'}(X,W')f_{k'}(X,W'')|X)]$$

  where with W' and W'' are two independent copies of W.

- If $i = 3(k-1)+2, j = 3(k'-1)+2, (k,k') \in [\![1,l]\!]^2$, then we obtain

$$(\Sigma_B)_{i,j} = \mathbb{E}[\mathrm{Cov}(f_k(\mathbf{X}); f_{k'}(\mathbf{X})|X)].$$

- If $i = 3(k-1)+3, j = 3(k'-1)+3, (k,k') \in [\![1,l]\!]^2$, then we obtain

$$(\Sigma_B)_{i,j} = \mathbb{E}[\mathrm{Cov}(f_k(\mathbf{X})^2; f_{k'}(\mathbf{X})^2|X)].$$

- If $i = 3(k-1)+1, j = 3(k'-1)+2, (k,k') \in [\![1,l]\!]^2$, then we obtain

$$(\Sigma_B)_{i,j} = 2\mathbb{E}[\mathrm{Cov}(f_k(\mathbf{X})f_k(X,W'); f_{k'}(\mathbf{X})|X)].$$

- If $i = 3(k-1)+1, j = 3(k'-1)+3, (k,k') \in [\![1,l]\!]^2$, then we obtain

$$(\Sigma_B)_{i,j} = 2\mathbb{E}[\mathrm{Cov}(f_k(\mathbf{X})f_k(X,W'); f_{k'}(\mathbf{X})^2|X)].$$

- If $i = 3(k-1)+2, j = 3(k'-1)+3, (k,k') \in [\![1,l]\!]^2$, then we obtain

$$(\Sigma_B)_{i,j} = 2\mathbb{E}[\mathrm{Cov}(f_k(\mathbf{X}); f_{k'}(\mathbf{X})^2|X)].$$

Proof of Lemma 2

In order to prove the Central Limit Theorem of the random variable C_n, we need the following lemmas.

Lemma 3. *Let k and $l \in [\![0, n]\!]$. There exists a measurable set $\Pi \subset \Omega_W$ with $\mathbb{P}_W$-probability one such that for any $\omega_W \in \Pi$,*

$$\pi_n(\omega_W) = \frac{1}{n} \sum_{j=1}^{n-k\wedge l} \delta_{\left(\frac{j}{n+1}, \frac{j+1}{n+1}, \cdots, \frac{j+l}{n+1}, W_j(\omega_W), \cdots, W_{j+l}(\omega_W)\right)} \Rightarrow \pi = \mathcal{L}_{(X,\cdots,X)} \otimes \mathcal{L}_W \otimes \cdots \otimes \mathcal{L}_W,$$

as $n \to \infty$ where X has a continuous distribution on the interval $[a, b]$ with strictly positive density and $\Rightarrow$ stands for the weak convergence of measures. Here, $\mathcal{L}_{(X,\cdots,X)}$ stands for the joint distribution of the random vector $(X, \cdots, X)$ of size k and $\mathcal{L}_W \otimes \cdots \otimes \mathcal{L}_W$ for the distribution $\mathcal{L}_W$ tensorized l times.

The proof of this lemma is the same as in the supplementary material of [Gam+20].

Lemma 4. *Let $(E_i)_{i\geqslant 1}$ be a sequence of i.i.d. random variables with standard exponential distribution and let ψ be a bounded measurable function on $[a, b]$. We denote by q_j the $j/(n + 1)$-th quantile of X, that is $q_j = F_X^{-1}(j/(n + 1))$. We assume that the set of discontinuity points of ψ has null Lebesgue measure. Then, the sequence*

$$\left(n^{-1/2} \sum_{j=1}^{n-1} \psi(q_j)(E_j - 1) \right)_{n\in\mathbb{N}^*}$$

converges in distribution to a centered Gaussian law with asymptotic variance: $\sigma_\psi^2 = \int_{[a,b]} \psi^2(x) dF_X(x)$.

The proof of this lemma is done by straightforward computation of the cumulant of the exponential random variable E_i. The detailed proof can be found in [Gam+20] as well, albeit with X supposed to be uniform in $[0, 1]$

Remark 23. *The previous lemma extends to the case of a continuous function $\Psi = (\psi_i)$ valued in $\mathbb{R}^d$ ($d \geqslant 1$). In this case, the asymptotic covariance matrix Σ_Ψ is the Gram matrix $\left(\int_{[a,b]} \psi_i(x)\psi_j(x) dF_X(x);\ 1 \leqslant i, j \leqslant d \right)$. Indeed, the previous lemma holds for any linear combination of such random vector sequence. A direct computation of the asymptotic variance leads to the quadratic form built on Σ_Ψ.*

Finally, the lemma of interest is the following.

Lemma 5. *Let $(U, \mathbb{B}(U))$ be a Polish space where $\mathbb{B}(U)$ denotes the Borel σ algebra of U. We consider a sequence $(\chi_j)_{1\leqslant j\leqslant n,\ n\in\mathbb{N}^*}$ valued in U and Q a probability measure on $[a, b] \times U$. We assume that the sequence of empirical measures $\left(\frac{1}{n} \sum_{j=1}^{n-1} \delta_{q_j,\chi_j} \right)_{n\in\mathbb{N}^*}$ converges in distribution to Q.*

Let ψ be a bounded measurable real function on $[a,b]\times U$. We assume that the set of discontinuity points of ψ has null Q-probability. Then,

$$M_n = \frac{1}{\sqrt{n}}\sum_{j=1}^{n-1}\psi\left(q_j,\chi_j\right)\left(X_{\sigma_n(j)}-q_j\right)\xrightarrow[n\to\infty]{\mathcal{L}}\mathcal{N}\left(0,s_\psi^2\right),$$

where the asymptotic variance s_ψ^2 is given in Equation (A.33).

Proof of Lemma 5. Recall that the sequence (E_i) has been defined in Lemma 4. We have, using the mean value theorem for the first equality

$$\begin{aligned}
X_{\sigma_n(j)}-q_j &\stackrel{\mathcal{L}}{=} \left(U_{\sigma_n(j)}-\frac{j}{n+1}\right)\times\frac{1}{f_X(F_X^{-1}(D_j))}\\
&= \left(\frac{\sum_{i=1}^{j}E_i}{\sum_{i=1}^{n+1}E_i}-\frac{j}{n+1}\right)\times\frac{1}{f_X(F_X^{-1}(D_j))} = \frac{1}{f_X(F_X^{-1}(D_j))}\frac{1}{\frac{1}{n+1}\sum_{i=1}^{n+1}E_i}\left(\frac{1}{n+1}\sum_{i=1}^{j}E_i-\frac{j}{(n+1)^2}\sum_{i=1}^{n+1}E_i\right)\\
&= \frac{1}{f_X(F_X^{-1}(D_j))}\frac{1}{\frac{1}{n+1}\sum_{i=1}^{n+1}E_i}\left(\frac{1}{n+1}\sum_{i=1}^{j}(E_i-1)-\frac{j}{(n+1)^2}\sum_{i=1}^{n+1}(E_i-1)\right),
\end{aligned}$$

with D_j a random variable in the interval $[U_{\sigma_n(j)},\frac{j}{n+1}]$,so that

$$M_n \stackrel{\mathcal{L}}{=} \frac{1}{\sqrt{n}(n+1)}\frac{1}{\frac{1}{n+1}\sum_{i=1}^{n+1}E_i}\sum_{j=1}^{n-1}\psi\left(q_j\chi_j\right)\frac{1}{f_X(F_X^{-1}(D_j))}\left(\sum_{i=1}^{j}(E_i-1)-\frac{j}{n+1}\sum_{i=1}^{n+1}(E_i-1)\right).$$

Using the assumption on the empirical measure, we get

$$\begin{aligned}
\frac{1}{n}\sum_{j=1}^{n}\psi\left(q_j,\chi_j\right)\frac{1}{f_X(F_X^{-1}(D_j))}\frac{j}{n+1} \to I = &\int_{U\times[0,1]} u\psi(F_X^{-1}(u),\chi)\frac{1}{f_X(F_X^{-1}(u))}dQ(F^{-1}(u),\chi)\\
&\int_{U\times[a,b]} x\psi(x,\chi)F_X(x)dQ(x,\chi)
\end{aligned}$$

By the weak law of large numbers, the quantity $(1/(n+1))\sum_{i=1}^{n+1}E_i$ converges in probability to $\mathbb{E}[E_1]=1$. Hence, by Slutsky's lemma, we are led to consider the random vector

$$\begin{aligned}
V_n &= \frac{1}{\sqrt{n}}\begin{pmatrix}\frac{1}{n+1}\sum_{j=1}^{n-1}\psi\left(q_j,\chi_j\right)\frac{1}{f_X(F_X^{-1}(D_j))}\sum_{i=1}^{j}(E_i-1)\\ \sum_{i=1}^{n+1}(E_i-1)\end{pmatrix}\\
&= \frac{1}{\sqrt{n}}\begin{pmatrix}\sum_{i=1}^{n-1}\left(\frac{1}{n+1}\sum_{j=1}^{n-1}\psi\left(q_j,\chi_j\right)\frac{1}{f_X(F_X^{-1}(D_j))}\mathbb{1}_{i\leqslant j}\right)(E_i-1)\\ \sum_{i=1}^{n+1}(E_i-1)\end{pmatrix}
\end{aligned}$$

Notice that the first coordinate of V_n can be rewritten as (up to the normalizing factor $n^{-1/2}$)

$$\sum_{i=1}^{n-1}\left(\frac{1}{n+1}\sum_{j=1}^{n-1}\psi\left(\chi_j,j/n\right)\perp\!\!\!\perp_{i\leqslant j}\right)(E_i-1).$$

For $t \in [0,1]$, let $\phi(t) = \int_{U\times[t,b]} \psi(x,\chi) dQ(x,\chi)$. We can show that

$$\lim_n \sup_{t\in[a,b]} \left| \left(\frac{1}{n+1} \sum_{j=1}^{n-1} \psi\left(q_j \chi_j\right) \frac{1}{f_X(F_X^{-1}(D_j))} \perp\!\!\!\perp_{i\leqslant j} \right) - \phi(t) \right| = 0, \qquad \text{(A.32)}$$

using Dini's theorem to the positive and negative part of $\psi(q_j,\chi)$ Then, in our study, we may replace V_n by

$$\widehat{V}_n = \frac{1}{\sqrt{n}} \begin{pmatrix} \frac{1}{n+1}\sum_{i=1}^{n-1} \phi(q_i)(E_i - 1) \\ \sum_{i=1}^{n+1}(E_i - 1) \end{pmatrix}$$

since (A.32) implies that $\lim_{n\to\infty} \mathbb{E}\|V_n - \widehat{V}_n\|^2 = 0$. Using Remark 23, we obtain that the sequence $(\widehat{V}_n)_{n\in\mathbb{N}^*}$ converges in distribution to a centered Gaussian vector with covariance matrix

$$\begin{pmatrix} \int_a^b \phi^2(t) dF_X(t) & \int_a^b \phi(t) dF_X(t) \\ \int_a^b \phi^2(t) dF_X(t) dt & 1 \end{pmatrix}.$$

Finally, using the so-called delta method [Van00, Theorem 3.1], $(M_n)_{n\in\mathbb{N}^*}$ converges in distribution to a centered Gaussian variable with variance

$$s_\psi^2 = \int_a^b (\phi(t) - I)^2 dF_X(t). \qquad \text{(A.33)}$$

Note that this variance can be rewritten as

$$s_\psi^2 = \iint_{([a,b]\times U)^2} \psi(x_1,\chi_1)\psi(x_2,\chi_2)(F_X(x_1)\wedge F_X(x_2) - F_X(x_1)F_X(x_2)) dQ(x_1,\chi_1) dQ(x_2,\chi_2). \qquad \text{(A.34)}$$

□

We can now prove the Lemma 2. Let $\omega_W \in \Pi$ as defined above. The aim is to establish a CLT for $\sqrt{n}C_{n,j}(\cdot,\omega_W)$. To ease the reading, we omit the notation $(\cdot,\omega_W)$ as classically done in probability. First, dealing with the first coordinate of $C_{n,j}$ defined in (A.30), one has

$$\begin{aligned} f_1\left(W_{n,j+1}\right)\Delta_{n,j,1} + f_1\left(W_{n,j}\right)\Delta_{n,j+1,1} &= \left(f_1\left(W_{n,j}\right) + f_1\left(W_{n,j+1}\right)\right)\Delta_{n,j,1} + f_1\left(W_{n,j}\right)\left(\Delta_{n,j+1} - \Delta_{n,j,1}\right) \\ &= \left(X_{\sigma_n(j)} - q_{j,n}\right)\left(f_1\left(W_{n,j}\right) + f_1\left(W_{n,j+1}\right)\right) f_{1,x}\left(W_{n,j}\right) + f_1\left(W_{n,j}\right)\left(\Delta_{n,j+1} - \Delta_{n,j,1}\right) \\ &\quad + \frac{1}{2}\left(X_{\sigma_n(j)} - q_{j,n}\right)^2 \left(f_1\left(W_{n,j}\right) + f_1\left(W_{n,j+1}\right)\right) f_{1,xx}\left(\delta_{n,j,1}, W_j\right) \end{aligned}$$

using the expansion of $\Delta_{n,j,1}$ given in (A.28). By A.24 and using the boundedness of f_1 and $f_{1,xx}$, we get that

$$\frac{1}{n}\sum_{j=1}^{n-1} \left(X_{\sigma_n(j)} - q_{j,n}\right)^2 \left(f_1\left(W_{n,j}\right) + f_1\left(W_{n,j+1}\right)\right) f_{1,xx}\left(\delta_{n,j,1}, W_j\right)$$

is $O_{\mathbb{P}}(1/n)$. The same asymptotic behavior is observed for the telescopic sum

$$\frac{1}{n}\sum_{j=1}^{n-1} f_1(W_{n,j})\,(\Delta_{n,j+1,1}-\Delta_{n,j,1}).$$

So that, using also the expansion of the $\Delta_{n,j,k}$ given in (A.28), A.24, and the boundedness of f_k and $f_{k,xx}$ for any k, the study of C_n reduces to that of the random vector

$$\frac{1}{n}\sum_{j=1}^{n-1}\left(X_{\sigma_n(j)}-q_{j,n}\right)\begin{pmatrix} f_{1,x}(W_{n,j})\times(f_1(W_{n,j})+f_1(W_{n,j+1})) \\ f_{1,x}(W_{n,j}) \\ 2f_{1,x}(W_{n,j})\,f_1(W_{n,j+1}) \\ \vdots \\ f_{l,x}(W_{n,j})\times(f_l(W_{n,j})+f_l(W_{n,j+1})) \\ f_{l,x}(W_{n,j}) \\ 2f_{l,x}(W_{n,j})\,f_l(W_{n,j+1}) \end{pmatrix} \tag{A.35}$$

by the independence between σ_n and $W_1,\dots,W_n$. In that view, let us consider the following linear combination $\sum_{k=1}^{l} u_k(f_k(W_{n,j})+f_k(W_{n,j+1}))+v+2wf_k(W_{n,j+1})$, where $(u_1,v_1,w_1,\cdots,u_l,v_l,w_l)\in\mathbb{R}^{3l}$ and the empirical mean

$$\sum_{k=1}^{k}\frac{1}{n}\sum_{j=1}^{n-1}\left(X_{\sigma_n(j)}-q_{j,n}\right)f_{k,x}(W_{n,j})\times(u_k(f_k(W_{n,j})+f_k(W_{n,j+1}))+v_k+2wf_k(W_{n,j+1})). \tag{A.36}$$

Now it remains to apply the lemmas prove in the beginning of this section with $\xi_j=(W_j,W_{j+1})$ and $\psi=\psi_{uvw}$ with

$$\psi_{uvw}\left(\xi_j,\frac{j}{n+1},\frac{j+1}{n+1}\right)=\sum_{k=1}^{l} f_{k,x}(W_{n,j})\,(u_k(f_k(W_{n,j})+f_k(W_{n,j+1}))+v_k+2w_kf_k(W_{n,j+1})), \tag{A.37}$$

noticing that, as $n\to\infty$, $(1/n)\sum_{j=1}^{n-1}\delta_{\xi_j,j/(n+1),(j+1)/(n+1)}$ converges in distribution to $Q=\nu=\mathcal{L}_{(X,X)}\otimes\mathcal{L}_W\otimes\mathcal{L}_W$. Thus we deduce that the empirical mean in (A.36) converges in distribution for any 3l-uplet (u,v,w). Since any linear combination of the components of the random vector defined in (A.35) satisfies a CLT, so does the random vector itself. The proof of Lemma 2 is now complete, up to the computation of the asymptotic variance-covariance matrix Σ_C done in the paragraph that follows.

Computation of the asymptotic covariance matrix Σ_C

We use the explicit expression (4) in the proof of the lemma proved above of the asymptotic variance σ_ψ^2 with $Q=\nu=\mathcal{L}_{(X,X)}\otimes\mathcal{L}_W\otimes\mathcal{L}_W$ and with ψ given by (A.37). Then taking the values $(1,0,0,\cdots)$, $(0,1,0,\cdots)$ and $(0,0,1,\cdots)$ – so on

and so forth – leads to the diagonal terms of the asymptotic variance-covariance matrix Σ_C while solving a three-dimensional system of equations provides the remaining terms.

The terms of the covariance matrix are as follow:

- If $i = 3(k-1)+1, j = 3(k'-1)+1, (k,k') \in [\![1,l]\!]^2$, we have

$$\begin{aligned}\Sigma_C^{i,j} = \\ \mathbb{E}\left[\left(f_k(X_1,W_1)+f_k(X_1,W_1')\right)\left(f_{k'}(X_2,W_2)+f_{k'}(X_2,W_2')\right)f_{k,x}(X_1,W_1)f_{k',x}(X_2,W_2)(F_X(X_1)\wedge F_X\right. \\ -\mathbb{E}\left[\left(f_k(X,W)+f_k(X,W')\right)f_{k,x}(X,W)F_X(X)\right]\mathbb{E}\left[\left(f_{k'}(X,W)+f_{k'}(X,W')\right)f_{k',x}(X,W)F_X(X)\right].\end{aligned}$$

- If $i = 3(k-1)+2, j = 3(k'-1)+2, (k,k') \in [\![1,l]\!]^2$, we have

$$\begin{aligned}\Sigma_C^{i,j} &= \mathbb{E}\left[f_{k,x}(X_1,W_1)f_{k',x}(X_2,W_2)(F_X(X_1)\wedge F_X(X_2))\right] \\ &\quad - \mathbb{E}\left[f_{k,x}(X,W)F_X(X)\right]\mathbb{E}\left[f_{k',x}(X,W)F_X(X)\right].\end{aligned}$$

- If $i = 3(k-1)+3, j = 3(k'-1)+3, (k,k') \in [\![1,l]\!]^2$, we have

$$\begin{aligned}\Sigma_C^{i,j} &= 4\mathbb{E}\left[f_k(X_1,W_1')f_{k'}(X_2,W_2')f_{k,x}(X_1,W_1)f_{k',x}(X_2,W_2)(F_X(X_1)\wedge F_X(X_2))\right] \\ &- 4\mathbb{E}\left[f_k(X,W')f_{k,x}(X,W)F_X(X)\right]\mathbb{E}\left[f_{k'}(X,W')f_{k',x}(X,W)F_X(X)\right].\end{aligned}$$

- If $i = 3(k-1)+1, j = 3(k'-1)+2, (k,k') \in [\![1,l]\!]^2$, we have

$$\begin{aligned}\Sigma_C^{i,j} &= \mathbb{E}\left[\left(f_k(X_1,W_1)+f_k(X_1,W_1')\right)f_{k,x}(X_1,W_1)f_{k',x}(X_2,W_2)(F_X(X_1)\wedge F_X(X_2))\right] \\ -\mathbb{E}&\left[\left(f_k(X,W)+f_k(X,W')\right)f_{k,x}(X,W)F_X(X)\right]\mathbb{E}\left[f_{k',x}(X,W)F_X(X)\right].\end{aligned}$$

- If $i = 3(k-1)+1, j = 3(k'-1)+3, (k,k') \in [\![1,l]\!]^2$, we have

$$\begin{aligned}\Sigma_C^{i,j} = \\ 2\mathbb{E}\left[\left(f_k(X_1,W_1)+f_k(X_1,W_1')\right)f_{k'}(X_2,W_2')f_{k,x}(X_1,W_1)f_{k',x}(X_2,W_2)(F_X(X_1)\wedge F_X(X_2))\right] \\ -2\mathbb{E}\left[\left(f_k(X,W)+f_k(X,W')\right)f_{k,x}(X,W)F_X(X)\right]\mathbb{E}\left[f_{k'}(X,W')f_{k',x}(X,W)F_X(X)\right].\end{aligned}$$

- If $i = 3(k-1)+2, j = 3(k'-1)+3, (k,k') \in [\![1,l]\!]^2$, we have

$$\begin{aligned}\Sigma_C^{i,j} &= 2\mathbb{E}\left[f_{k'}(X_2,W_2')f_{k,x}(X_1,W_1)f_{k',x}(X_2,W_2)(F_X(X_1)\wedge F_X(X_2))\right] \\ &\quad - 2\mathbb{E}\left[f_{k,x}(X,W)F_X(X)\right]\mathbb{E}\left[f_{k'}(X,W')f_{k',x}(X,W)F_X(X)\right].\end{aligned}$$

Asymptotic variance σ^2 of Theorem 4.1.3

We have proved yet that

$$\sqrt{n}\left(\begin{pmatrix}B_n\\C_n\end{pmatrix}-\begin{pmatrix}m_B\\0\end{pmatrix}\right)\xrightarrow[n\to\infty]{\mathcal{L}}\mathcal{N}_{6l}\left(0,\begin{pmatrix}\Sigma_B & 0\\0 & \Sigma_C\end{pmatrix}\right),$$

where the explicit expressions of m_B, Σ_B and Σ_C are given in (A.31) of Lemma 1, Appendices A.5.5 and A.5.5 respectively. Applying the so-called delta method [Van00, Theorem 3.1] to the linear function $f(x, y) = x + y$, we conclude that

$$\sqrt{n}(Z_n - m_B) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}_{3l}\left(0, \Sigma_B + \Sigma_C\right) \tag{A.38}$$

Further, we notice that $\bar{\xi}_n(f, \mathbf{X}^{(n)}) \stackrel{\mathcal{L}}{=} \Psi(Z_n)$ with $\Psi(x_1, y_1, z_1, \cdots, x_l, y_l, z_l) = \sum(x_i - y_i^2)/(\sum z_i - y_i^2)$. The so-called delta method [Van00, Theorem 3.1] then gives

$$\sqrt{N}\left(\bar{\xi}_n(\mathbf{f}, \mathbf{X}^{(n)}) - \bar{S}_1(\mathbf{f})\right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}_1(0, \sigma^2)$$

where $\bar{S}_1(\mathbf{f})$ is the first-order multivariate Sobol' index with respect to X and $\sigma^2 = g^\top(\Sigma_B + \Sigma_C)g$ with $g = \nabla\Psi(m_B)$. By assumption $\mathrm{Var}(Y) \neq 0$, Ψ is differentiable at m_B and we will see in the sequel that $g^\top(\Sigma_B + \Sigma_C)g \neq 0$, so that the application of the delta method is justified. By differentiation, we get that, for any x, y, and z so that $z \neq y^2$:

$$\frac{\partial\psi}{\partial x_i} = \frac{1}{\sum z_i - y_i^2} \tag{A.39}$$

$$\frac{\partial\psi}{\partial y_i} = \frac{-2y_i\left(\sum z_i - x_i\right)}{\left(\sum z_i - y_i^2\right)^2} \tag{A.40}$$

$$\frac{\partial\psi}{\partial z_i} = \frac{-\sum x_i - y_i^2}{\left(\sum z_i - y_i^2\right)^2} \tag{A.41}$$

$$\tag{A.42}$$

so that

$$g = \nabla\Psi(m_B) = \frac{1}{\mathrm{Var}(\mathbf{f}(\mathbf{X}))}\begin{pmatrix} 1 \\ 2\mathbb{E}[f_1(\mathbf{X})](\bar{S}_1(\mathbf{f}) - 1) \\ -\bar{S}_1(\mathbf{f}) \\ \vdots \\ 1 \\ 2\mathbb{E}[f_l(\mathbf{X})](\bar{S}_1(\mathbf{f}) - 1) \\ -\bar{S}_1(\mathbf{f}) \end{pmatrix}.$$

Hence the asymptotic variance σ^2 in Theorem 4.1.3 is finally given by $\sigma^2 = g^\top\left(\Sigma_B + \Sigma_C\right)g$ where Σ_B and Σ_C have been defined in Appendices A.5.5 and A.5.5 respectively. This concludes the proof.