Fairness seen as Global Sensitivity Analysis

Clément Bénesse, Fabrice Gamboa, Jean-Michel Loubes, Thibaut Boissin

Machine Learning manuscript No.
(will be inserted by the editor)

Fairness seen as Global Sensitivity Analysis

Clément Bénesse · Fabrice Gamboa · Jean-Michel Loubes · Thibaut Boissin

Received: date / Accepted: date

Abstract Ensuring that a predictor is not biased against a sensitive feature is the goal of fair learning. Meanwhile, Global Sensitivity Analysis (GSA) is used in numerous contexts to monitor the influence of any feature on an output variable. We merge these two domains, Global Sensitivity Analysis and Fairness, by showing how Fairness can be defined using a special framework based on Global Sensitivity Analysis and how various usual indicators are common between these two fields. We also present new Global Sensitivity Analysis indices, as well as rates of convergence, that are useful as fairness proxies.

Keywords Global Sensitivity Analysis · Fairness · Sobol' indices · Cramér-von-Mises indices · Disparate Impact

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. A large number of 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 [9, 14, 38] and [2] and references therein for a presentation of different fairness criteria. Most of these definitions stem back

C. Bénesse
Institut de Mathématiques de Toulouse
E-mail: clement.benesse@math.univ-toulouse.fr

F. Gamboa & J-M. Loubes
Institut de Mathématiques de Toulouse

T. Boissin
IRIT Saint-Exupéry, Toulouse

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 $f(\mathbf{X})$ that predicts a target Y, enables to detect unfair algorithmic treatments. Then, ensuring that predictors are fair is achieved by controlling previous measures, as done in [2, 8, 20, 22, 35, 46]. If this notion has been extensively studied for classification, recent work tackle the regression case as in [10, 22, 27] or [30].

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 [12, 17, 21, 25, 26] 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 [21, 26, 33] 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 [24] 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 [33] 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. Then, we propose 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. 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). We give estimates for the extended Sobol' and Cramér-von-Mises indices, along with respectively asymptotic normality (Theorem 1). We then 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 3).

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 often are 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 [12] or [25] 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 [1, 7, 17].

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 [17, 36] 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 3. We also prove asymptotic normality for extended Sobol' indices, which is a first to the best of our knowledge.

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 [43], these indices compare, thanks to the Hoeffding decomposition [44], 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 [44] on $f(\mathbf{X})$ – sometimes also called

ANOVA-decomposition – so that we may write

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

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

$$V = \sum_{k=1}^{p} V_k + \sum_{k_2 > k_1}^{p} V_{k_1,k_2} + \cdots + V_{1,\cdots,p}. \tag{2}$$

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:

$$S_{(m)} := \frac{V_{(m)}}{V}, \tag{3}$$

which is the expression of the so-called Sobol' sensitivity indices. The index S_k quantifies the proportion of the output's variance caused by the input X_k on its own. The index $S_{(m)}, k \in (m)$ 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.

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 [33], 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 $p_{X_{\sim i}} p_{X_i}$ – in this case, $U_{\sim p}$ is denoted $U_{\sim p}^{i+1}$. In the Appendix A, 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 1.

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

$$Sob_i := \frac{\mathrm{Var}[\mathbb{E}[f(\mathbf{X})|X_i]]}{\mathrm{Var}[f(\mathbf{X})]} \tag{4}$$

$$SobT_i := \frac{\mathbb{E}[\mathrm{Var}[f(\mathbf{X})|Z_i]]}{\mathrm{Var}[f(\mathbf{X})]} \tag{5}$$

$$Sob_i^{ind} := \frac{\mathrm{Var}[\mathbb{E}[f(\mathbf{X})|Z_i]]}{\mathrm{Var}[f(\mathbf{X})]} \tag{6}$$

$$SobT_i^{ind} := \frac{\mathbb{E}[\mathrm{Var}[f(\mathbf{X})|X_{\sim i}]]}{\mathrm{Var}[f(\mathbf{X})]}, \tag{7}$$

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 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.

Remark 1 If the features are independent, then for all $i \in [\![1, \cdots, p]\!]$, $Sob_i^{ind} = Sob_i$ and $SobT_i^{ind} = SobT_i$. The proof comes from the fact that in the independent case, we have $U_1^i = U_p^{i+1}$.

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

$$0 \leq Sob_i^{ind} \leq Sob_i \leq SobT_i \leq 1 \quad \text{and} \quad 0 \leq Sob_i^{ind} \leq SobT_i^{ind} \leq SobT_i \leq 1.$$

We refer to [33] and to the law of total variance for the proof. Note that, in general, there are no inequalities between Sob_i and $SobT_i^{ind}$.

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 $Sob_1 = SobT_1 = Sob_1^{ind} = SobT_1^{ind} = 0.5$, 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 $Sob_1 = SobT_1 = (\alpha^2 V(x_1))/(\alpha^2 V(x_1) + \sigma^2)$ while $Sob_1^{ind} = SobT_1^{ind} = 0$. 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.

Table 1: Sobol' indices: what is taken into account and what is not.

Sobol' indices		
	Correlation between variables	Joint contributions
Sob_i	✓	✗
$SobT_i$	✓	✓
Sob_i^{ind}	✗	✗
$SobT_i^{ind}$	✗	✓

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, $Sob_1 = Sob_1^{ind} = 0 = Sob_2 = Sob_2^{ind}$ while $SobT_1 = SobT_1^{ind} = 1 = SobT_2 = SobT_2^{ind}$ 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 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_i and $SobT_i$ or between Sob_i^{ind} and $SobT_i^{ind}$ 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_i^{ind} and Sob_i or between $SobT_i^{ind}$ and $SobT_i$ is big, a large part of the influence of the feature X_i will be through its intrinsic influence on other features.

These indices can be rewritten as follow, by using the Lévy-Rosemblatt theorem:

$$Sob_i := \frac{\mathrm{Var}[\mathbb{E}[g_i(\mathbf{U}^i)|U_1^i]]}{\mathrm{Var}[g_i(\mathbf{U}^i)]} \tag{8}$$

$$SobT_i := \frac{\mathbb{E}[\mathrm{Var}[g_i(\mathbf{U}^i)|U_{\sim 1}^i]]}{\mathrm{Var}[g_i(\mathbf{U}^i)]} \tag{9}$$

$$Sob_i^{ind} := \frac{\mathrm{Var}[\mathbb{E}[g_{i+1}(\mathbf{U}^{i+1})|U_p^{i+1}]]}{\mathrm{Var}[g_{i+1}(\mathbf{U}^{i+1})]} \tag{10}$$

$$SobT_i^{ind} := \frac{\mathbb{E}[\mathrm{Var}[g_{i+1}(\mathbf{U}^{i+1})|U_{\sim p}^{i+1}]]}{\mathrm{Var}[g_{i+1}(\mathbf{U}^{i+1})]}, \tag{11}$$

as explained in detail in [33, 34] or in Appendix B. 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 1 *Each index $\mathcal{S}$ in the equations* (4) *to* (7) *can be estimated by its empirical counter part $\mathcal{S}_n$ 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 B.*

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 [17], an answer to this lack of distributional information that will be of use later in a fairness framework – see Section 3.

2.3.1 Classical Cramér-von-Mises indices

The Cramér-von-Mises indices are based on the whole distribution of $f(\mathbf{X})$. They are defined (see [17]), for every input i, as follow:

$$CVM_i := \frac{\int_{\mathbb{R}} \mathbb{E}\left[(\mu(t) - \mu^i(t))^2\right] d\mu(t)}{\int_{\mathbb{R}} \mu(t)(1-\mu(t))d\mu(t)} \tag{12}$$

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

$$CVM_i = \frac{\int \mathrm{Var}(\mathbb{E}\left[\mathbb{1}_{f(\mathbf{X})\leq t}|X_i\right])d\mu(t)}{\int \mathrm{Var}(\mathbb{1}_{f(\mathbf{X})\leq t})d\mu(t)}. \tag{13}$$

As before, these indices extend to the multivariate case. Simple estimators have been proposed [7, 17], 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 (13) as:

$$CVM_i = \int Sob_i(\mathbb{1}_{f(\mathbf{X})\leq t}) \frac{\mathrm{Var}(\mathbb{1}_{f(\mathbf{X})\leq t})}{\int \mathrm{Var}(\mathbb{1}_{f(\mathbf{X})\leq t})d\mu(t)} d\mu(t). \tag{14}$$

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 enable to capture information about the distribution of the outcome instead of moment-related information.

2.3.2 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 [1]. This new set of indices will capture the influence of a feature independently of the rest of the features.

Definition 1 For every input i, we define the independent Cramér-von-Mises indices as:

$$CVM_i^{ind} := \frac{\int \mathbb{E}(Var(\mathbb{1}_{f(\mathbf{X})\leq t}|X_{\sim i}))d\mu(t)}{\int Var(\mathbb{1}_{f(\mathbf{X})\leq t})d\mu(t)} \tag{15}$$

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 (7), 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 D by the mean of estimates $\widehat{CVM}_i$. Similarly to Theorem 1, we have the following theorem.

Theorem 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 [18]. Note that new estimation procedures can be efficient with little data, as mentioned in [17], which will be helpful for measuring intersectional fairness in the following Section.

3 Fairness

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 realisation 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 two main common definitions of fairness are the following

- *Statistical Parity*, see for instance in [14], 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 from 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 $i = 0, 1$.

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 or described in [12, 25] are suitable indicators to quantify fairness as follows and these definitions can be extended to continuous predictors and continuous Y.

Definition 2 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 $\Lambda(\Phi(Y, \mathbf{X}), S)) = \mathrm{Var}(\mathbb{E}[f(\mathbf{X})|S])$. 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 [20].

Example 2 (Avoiding Disparate Treatment) The so-called *Avoiding Disparate Treatment* fairness is achieved by taking $\Lambda(\Phi(Y, \mathbf{X}), S)) = \mathbb{E}[\mathrm{Var}(f(\mathbf{X})|X)]$. This corresponds to the GSA measure $SobT_S(f(\mathbf{X}))$. 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 $\Lambda(\Phi(Y, \mathbf{X}), S)) = \mathbb{E}[\mathrm{Var}(\mathbb{E}[f(\mathbf{X})|S, Y]|Y)]$. 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 $\Lambda(\Phi(Y, \mathbf{X}), S)) = \mathrm{Var}(\mathbb{E}[\ell(f(\mathbf{X}), Y)|S])$ 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, [24] uses Shapley values, [32] uses HSIC measures, [19] 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.

Table 2: Common fairness definitions and associated GSA measures

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

In Table 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.

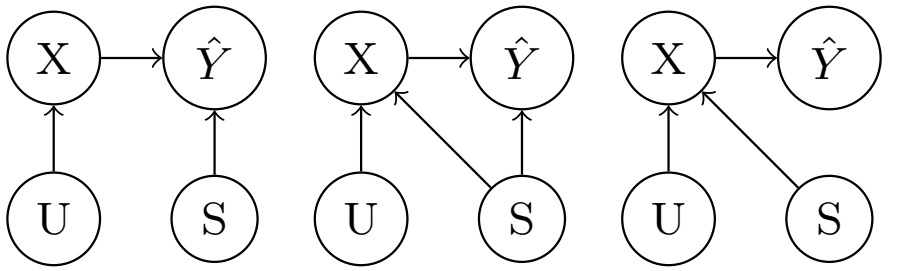
Remark 5 Note that many fairness measures are defined using discrete or binary sensitive variable. The GSA framework enables to handle continuous variables without additional difficulties. Moreover using kernel methods, GSA indices can be defined for a larger and more "exotic" variety of variables such as graphs or trees, for instance. In particular HSIC (see in [3, 12, 23, 36, 42]) is a kernel-based GSA measure that has been used in fairness.

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 [32]. 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



(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 both another variable X and the outcome $\hat{Y}$.

Fig. 1: Examples of representation of causal models with directed acyclic graphs.

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 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 [36]. 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.

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 [5, 39].

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 [5] and are illustrated in the various DAGs of Figure 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 [5, 29].

Example 6 (Causal graphs [41]) In this example, we specify three causal models and illustrate the previous proposition.

In Graph 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 1b, we have no information about the influence of S on the outcome.

In Graph 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

Table 3: Sobol' indices: what is taken into account and what is not.

Sobol' indices		
	Correlation between variables	Joint contributions
Sob_i	✓	✗
$SobT_i$	✓	✓
Sob_i^{ind}	✗	✗
$SobT_i^{ind}$	✗	✓

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 [16, 28].

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 [11]. Some recent works provide extensions of fairness measures to take into account the bias amplification due to intersectionality. We refer for instance to [37] or [15]. 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 3 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.

Table 4: Synthetic experiments based on causal DAGs – Figure 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)".

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.

4 Experiments

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 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 4 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

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 [13] and the COMPAS dataset.

4.2.1 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

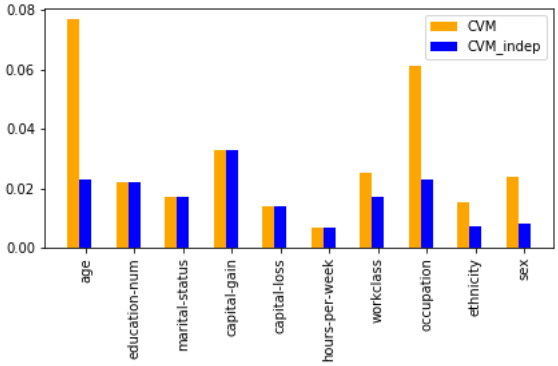
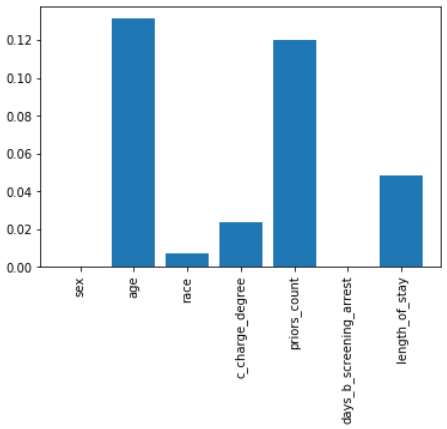


Fig. 2: Cramér-von-Mises and independent Cramér-von-Mises indices for the Adult dataset.

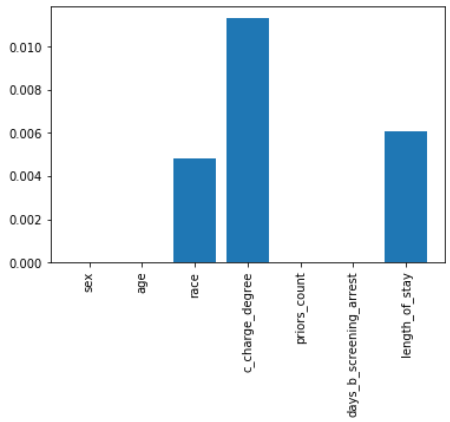
using an Extreme Gradient Boosting Procedure are shown in Figure 2. We used the same pre-process as [4] for the choice of variables.
If we look at the independent Cramér-von-Mises, we quantify the direct influence of a variable . We recover the influent indicators – "capital gain", "education-number", "age", "occupation"... – given by other studies [4, 16].
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 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 [4].

4.2.2 COMPAS dataset

The so-called COMPAS dataset, gathered by ProPublica described for instance in in [45] , 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



(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.

Fig. 3: Cramér-von-Mises indices for the COMPAS dataset.

analysed this dataset with Cramér-von-Mises indices in order to quantify fairness exhibited by the COMPAS algorithm. The results are shown in Figure 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.

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.

Acknowledgements Research partially supported by the AI Interdisciplinary Institute ANITI, which is funded by the French "Investing for the Future – PIA3" program under the Grant agreement ANR-19-PI3A-0004.

Conflict of interest

The authors declare that they have no conflict of interest.

References

1. Azadkia, M., Chatterjee, S.: A simple measure of conditional dependence. arXiv preprint arXiv:1910.12327 (2019)
2. del Barrio, E., Gordaliza, P., Loubes, J.M.: Review of mathematical frameworks for fairness in machine learning. arXiv preprint arXiv:2005.13755 (2020)
3. Berlinet, A., Thomas-Agnan, C.: A collection of examples. In: Reproducing Kernel Hilbert Spaces in Probability and Statistics, pp. 293–343. Springer (2004)
4. Besse, P., del Barrio, E., Gordaliza, P., Loubes, J.M., Risser, L.: A survey of bias in machine learning through the prism of statistical parity. The American Statistician **0**(ja), 1–25 (2021). DOI 10.1080/00031305.2021.1952897. URL `https://doi.org/10.1080/00031305.2021.1952897`
5. Bongers, S., Forré, P., Peters, J., Schölkopf, B., Mooij, J.M.: Foundations of structural causal models with cycles and latent variables. arXiv preprint arXiv:1611.06221 (2020)
6. Carlier, G., Galichon, A., Santambrogio, F.: From knothe's transport to brenier's map and a continuation method for optimal transport. SIAM Journal on Mathematical Analysis **41**(6), 2554–2576 (2010)
7. Chatterjee, S.: A new coefficient of correlation. Journal of the American Statistical Association pp. 1–21 (2020)
8. Chiappa, S., Jiang, R., Stepleton, T., Pacchiano, A., Jiang, H., Aslanides, J.: A general approach to fairness with optimal transport. In: AAAI, pp. 3633–3640 (2020)
9. Chouldechova, A.: Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big data **5**(2), 153–163 (2017)
10. Chzhen, E., Denis, C., Hebiri, M., Oneto, L., Pontil, M.: Fair Regression via Plug-in Estimator and Recalibration With Statistical Guarantees. Advances in Neural Information Processing Systems (2020)
11. Crenshaw, K.: Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. u. Chi. Legal f. p. 139 (1989)
12. Da Veiga, S.: Global sensitivity analysis with dependence measures. Journal of Statistical Computation and Simulation **85**(7), 1283–1305 (2015). DOI 10.1080/00949655.2014.945932. URL `https://hal.archives-ouvertes.fr/hal-01128666`
13. Dua, D., Graff, C.: UCI machine learning repository (2017). URL `http://archive.ics.uci.edu/ml`
14. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. In: Proceedings of the 3rd innovations in theoretical computer science conference, pp. 214–226. ACM (2012)
15. Foulds, J.R., Islam, R., Keya, K.N., Pan, S.: An intersectional definition of fairness. In: 2020 IEEE 36th International Conference on Data Engineering (ICDE), pp. 1918–1921. IEEE (2020)
16. Frye, C., Rowat, C., Feige, I.: Asymmetric shapley values: incorporating causal knowledge into model-agnostic explainability. Advances in Neural Information Processing Systems **33** (2020)
17. Gamboa, F., Gremaud, P., Klein, T., Lagnoux, A.: Global sensitivity analysis: a new generation of mighty estimators based on rank statistics. arXiv preprint arXiv:2003.01772 (2020)
18. Gamboa, F., Klein, T., Lagnoux, A.: Sensitivity analysis based on cramér–von mises distance. SIAM/ASA Journal on Uncertainty Quantification **6**(2), 522–548 (2018)

19. Ghassami, A., Khodadadian, S., Kiyavash, N.: Fairness in supervised learning: An information theoretic approach. In: 2018 IEEE International Symposium on Information Theory (ISIT), pp. 176–180. IEEE (2018)
20. Gordaliza, P., Del Barrio, E., Fabrice, G., Loubes, J.M.: Obtaining fairness using optimal transport theory. In: International Conference on Machine Learning, pp. 2357–2365 (2019)
21. Grandjacques, M.: Analyse de sensibilité pour des modèles stochastiques à entrées dépendantes: application en énergétique du bâtiment. Ph.D. thesis, Grenoble Alpes (2015)
22. Grari, V., Ruf, B., Lamprier, S., Detyniecki, M.: Fairness-aware neural réyni minimization for continuous features (2019)
23. Gretton, A., Herbrich, R., Smola, A., Bousquet, O., Schölkopf, B.: Kernel methods for measuring independence. Journal of Machine Learning Research **6**(Dec), 2075–2129 (2005)
24. Hickey, J.M., Stefano, P.G.D., Vasileiou, V.: Fairness by explicability and adversarial shap learning (2020)
25. Iooss, B., Lemaître, P.: A review on global sensitivity analysis methods. In: Uncertainty management in simulation-optimization of complex systems, pp. 101–122. Springer (2015)
26. Jacques, J., Lavergne, C., Devictor, N.: Sensitivity analysis in presence of model uncertainty and correlated inputs. Reliability Engineering & System Safety **91**(10-11), 1126–1134 (2006)
27. Jeremie Mary Clement Calauzenes, N.E.K.: Fairness-aware learning for continuous attributes and treatments (2019)
28. Kilbertus, N., Carulla, M.R., Parascandolo, G., Hardt, M., Janzing, D., Schölkopf, B.: Avoiding discrimination through causal reasoning. In: Advances in Neural Information Processing Systems, pp. 656–666 (2017)
29. de Lara, L., González-Sanz, A., Asher, N., Loubes, J.M.: Counterfactual models: The mass transportation viewpoint (2021)
30. Le Gouic, T., Loubes, J.M., Rigollet, P.: Projection to fairness in statistical learning. arXiv e-prints pp. arXiv–2005 (2020)
31. Lévy, P.: Théorie de l'addition des variables aléatoires, vol. 1. Gauthier-Villars (1954)
32. Li, Z., Perez-Suay, A., Camps-Valls, G., Sejdinovic, D.: Kernel dependence regularizers and gaussian processes with applications to algorithmic fairness. arXiv preprint arXiv:1911.04322 (2019)
33. Mara, T.A., Tarantola, S.: Variance-based sensitivity indices for models with dependent inputs. Reliability Engineering & System Safety **107**, 115–121 (2012)
34. Mara, T.A., Tarantola, S., Annoni, P.: Non-parametric methods for global sensitivity analysis of model output with dependent inputs. Environmental modelling & software **72**, 173–183 (2015)
35. Mary, J., Calauzènes, C., El Karoui, N.: Fairness-aware learning for continuous attributes and treatments. In: International Conference on Machine Learning, pp. 4382–4391 (2019)
36. Meynaoui, A., Marrel, A., Laurent, B.: New statistical methodology for second level global sensitivity analysis. arXiv preprint arXiv:1902.07030 (2019)
37. Morina, G., Oliinyk, V., Waton, J., Marusic, I., Georgatzis, K.: Auditing and achieving intersectional fairness in classification problems. arXiv preprint arXiv:1911.01468 (2019)
38. Oneto, L., Chiappa, S.: Recent Trends in Learning From Data. Springer (2020)
39. Pearl, J.: Causality. Cambridge university press (2009)
40. Rosenblatt, M.: Remarks on a multivariate transformation. Ann. Math. Statist. **23**(3), 470–472 (1952). DOI 10.1214/aoms/1177729394. URL `https://doi.org/10.1214/aoms/1177729394`
41. Rothenhäusler, D., Meinshausen, N., Bühlmann, P., Peters, J.: Anchor regression: heterogeneous data meets causality. arXiv preprint arXiv:1801.06229 (2018)
42. Smola, A., Gretton, A., Song, L., Schölkopf, B.: A hilbert space embedding for distributions. In: International Conference on Algorithmic Learning Theory, pp. 13–31. Springer (2007)
43. Sobol', I.M.: On sensitivity estimation for nonlinear mathematical models. Matematicheskoe modelirovanie **2**(1), 112–118 (1990)
44. Van der Vaart, A.W.: Asymptotic statistics, vol. 3. Cambridge university press (2000)
45. Washington, A.L.: How to argue with an algorithm: Lessons from the compas-propublica debate. Colo. Tech. LJ **17**, 131 (2018)
46. Williamson, R.C., Menon, A.K.: Fairness risk measures. arXiv preprint arXiv:1901.08665 (2019)

A 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 3 (Lévy-Rosemblatt theorem, [31, 40]) *: there is a bijection (denoted "RT" for Rosemblatt transform) between $p(\mathbf{X})$ and p independent uniform random variables*

$$(X_i, (X_{i+1}|X_i), \cdots, (X_{i-1}|X_{\sim(i-1)})) \sim p_{\mathbf{X}} \xrightarrow{RT} (U_1^i, \cdots, U_p^i) \sim \mathcal{U}^p(0,1). \tag{16}$$

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 encapsules 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

$$z_1 = F_{X_1}^{-1}(u_1^1) \tag{17}$$

$$z_2 = F_{X_2|X_1=x_1}^{-1}(u_2^1). \tag{18}$$

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 7 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 8 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 9 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.

B 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:

$$Sob_i = \frac{V[\mathbb{E}[g_i(\mathbf{U}^i)|U_1^i]]}{V[g_i(\mathbf{U}^i)]} = \frac{V[\mathbb{E}[f(\mathbf{X})|X_i]]}{V[f(\mathbf{X})]} \tag{19}$$

$$SobT_i = \frac{\mathbb{E}[V[g_i(\mathbf{U}^i)|U_{\sim 1}^i]]}{V[g_i(\mathbf{U}^i)]} = \frac{\mathbb{E}[V[f(\mathbf{X})|Z_i]]}{V[f(\mathbf{X})]} \tag{20}$$

$$Sob_i^{ind} = \frac{V[\mathbb{E}[g_{i+1}(\mathbf{U}^{i+1})|U_p^{i+1}]]}{V[g_{i+1}(\mathbf{U}^{i+1})]} = \frac{V[\mathbb{E}[f(\mathbf{X})|Z_i]]}{V[f(\mathbf{X})]} \tag{21}$$

$$SobT_i^{ind} = \frac{\mathbb{E}[V[g_{i+1}(\mathbf{U}^{i+1})|U_{\sim p}^{i+1}]]}{V[g_{i+1}(\mathbf{U}^{i+1})]} = \frac{\mathbb{E}[V[f(\mathbf{X})|X_{\sim i}]]}{V[f(\mathbf{X})]} \tag{22}$$

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:

$$\begin{aligned}
(u_1^i, \cdots, u_p^i) &\xrightarrow{IRT} \mathbf{x} = (x_i, \cdots, x_{i-1}) \sim p(\mathbf{X}), \\
(u_1^{i\prime}, \cdots, u_p^{i\prime}) &\xrightarrow{IRT} \mathbf{x}' = (x_i', \cdots, x_{i-1}') \sim p(\mathbf{X}), \\
(u_1^i, u_2^{i\prime}, \cdots, u_p^{i\prime}) &\xrightarrow{IRT} \mathbf{x}^i = (x_i, x_{i+1}' \cdots, x_{i-1}') \sim p(X_i)p(X_{\sim i}|X_i), \\
(u_1^{i\prime}, \cdots, u_{p-1}^{i\prime}, u_p^i) &\xrightarrow{IRT} \mathbf{x}^{i-1} = (x_i', x_{i+1}' \cdots, x_{i-1}) \sim p(X_{\sim i-1})p(X_{i-1}|X_{\sim i-1}).
\end{aligned} \tag{23}$$

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:

$$\begin{aligned}
\widehat{Sob_i} &= \frac{\frac{1}{N}\sum_{k=1}^N f(\mathbf{x}_k) \times \left(f(\mathbf{x}_k^i) - f(\mathbf{x}_k')\right)}{\hat{V}} \\
\widehat{SobT_i^{ind}} &= \frac{\frac{1}{N}\sum_{k=1}^N \left(f(\mathbf{x}_k^{i-1}) - f(\mathbf{x}_k')\right)^2}{2\hat{V}} \\
\widehat{Sob_{i-1}^{ind}} &= \frac{\frac{1}{N}\sum_{k=1}^N f(\mathbf{x}_k) \times \left(f(\mathbf{x}_k^{i-1}) - f(\mathbf{x}_k')\right)}{\hat{V}} \\
\widehat{SobT_i} &= \frac{\frac{1}{N}\sum_{k=1}^N \left(f(\mathbf{x}_k^i) - f(\mathbf{x}_k')\right)^2}{2\hat{V}},
\end{aligned} \tag{24}$$

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 $\mathbf{x}^*$.

C Central Limit Theorem for Sobol' indices

We recall the theorem 1 we presented in Section 2.

Theorem 4 *Each index $\mathcal{S}$ in the equations (4) to (7) 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:*

$$\sqrt{n}(\mathcal{S}_n - \mathcal{S}) \xrightarrow{D} \mathcal{N}(0, \sigma^2) \tag{25}$$

with σ^2 depending on which index we study.

We propose to study the central limit theorem for the estimator of the index Sob_i proposed in Appendix B. Note that the result is the same for other estimators of the Sobol' indices proposed in the same section.

If we denote

$$Z_n = \begin{pmatrix} n^{-1}\sum f(X_{i,k}, X_{\sim i,k}) f(X_{i,k}, X'_{\sim i,k}) \\ n^{-1}\sum f(X_{i,k}, X_{\sim i,k}) f(X'_{i,k}, X'_{\sim i,k}) \\ n^{-1}\sum f(X_{i,k}, X_{\sim i,k}) \\ n^{-1}\sum f^2(X_{i,k}, X_{\sim i,k}) \end{pmatrix} \tag{26}$$

then the estimator $\widehat{Sob_i}$ of the Sobol' index Sob_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 [44], we obtain the convergence of $h(Z_n)$ to $h(Z) = Sob_i$

$$\sqrt{n}\left(\widehat{Sob_i} - Sob_i\right) \to \mathcal{N}(0, \nabla h(\beta) \Sigma \nabla h(\beta)^T), \tag{27}$$

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

$$\Sigma = \begin{pmatrix} \sigma_{11}^2 & \sigma_{12}^2 & \sigma_{13}^2 & \sigma_{14}^2 \\ \sigma_{12}^2 & \sigma_{22}^2 & 0 & 0 \\ \sigma_{13}^2 & 0 & \sigma_{33}^2 & \sigma_{34}^2 \\ \sigma_{14}^2 & 0 & \sigma_{34}^2 & \sigma_{44}^2 \end{pmatrix} \tag{28}$$

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

$$\begin{aligned}
\sigma_{11}^2 =& \mathrm{Var}(f(X, X_{\sim i}) f(X, X'_{\sim i})) \\
\sigma_{12}^2 =& \mathbb{E}[f^2(X, X_{\sim i}) f(X, X'_{\sim i}) f(X', X'_{\sim i})] \\
\sigma_{13}^2 =& \mathbb{E}[f^2(X, X_{\sim i}) f(X, X'_{\sim i})] \\
\sigma_{14}^2 =& \mathbb{E}[f^3(X, X_{\sim i}) f(X, X'_{\sim i}) f(X', X'_{\sim i})] - \mathbb{E}[f^2(X, X_{\sim i})]\mathbb{E}[f(X, X'_{\sim i}) f(X, X'_{\sim i})] \\
\sigma_{22}^2 =& \mathrm{Var}(f(X, X_{\sim i}))^2 \\
\sigma_{33}^2 =& \mathrm{Var}(f(X, X_{\sim i})) \\
\sigma_{34}^2 =& \mathbb{E}[f^3(X, X_{\sim i})] \\
\sigma_{44}^2 =& \mathbb{E}[f^4(X, X_{\sim i}) - \mathbb{E}[f^2(X, X_{\sim i})]^2.
\end{aligned} \tag{29}$$

D 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 (15).

The first one is to use the fact that

$$\begin{aligned}
U(Y, X_i | \mathbf{Z}) &= \frac{\int \mathbb{E}(\mathrm{Var}(\mathbb{E}\left[\mathbb{1}_{Y \le t} | X_i, \mathbf{Z}\right] | \mathbf{Z})) d\mu(t))}{\int \mathrm{Var}(\mathbb{1}_{Y \le t}) d\mu(t))} \\
&= T(Y, X_i | \mathbf{Z}) \times (1 - T(Y, \mathbf{Z})).
\end{aligned} \tag{30}$$

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 [1].

Consider a triple of random variables (X, Z, Y) and an i.i.d sample $(X_i, Z_i, Y_i)_{1\leq i\leq n}$. 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 $Y_j \leq Y_i$.

The correlation coefficient defined in [1] is defined as:

$$T_n(Y, X|Z) = \frac{\sum_{i=1}^n \left(\min\{R_i, R_{M(i)}\} - \min\{R_i, R_{N(i)}\}\right)}{\sum_{i=1}^n \left(R_i - \min\{R_i, R_{N(i)}\}\right)}. \tag{31}$$

The authors of [1] 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

$$U_n(Y, X_i|X_{\sim i}) = T_n(Y, X_i|X_{\sim i}) \times (1 - T_n(Y, X_{\sim i}))\,. \tag{32}$$

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 [1] and [7]. We have the two following convergences almost surely:

$$Q_n(Y, X|Z) = n^{-2}\sum_{j=1}^n \left(\min\{R_j, R_{M(j)}\} - \min\{R_j, R_{N(j)}\}\right) \to \int \mathbb{E}(\mathrm{Var}(\mathbb{E}\left[\mathbb{1}_{Y\leq t}|X, Z\right]|Z))d\mu(t)) \tag{33}$$

$$S_n(Y) = n^{-3}\sum_{j=1}^n L_j(n - L_j) \to \int \mathrm{Var}(\mathbb{1}_{Y\leq t})d\mu(t)) \tag{34}$$

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

Proposition 4 (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.

E Proofs

E.1 Proof of Theorem 3

Proof Indeed, we can always write

$$p_{\mathbf{X}} = p_{X_i} \times p_{X_{i+1}|X_i} \times \cdots \times p_{X_{i-1}|X_{\sim(i-1)}}. \tag{35}$$

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 [6] – and by doing so for each variable defined above, we have the so-called Levy-Rosenblatt transform, denoted here as RT, that is:

$$(X_i, (X_{i+1}|X_i), \cdots, (X_{i-1}|X_{\sim(i-1)})) \sim p_{\mathbf{X}} \xrightarrow{RT} (U_1^i, \cdots, U_p^i) \sim \mathcal{U}^p(0, 1). \tag{36}$$

E.2 Proof of Examples following 2

Proof We will show here how each definition of fairness and GSA measure presented in Table 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.

E.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 $SobT^{ind}_S = 0$ then $\mathrm{Var}(\psi_S(S) + \psi_{S,X}(S) \times \psi_{X,S}(X(S))) = 0$. By orthogonality in the Hoeffding decomposition, $\mathrm{Var}(\psi_S(S)) = \mathrm{Var}(\psi_{S,X}(S) \times \psi_{X,S}(X(S))) = 0$, 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 $\mathrm{Var}(\psi'_S(S(X)) + \psi'_{S,X}(X) \times \psi'_{X,S}(S(X))) = 0$. By orthogonality in the Hoeffding decomposition, $\mathrm{Var}(\psi'_S(S(X))) = \mathrm{Var}(\psi'_{S,X}(X) \times \psi'_{X,S}(S(X)) = 0$, which lead to $\psi'_S(S(X)) = \psi'_{S,X}(X) \times \psi'_{X,S}(S(X)) = 0$. It holds that $Y = \psi'_X(X)$.

E.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 $f(X) = S_1 \times S_2$.