Fairness seen as Global Sensitivity Analysis

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

FAIRNESS SEEN AS GLOBAL SENSITIVITY ANALYSIS

Clément Bénesse*
Institut de Mathématiques de Toulouse
Université de Toulouse III
Toulouse, France
clement.benesse@math.univ-toulouse.fr

Fabrice Gamboa*
Institut de Mathématiques de Toulouse
Université de Toulouse III
Toulouse, France

Jean-Michel Loubes *
Institut de Mathématiques de Toulouse
Université de Toulouse III
Toulouse, France

Thibaut Boissin*
IRT Saint Exupery
Toulouse, France

ABSTRACT

Ensuring that a predictor is not biased against a sensible feature is the key of Fairness learning. Conversely, Global Sensitivity Analysis is used in numerous contexts to monitor the influence of any feature on an output variable. We reconcile these two domains by showing how Fairness can be seen as a special framework of 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.

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 [13, 7, 31] and [11] 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 sensible 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 [29, 39, 19, 17, 11, 6]. If this notion has been extensively studied for classification, recent work tackle the regression case as in [19, 24, 8] or [26].

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 [23, 10, 22, 18, 16] 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 a usual assumption in this field is to suppose the inputs to be independent, some works [23, 28, 18] remove this assumption to go further in the understanding of the

*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.

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 [21] 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 [28] 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 2.1). 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 [10] or [22] 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, Y. For this, we compare, for a feature X_i and the output Y, the probability distribution $\mathbb{P}_{X_i,Y}$ and the product probability distribution $\mathbb{P}_{X_i}\mathbb{P}_Y$ 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 indexes, 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, 16] and [5].

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 [36], these indices compare, thanks to the Hoeffding decomposition [37], 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 $Y = f(\mathbf{X}) = f(X_1, \cdots, X_p)$ where Y is an unidimensionnal square-integrable output, f an algorithm considered as a black-box and $X_1, \cdots, X_p$ inputs, with p the number of parameters. We denote by $p_{\mathbf{X}}$ the distribution of $\mathbf{X}$. For now, we suppose the different inputs to be independent, meaning that $p_{\mathbf{X}} = \otimes_{k=1}^{p} p_{X_k}$. Then, we can use the Hoeffding decomposition [37] on Y – sometimes also called ANOVA-decomposition – so that we may write

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

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 $V = \text{Var}(Y)$ and $V_s = \text{Var}(f_s(\mathbf{X}_s))$. 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 Y that is only due to the variable X_1 – that is, how much Y 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 Y 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_k by V, we get the expression of the so-called Sobol' sensitivity indices. These indices quantify the proportion of the output's variance caused by the input X_k, jointly or not with other inputs.

2.2 Sobol' indices for non-independent inputs

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. When the inputs are not independent, we need to duplicate each index in order to differentiate whether influences caused by correlations between inputs are taken into account or not. Introduced in this framework by [28], we use the Lévy-Rosemblatt theorem to create two mappings of interest. We denote by $\sim i$ every index other than i. We create a mapping between p independent uniform random variables U and the variables $\mathbf{X}$ either by mapping $p_{U_1^i} p_{U_{\sim 1}^i}$ to $p_{X_i} p_{X_{\sim i}|X_i}$ or by mapping $p_{U_{\sim p}^{i+1}} p_{U_p^{i+1}}$ to $p_{X_{\sim i}} p_{X_i}$. In the Annex 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:

$$Sob_i = \frac{V[\mathbb{E}[g_i(\mathbf{U}^i)|U_1^i]]}{V[g_i(\mathbf{U}^i)]} = \frac{V[\mathbb{E}[f(\mathbf{X})|X_i]]}{V[f(\mathbf{X})]} \tag{3}$$

$$SobT_i = \frac{\mathbb{E}[V[g_i(\mathbf{U}^i)|U_{\sim 1}^i]]}{V[g_i(\mathbf{U}^i)]} = \frac{\mathbb{E}[V[f(\mathbf{X})|Z_i]]}{V[f(\mathbf{X})]} \tag{4}$$

$$Sob_i^{ind} = \frac{V[\mathbb{E}[g_{i+1}(\mathbf{U}^{i+1})|U_p^{i+1}]]}{V[g_{i+1}(\mathbf{U}^{i+1})]} = \frac{V[\mathbb{E}[f(\mathbf{X})|Z_i]]}{V[f(\mathbf{X})]} \tag{5}$$

$$SobT_i^{ind} = \frac{\mathbb{E}[V[g_{i+1}(\mathbf{U}^{i+1})|U_{\sim p}^{i+1}]]}{V[g_{i+1}(\mathbf{U}^{i+1})]} = \frac{\mathbb{E}[V[f(\mathbf{X})|X_{\sim i}]]}{V[f(\mathbf{X})]}, \tag{6}$$

where the random variable Z_i has the distribution $p_{X_i|X_{\sim i}}$ and is equal to $F^{-1}_{X_i|X_{\sim i}}(U_p^{i+1})$.

Remark 1. *If the features are independent, then for all $i \in [\![1, \cdots, p]\!]$, $U_1^i = U_p^{i+1}$. This leads to the equalities $Sob_i^{ind} = Sob_i$ and $SobT_i^{ind} = SobT_i$.*

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 2. *Thanks to the law of total variance – see [28], various bounds and equality can be found between indices, echoing the various properties previously described for Sobol' indices with independent inputs. For instance, for all $i \in \{1, \cdots, p\}$, $0 \leq Sob_i^{ind} \leq Sob_i \leq SobT_i \leq 1$ and $0 \leq Sob_i^{ind} \leq SobT_i^{ind} \leq SobT_i \leq 1$. Note that, in general, there are no inequalities between Sob_i and $SobT_i^{ind}$.*

In order to better understand these indices, we present the three typical ways for a feature to modify the output of an algorithm.

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

SOBOL' INDICES		
	CORRELATION BETWEEN VARIABLES	JOINED CONTRIBUTIONS
Sob_i	✓	✗
$SobT_i$	✓	✓
Sob_i^{ind}	✗	✗
$SobT_i^{ind}$	✗	✓

1. 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 $f(\mathbf{x}) = x_1$ and x_1 independent to the rest of the variables. In this example, we would have $Sob_1 = SobT_1 = Sob_1^{ind} = SobT_1^{ind} = 1$.
2. Secondly, 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 $Sob_1^{ind} = SobT_1^{ind} = 0$.
3. 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 $f(\mathbf{x}) = x_1 \times x_2$. There, the distinction will be 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 correlation or joined contribution 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 joined 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.

We provide Monte-Carlo estimation of the extended Sobol' indices in the Annex B. 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 2.1. *Each index $\mathcal{S}$ in the equations* (3) *to* (6) *can be estimated by their 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 Annex B.*

2.3 Cramér-von-Mises indices

Sobol' indices are based on a decomposition of the variance, and therefore only quantify the second order influence of the inputs. 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 the means of divergences, or measures of dissimilarity between distributions for example. We recall here the Cramér-von-Mises indices [16], an answer to this issue 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 Y. We denote by $(X_i, i = 1, \cdots, p)$ the inputs and by $(X^j, Y^j), j = 1, \cdots, n$ a n-sample of both inputs and outputs. They are defined (see [16]), for every input i, as follow:

$$CVM_i = \frac{\int_{\mathbb{R}} \mathbb{E}\left[(\mu(t) - \mu^i(t))^2\right] d\mu(t)}{\int_{\mathbb{R}} \mu(t)(1 - \mu(t)) d\mu(t)} \tag{7}$$

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

This equation can be rewritten as

$$CVM_i = \frac{\int \text{Var}(\mathbb{E}\left[\mathbb{1}_{Y \leq t}|X_i\right])d\mu(t)}{\int \text{Var}(\mathbb{1}_{Y \leq t})d\mu(t)}. \tag{8}$$

As before, these indices extend to the multivariate case. Simple estimators have been proposed [5, 16], and are based on permutations and rankings. We will recall the estimation in a following section.

Remark 3. *The Cramér-von-Mises index is an extension of the Sobol' index to quantify more than just the second-order influence of the inputs on the output.*

If we recall the definition of the Sobol' index Sob_i, it is the ratio between the variance $\text{Var}(\mathbb{E}[Y|X_i])$ explained by the feature X_i and the total variance of the algorithm $\text{Var}(Y)$. However, as mentioned earlier, the issue with this comparison of variances is that only second-order influence is quantified, which can be limiting. Some way to capture more complex influences on the whole distribution is to work with level lines of Y, that is to look at $Sob_i(\mathbb{1}_{Y \leq t})$ and to integrate on the whole range of admissible t. This lead to the definition of the Cramér-von-Mises index defined in equation (8).

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

$$CVM_i^{ind} = \frac{\int \mathbb{E}(\text{Var}(\mathbb{1}_{Y \leq t}|X_{\sim i}))d\mu(t)}{\int \text{Var}(\mathbb{1}_{Y \leq t})d\mu(t)} \tag{9}$$

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 (6), 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.

Estimation of these indices is given in Annex D.

3 Fairness

3.1 Sensitivity Index as Fairness measures

In this section, we provide a probabilistic framework to unify all the various Fairness definitions as Global Sensitivity Indices. Several measures of fairness corresponding to different definitions of fairness have been proposed in the machine learning literature. The *Statistical Parity* see for instance in [13], requires that the algorithm f, predicting a target Y, has similar outputs for all the values of S in the sense that $\mathbb{P}(f(\mathbf{X}) = 1|S) = \mathbb{P}(f(\mathbf{X}) = 1)$ 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 fairness here implies that $f(\mathbf{X}) \perp\!\!\!\perp S|Y$. 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 theorem proves that GSA measures as defined in 2 or described in [10, 22] are suitable indicators to quantify fairness as follows and that these definitions can be extended to continuous predictors and continuous Y.

Table 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	$\text{VAR}(\mathbb{E}[\ell(f(\mathbf{X}),Y)\|S]) \to Sob_S(\ell(f(\mathbf{X}),Y))$

Definition 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.*

Theorem 3.1. *Let Φ be a function of the features and Γ be a GSA measure for Φ and S. Then, Γ induces a Fairness measure defined as $\Lambda(\Phi(Y,\mathbf{X}),S) = \Gamma(\Phi(Y,\mathbf{X}),S)$.*

Among well known fairness measures, we point out that we immediately recover two main fairness measures used in the fair learning literature. GSA fairness measures can be computed for different function Φ 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.

- In the classification case, set $\Phi(Y,\mathbf{X}) = \mathbb{1}_{f(\mathbf{X})=1}$. The GSA measure quantifies the influence of a random variable on the predictions given by the function f. The disparate impact defined as $DI(f) = |\mathbb{P}(f(\mathbf{X})=1|S=0) - \mathbb{P}(f(\mathbf{X})=1|S=1)|$ used top measure fairness in this case can be computed as $DI(f) = \Lambda(\Phi(Y,f(\mathbf{X})))$ with $\Lambda(t,S) = |\mathbb{E}[t|S=0] - \mathbb{E}[t|S=1]|$.
- A second possibility is to take $\Phi(Y,\mathbf{X}) = \ell(f(\mathbf{X}),Y)$, with ℓ a chosen loss function. In this case, the GSA measure will quantify the influence of the sensitive variable on the performance of the predictor. Here Avoiding Disparate Mistreatment measured as $ADM = |\mathbb{P}(f(\mathbf{X}) \neq Y|S=1) - \mathbb{P}(f(\mathbf{X}) \neq Y|S=0)|$ is a GSA index by setting $ADM = \Lambda(\Phi(Y,f(\mathbf{X})))$ with $\Lambda(x) = \Lambda(t,S) = |\mathbb{E}[t|S=0] - \mathbb{E}[t|S=1]|$ and $\Phi(Y,\mathbf{X}) = \mathbb{1}_{f(\mathbf{X})\neq Y}$.

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

Proposition 1 (Classical Fairness definitions and associated GSA measures)**.** *For any fairness definition of Table 2, there exist a GSA measure Γ for a certain Φ and the sensible feature S such that a predictor is fair if and only if $\Gamma(\Phi(Y,X,S),S) = 0$.*

In Table 2, we give the different indexes associated to classical studied fairness definitions, given for the special case where S and the predictor are both binary.

Remark 5. *Many fairness measures are defined using discrete or binary sensitive variable. The GSA framework enables to handle continuous variables. Moreover using kernel methods, GSA indices based on kernels can be used for a large variety of variables such as graphs or trees, for instance. In particular HSIC (see in [10, 2, 20, 35]) is a GSA measure that can be used in fairness.*

3.2 Applications to Causal Models

Fairness is often measured using the outcome of an algorithm, yet from a legal point of view, discrimination is often related to the causal effect of a variable. Causality is often modeled using DAG – directed acyclic graphs [32]. Actually, the graph structure of a DAG enables to visualize the different interactions variables can have with each other.

In this subsection, we show how to address causal notions of fairness using the GSA framework, illustrated by two examples.

Example 1 (Causal graphs [34])**.** *Lets consider a situation in which a protected variable S influence others variables X, conjointly with an unobserved variable U. Then, the couple (X,S) are the inputs of a predictor Y corresponding to*

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

where ϕ and ψ are some unknown functions. These equations are a consequence of the unique solvability of acyclic models [3] and are illustrated in the various DAGs of Figure 1.

The influence of S on the predictor Y is then two-fold as S can change directly the outcome Y or through its influence on X.

Example 2 (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 Y the admission decision. The gender should not directly influence any admission decision 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 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 Y. This issue on unresolved discrimination is tackled in [25, 15].*

In many practical cases, the causal graph is unknown and we need indices to quantify causality. We will show how GSA can quantify causal influence following DAG structure, in particular , the Total Sobol and the Total Independent Sobol indices. Different GSA indices will correspond to different paths from S to Y. Two type of relationships can be measured, represented either by a path from S directly to Y or a path from S to another variable X that influences itself the predictor Y.

Proposition 2. *The condition $SobT_S^{ind} = 0$ implies that the direct path from S to Y is non-existent. Similarly, the condition $SobT_S = 0$ implies that every path from S to Y is non-existent.*

Proof. The proof is a direct consequence of the Hoeffding decomposition of the function $Y = \psi(X, S)$ and will be developed in the Annex. □

In the case where $SobT_S^{ind} = 0$, S can still be influential through X but not on its own. This type of fairness corresponds to *unresolved discrimination*. However, if $SobT_S = 0$, S is not influential on the outcome at all.

Remark 6. *Fairness such as unresolved discrimination, sought after in cases such as Example 2, correspond to the fact that S cannot be influential on its own, only through the other inputs X. This is equivalent to the non-existence of a direct path from S to Y, i.e. the condition $SobT_S^{ind} = 0$.*

3.3 Quantifying intersectionality unfairness with GSA index

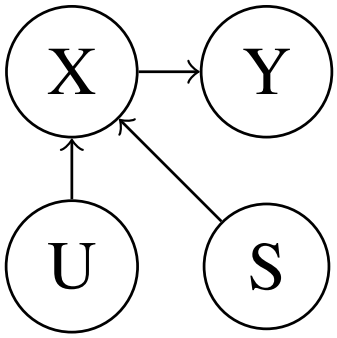
Most of fairness results are stated in the case where there is only one sensitive variable. Yet in many cases, the bias and the resulting possible discrimination are the result of multiple sensitive variables. This situation is known as intersectionality, when the level of discrimination of an intersection of several minority groups is worse than the discrimination present in each group as presented in [9]. Some very recent works provide extensions of fairness measures to take into account the bias amplification due to intersectionality. We refer for instance to [30] or [14]. 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.
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 understood 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.

Intersectionality fairness is obtained when the variables S_1 and S_2 do not have any joined influence on the output of the algorithm.

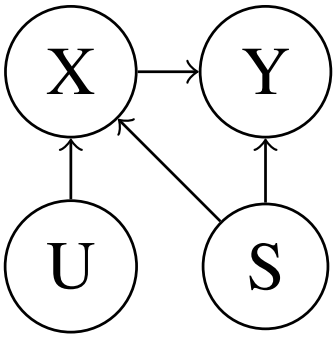
Definition 3.2. *Let $S_1, S_2, \cdots, S_m$ be sensitive features. It is said that an algorithm output is intersectionaly fair if $\Gamma(\Phi(X, S_1, \cdots, S_m); (S_1, \cdots, S_m)) = 0$. This constraint can be relaxed to $\Gamma(\Phi(X, S_1, \cdots, S_m); (S_1, \cdots, S_m)) \leq \varepsilon$ with ε small for approximate intersectionality fairness.*

Remark 7. *Intersectionality fairness is different than classical fairness. Classical fairness is usually interested only on 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. However, we will see with Sobol' indices that sometimes, a single fairness definition can cover both types of fairness.*

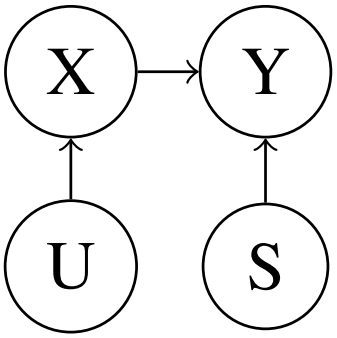
Example 3. *We mentioned above that classical fairness definitions stem back from the Sobol' indices and that these indices can be readily interpreted. Let $Y = f(X, S_1, S_2)$, with S_1 and S_2 two sensitive features.*



(a) Causal generative model in which a protected variable S influence both another variable X and the outcome Y.



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



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

Figure 1: Examples of represention of causal models with directed acyclic graphs.

Table 3: Synthetic experiments based of causal DAGs – Figure 1

	S	ST	S^{ind}	ST^{ind}
$Y = 2 \times X$				
X	1.00	1.00	0.75	0.75
S	0.24	0.25	0.00	0.00
$Y = 0.7 \times X + 0.3 \times S$				
X	0.91	0.92	0.51	0.52
S	0.52	0.54	0.07	0.09
$Y = 0.7 \times X + 0.3 \times S$				
X	0.78	0.84	0.81	0.82
S	0.13	0.17	0.14	0.15

There are four different Sobol' indices associated with (S_1, S_2), as seen in the previous Section. Each of these indices will lead to an intersectionality fairness definition. However, these definitions will not be equivalent. We refer back to Table 1 for the various influences Sobol' indices take or do not take into account. The Total Sobol' index is, out of the four Sobol' indices, the only one to take into account every possible influence coming from (S_1, S_2). It will therefore lead to the more restrictive intersectional fairness.

It is possible to compare some selected classical fairness definitions with some intersectionality fairness definitions.

Proposition 3. *Let (S_1, S_2) be two sensitive features. To be fair in the sense of Avoiding Disparate Treatment for S_1 implies intersectional fairness for (S_1, S_2). However, to be fair in the sense of Disparate Impact for S_1 do not quantify any intersectional fairness.*

Proof. Because of the various bounds on Sobol' indices explained in previous Section, we know that $SobT_{S_1,S_2} \leq SobT_{S_1}$. Since $SobT_{S_1}$ is the GSA measure associated with *Avoiding Disparate Treatment*, we have the first result. 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. □

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 assume a Gaussian model. In each experiment $j, j \in \{1, 2, 3\}$, (X, S, U) are random variables drawn after a

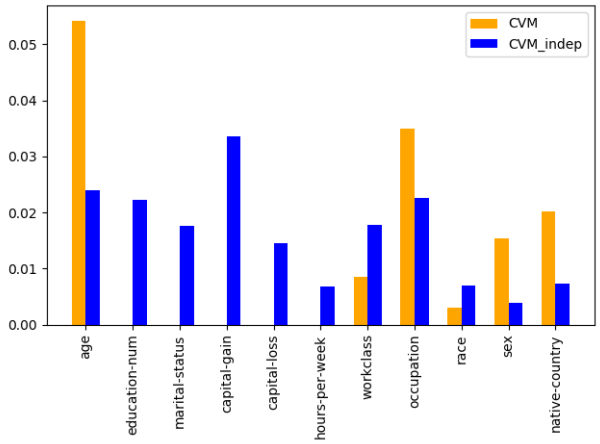


Figure 2: Cramér-von-Mises and independent Cramér-von-Mises indices for the Adult dataset.

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 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 sensible 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 2 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.

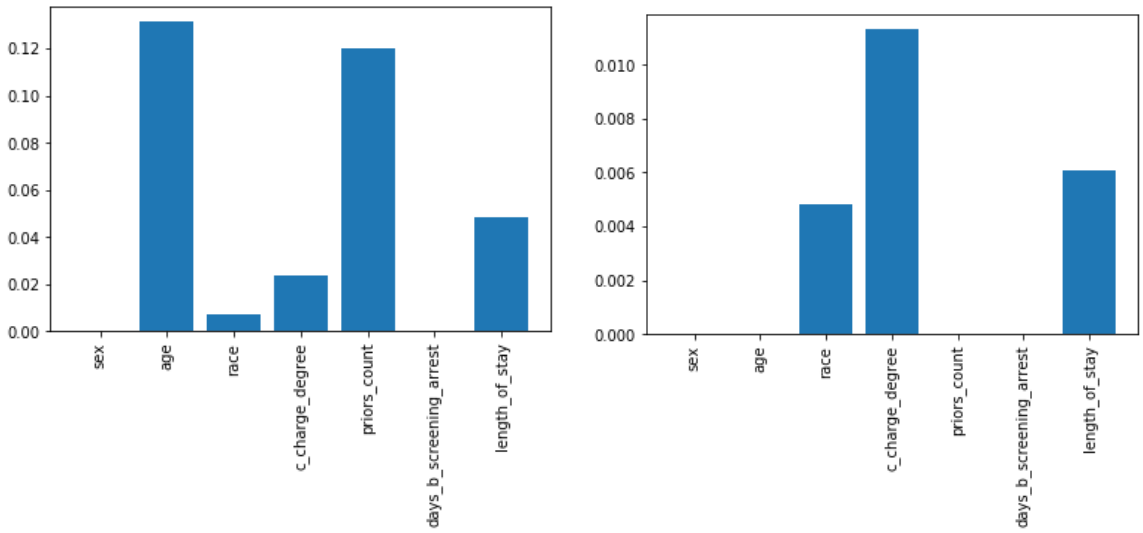
In the third experiment, S and X are independent and S can only influence the outcome directly. This is the framework of classical Global Sensitivity Analysis. In this case, non-independent and independent Sobol' indices are equal, as mentioned in Remark 1

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 [12] 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 and in particular the sensitive attributes mostly considered : gender and country of origin. The results for a classifier obtained for an algorithm built using an Extreme Gradient Boosting Procedure are shown in Figure 2.
First, we point out that we recover the influence of variables, which have already been shown in previous



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

works to be the most important for the explanation of such algorithm. Our two indices provide a different information. The independent Cramér-von-Mises index describes the direct effect of a variable in the outcome of the algorithm without any correlation from other variables. Meanwhile, the Cramér-von-Mises indices include the fact that inputs are not independent. Hence when the variables sex which stands for the gender and native country have CVM high index, this shows the direct discrimination of the algorithmic decision, while the race does not impact the decision. The independent CVM highlights the interactions of the variables on the others and may lead to indirect effect. Here among the sensitive variables, the variable *race* is the one which may affect the most the algorithm through its effect on the other features.

4.2.2 COMPAS dataset

The so-called COMPAS dataset, gathered by ProPublica described for instance in in [38] , 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 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.

References

[1] Mona Azadkia and Sourav Chatterjee. A simple measure of conditional dependence. *arXiv preprint arXiv:1910.12327*, 2019.

[2] Alain Berlinet and Christine Thomas-Agnan. A collection of examples. In *Reproducing Kernel Hilbert Spaces in Probability and Statistics*, pages 293–343. Springer, 2004.

[3] Stephan Bongers, Patrick Forré, Jonas Peters, Bernhard Schölkopf, and Joris M Mooij. Foundations of structural causal models with cycles and latent variables. *arXiv preprint arXiv:1611.06221*, 2020.

[4] Guillaume Carlier, Alfred Galichon, and Filippo Santambrogio. 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.

[5] Sourav Chatterjee. A new coefficient of correlation. *Journal of the American Statistical Association*, pages 1–21, 2020.

[6] Silvia Chiappa, Ray Jiang, Tom Stepleton, Aldo Pacchiano, Heinrich Jiang, and John Aslanides. A general approach to fairness with optimal transport. In *AAAI*, pages 3633–3640, 2020.

[7] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

[8] Evgenii Chzhen, Christophe Denis, Mohamed Hebiri, Luca Oneto, and Massimiliano Pontil. Fair Regression via Plug-in Estimator and Recalibration With Statistical Guarantees. *Advances in Neural Information Processing Systems*, March 2020.

[9] Kimberlé Crenshaw. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *u. Chi. Legal f.*, page 139, 1989.

[10] Sébastien Da Veiga. Global sensitivity analysis with dependence measures. *Journal of Statistical Computation and Simulation*, 85(7):1283–1305, May 2015.

[11] Eustasio del Barrio, Paula Gordaliza, and Jean-Michel Loubes. Review of mathematical frameworks for fairness in machine learning. *arXiv preprint arXiv:2005.13755*, 2020.

[12] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[13] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226. ACM, 2012.

[14] James R Foulds, Rashidul Islam, Kamrun Naher Keya, and Shimei Pan. An intersectional definition of fairness. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, pages 1918–1921. IEEE, 2020.

[15] Christopher Frye, Colin Rowat, and Ilya Feige. Asymmetric shapley values: incorporating causal knowledge into model-agnostic explainability. *Advances in Neural Information Processing Systems*, 33, 2020.

[16] Fabrice Gamboa, Pierre Gremaud, Thierry Klein, and Agnès Lagnoux. Global sensitivity analysis: a new generation of mighty estimators based on rank statistics. *arXiv preprint arXiv:2003.01772*, 2020.

[17] Paula Gordaliza, Eustasio Del Barrio, Gamboa Fabrice, and Jean-Michel Loubes. Obtaining fairness using optimal transport theory. In *International Conference on Machine Learning*, pages 2357–2365, 2019.

[18] Mathilde Grandjacques. *Analyse de sensibilité pour des modèles stochastiques à entrées dépendantes: application en énergétique du bâtiment*. PhD thesis, Grenoble Alpes, 2015.

[19] Vincent Grari, Boris Ruf, Sylvain Lamprier, and Marcin Detyniecki. Fairness-aware neural réyni minimization for continuous features, 2019.

[20] Arthur Gretton, Ralf Herbrich, Alexander Smola, Olivier Bousquet, and Bernhard Schölkopf. Kernel methods for measuring independence. *Journal of Machine Learning Research*, 6(Dec):2075–2129, 2005.

[21] James M. Hickey, Pietro G. Di Stefano, and Vlasios Vasileiou. Fairness by explicability and adversarial shap learning, 2020.

[22] Bertrand Iooss and Paul Lemaître. A review on global sensitivity analysis methods. In *Uncertainty management in simulation-optimization of complex systems*, pages 101–122. Springer, 2015.

[23] Julien Jacques, Christian Lavergne, and Nicolas Devictor. Sensitivity analysis in presence of model uncertainty and correlated inputs. *Reliability Engineering & System Safety*, 91(10-11):1126–1134, 2006.

[24] Noureddine El Karoui Jeremie Mary, Clement Calauzenes. Fairness-aware learning for continuous attributes and treatments, 2019.

[25] Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems*, pages 656–666, 2017.

[26] Thibaut Le Gouic, Jean-Michel Loubes, and Philippe Rigollet. Projection to fairness in statistical learning. *arXiv e-prints*, pages arXiv–2005, 2020.

[27] Paul Lévy. *Théorie de l'addition des variables aléatoires*, volume 1. Gauthier-Villars, 1954.

[28] Thierry A Mara and Stefano Tarantola. Variance-based sensitivity indices for models with dependent inputs. *Reliability Engineering & System Safety*, 107:115–121, 2012.

[29] Jérémie Mary, Clément Calauzènes, and Noureddine El Karoui. Fairness-aware learning for continuous attributes and treatments. In *International Conference on Machine Learning*, pages 4382–4391, 2019.

[30] Giulio Morina, Viktoriia Oliinyk, Julian Waton, Ines Marusic, and Konstantinos Georgatzis. Auditing and achieving intersectional fairness in classification problems. *arXiv preprint arXiv:1911.01468*, 2019.

[31] Luca Oneto and S Chiappa. *Recent Trends in Learning From Data*. Springer, 2020.

[32] Judea Pearl. *Causality*. Cambridge university press, 2009.

[33] Murray Rosenblatt. Remarks on a multivariate transformation. *Ann. Math. Statist.*, 23(3):470–472, 09 1952.

[34] Dominik Rothenhäusler, Nicolai Meinshausen, Peter Bühlmann, and Jonas Peters. Anchor regression: heterogeneous data meets causality. *arXiv preprint arXiv:1801.06229*, 2018.

[35] Alex Smola, Arthur Gretton, Le Song, and Bernhard Schölkopf. A hilbert space embedding for distributions. In *International Conference on Algorithmic Learning Theory*, pages 13–31. Springer, 2007.

[36] Il'ya Meerovich Sobol'. On sensitivity estimation for nonlinear mathematical models. *Matematicheskoe modelirovanie*, 2(1):112–118, 1990.

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

[38] Anne L Washington. How to argue with an algorithm: Lessons from the compas-propublica debate. *Colo. Tech. LJ*, 17:131, 2018.

[39] Robert C Williamson and Aditya Krishna Menon. 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 A.1 (Lévy-Rosemblatt theorem,[27, 33])**.** *: there is a bijection (denoted "RT" for Rosemblatt transform) between $p(\mathbf{X})$ and p independent uniform random variables*

$$(X_i, (X_{i+1}|X_i), \cdots, (X_{i-1}|X_{\sim(i-1)})) \sim p_{\mathbf{X}} \xrightarrow{RT} (U_1^i, \cdots, U_p^i) \sim \mathcal{U}^p(0,1). \tag{10}$$

Example 4. *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

$$z_1 = F_{X_1}^{-1}(u_1^1) \tag{11}$$

$$z_2 = F_{X_2|X_1=x_1}^{-1}(u_2^1). \tag{12}$$

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 8. *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 9. *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 10. *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:

$$Sob_i = \frac{V[\mathbb{E}[g_i(\mathbf{U}^i)|U_1^i]]}{V[g_i(\mathbf{U}^i)]} = \frac{V[\mathbb{E}[f(\mathbf{X})|X_i]]}{V[f(\mathbf{X})]} \tag{13}$$

$$SobT_i = \frac{\mathbb{E}[V[g_i(\mathbf{U}^i)|U_{\sim 1}^i]]}{V[g_i(\mathbf{U}^i)]} = \frac{\mathbb{E}[V[f(\mathbf{X})|Z_i]]}{V[f(\mathbf{X})]} \tag{14}$$

$$Sob_i^{ind} = \frac{V[\mathbb{E}[g_{i+1}(\mathbf{U}^{i+1})|U_p^{i+1}]]}{V[g_{i+1}(\mathbf{U}^{i+1})]} = \frac{V[\mathbb{E}[f(\mathbf{X})|Z_i]]}{V[f(\mathbf{X})]} \tag{15}$$

$$SobT_i^{ind} = \frac{\mathbb{E}[V[g_{i+1}(\mathbf{U}^{i+1})|U_{\sim p}^{i+1}]]}{V[g_{i+1}(\mathbf{U}^{i+1})]} = \frac{\mathbb{E}[V[f(\mathbf{X})|X_{\sim i}]]}{V[f(\mathbf{X})]} \tag{16}$$

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:

$$\begin{aligned}
(u_1^i, \cdots, u_p^i) &\xrightarrow{IRT} \mathbf{x} = (x_i, \cdots, x_{i-1}) \sim p(\mathbf{X}), \\
(u_1^{i\prime}, \cdots, u_p^{i\prime}) &\xrightarrow{IRT} \mathbf{x}' = (x'_i, \cdots, x'_{i-1}) \sim p(\mathbf{X}), \\
(u_1^i, u_2^{i\prime}, \cdots, u_p^{i\prime}) &\xrightarrow{IRT} \mathbf{x}^i = (x_i, x'_{i+1} \cdots, x'_{i-1}) \sim p(X_i)p(X_{\sim i}|X_i), \\
(u_1^{i\prime}, \cdots, u_{p-1}^{i\prime}, u_p^i) &\xrightarrow{IRT} \mathbf{x}^{i-1} = (x'_i, x'_{i+1} \cdots, x_{i-1}) \sim p(X_{\sim i-1})p(X_{i-1}|X_{\sim i-1}).
\end{aligned} \tag{17}$$

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:

$$\begin{aligned}
\widehat{Sob_i} &= \frac{\frac{1}{N}\sum_{k=1}^N f(\mathbf{x}_k) \times \left(f(\mathbf{x}_k^i) - f(\mathbf{x}'_k)\right)}{\hat{V}} \\
\widehat{SobT_i^{ind}} &= \frac{\frac{1}{N}\sum_{k=1}^N \left(f(\mathbf{x}_k^{i-1}) - f(\mathbf{x}'_k)\right)^2}{2\hat{V}} \\
\widehat{Sob_{i-1}^{ind}} &= \frac{\frac{1}{N}\sum_{k=1}^N f(\mathbf{x}_k) \times \left(f(\mathbf{x}_k^{i-1}) - f(\mathbf{x}'_k)\right)}{\hat{V}} \\
\widehat{SobT_i} &= \frac{\frac{1}{N}\sum_{k=1}^N \left(f(\mathbf{x}_k^{i}) - f(\mathbf{x}'_k)\right)^2}{2\hat{V}},
\end{aligned} \tag{18}$$

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 2.1 we presented in Section 2.

Theorem C.1. *Each index $\mathcal{S}$ in the equations* (3) *to* (6) *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:*

$$\sqrt{n}(\mathcal{S}_n - \mathcal{S}) \xrightarrow{D} \mathcal{N}(0, \sigma^2) \tag{19}$$

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

$$Z_n = \begin{pmatrix} n^{-1}\sum f(X_{i,k}, X_{\sim i,k}) f(X_{i,k}, X'_{\sim i,k}) \\ n^{-1}\sum f(X_{i,k}, X_{\sim i,k}) f(X'_{i,k}, X'_{\sim i,k}) \\ n^{-1}\sum f(X_{i,k}, X_{\sim i,k}) \\ n^{-1}\sum f^2(X_{i,k}, X_{\sim i,k}) \end{pmatrix} \tag{20}$$

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 [37], we obtain the convergence of $h(Z_n)$ to $h(Z) = Sob_i$

$$\sqrt{n}\left(\widehat{Sob_i} - Sob_i\right) \to \mathcal{N}(0, \nabla h(\beta) \Sigma \nabla h(\beta)^T), \tag{21}$$

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

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

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

$$\begin{aligned}
\sigma_{11}^2 =& \text{Var}(f(X, X_{\sim i}) f(X, X'_{\sim i})) \\
\sigma_{12}^2 =& \mathbb{E}[f^2(X, X_{\sim i}) f(X, X'_{\sim i}) f(X', X'_{\sim i})] \\
\sigma_{13}^2 =& \mathbb{E}[f^2(X, X_{\sim i}) f(X, X'_{\sim i})] \\
\sigma_{14}^2 =& \mathbb{E}[f^3(X, X_{\sim i}) f(X, X'_{\sim i}) f(X', X'_{\sim i})] - \mathbb{E}[f^2(X, X_{\sim i})]\mathbb{E}[f(X, X'_{\sim i}) f(X, X'_{\sim i})] \\
\sigma_{22}^2 =& \text{Var}(f(X, X_{\sim i}))^2 \\
\sigma_{33}^2 =& \text{Var}(f(X, X_{\sim i})) \\
\sigma_{34}^2 =& \mathbb{E}[f^3(X, X_{\sim i})] \\
\sigma_{44}^2 =& \mathbb{E}[f^4(X, X_{\sim i}) - \mathbb{E}[f^2(X, X_{\sim i})]^2.
\end{aligned} \tag{23}$$

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 (9).

The first one is to use the fact that

$$\begin{aligned}
U(Y, X_i | \mathbf{Z}) &= \frac{\int \mathbb{E}(\text{Var}(\mathbb{E}\left[\mathbb{1}_{Y \leq t} | X_i, \mathbf{Z}\right] | \mathbf{Z})) d\mu(t))}{\int \text{Var}(\mathbb{1}_{Y \leq t}) d\mu(t))} \\
&= T(Y, X_i | \mathbf{Z}) \times (1 - T(Y, \mathbf{Z})).
\end{aligned} \tag{24}$$

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:

$$T_n(Y, X|Z) = \frac{\sum_{i=1}^n \left(\min\{R_i, R_{M(i)}\} - \min\{R_i, R_{N(i)}\}\right)}{\sum_{i=1}^n \left(R_i - \min\{R_i, R_{N(i)}\}\right)}. \tag{25}$$

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

$$U_n(Y, X_i|X_{\sim i}) = T_n(Y, X_i|X_{\sim i}) \times (1 - T_n(Y, X_{\sim i}))\,. \tag{26}$$

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 [5]. We have the two following convergences almost surely:

$$Q_n(Y, X|Z) = n^{-2}\sum_{j=1}^n \left(\min\{R_j, R_{M(j)}\} - \min\{R_j, R_{N(j)}\}\right) \to \int \mathbb{E}(\text{Var}(\mathbb{E}\left[\mathbb{1}_{Y\leq t}|X, Z\right]|Z))d\mu(t)) \tag{27}$$

$$S_n(Y) = n^{-3}\sum_{j=1}^n L_j(n - L_j) \to \int \text{Var}(\mathbb{1}_{Y\leq t})d\mu(t)) \tag{28}$$

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 elementary using classical probability tools.

E Main proofs

E.1 Proof of Theorem A.1

Proof. Indeed, we can always write

$$p_{\mathbf{X}} = p_{X_i} \times p_{X_{i+1}|X_i} \times \cdots \times p_{X_{i-1}|X_{\sim(i-1)}}. \tag{29}$$

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

$$(X_i, (X_{i+1}|X_i), \cdots, (X_{i-1}|X_{\sim(i-1)})) \sim p_{\mathbf{X}} \xrightarrow{RT} (U_1^i, \cdots, U_p^i) \sim \mathcal{U}^p(0, 1). \tag{30}$$

□

E.2 Proof of Proposition 1

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 $\text{Var}(f(\mathbf{X})) = 1$. If we compute the Sobol' index of the predictor $f(\mathbf{X})$ for the protected variable S, we obtain:

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

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 $\text{Var}_S(E_X[f(\mathbf{X})|Y = i, S])$. Since we want this quantity to be equal to zero for each i, we can compute *Equality of Odds* with $\mathbb{E}_Y\text{Var}_S(E_X[f(\mathbf{X})|Y, S])$, 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 $\mathbb{E}_{\mathbf{X}\setminus S}\text{Var}\mathbb{E}_{\mathbf{X}\setminus S}[f(\mathbf{X})|\mathbf{X}]$. This can be simplified into $\mathbb{E}_{\mathbf{X}\setminus S}\text{Var}[f(\mathbf{X})|\mathbf{X} \setminus S]$, which is the Total Sobol' index of the predictor for the protected variable S.

□