Global Sensitivity Analysis: a new generation of mighty estimators based on rank statistics

Fabrice Gamboa, Pierre Gremaud, Thierry Klein, Agnès Lagnoux

Global Sensitivity Analysis: a new generation of mighty estimators based on rank statistics

Fabrice Gamboa[1], Pierre Gremaud[2], Thierry Klein[3], and Agnès Lagnoux[4]

[1]Institut de Mathématiques de Toulouse and ANITI; UMR5219. Université de Toulouse; CNRS. UT3, F-31062 Toulouse, France.
[2]Department of Mathematics. NC State University. Raleigh, North Carolina 27695, USA.
[3]Institut de Mathématiques de Toulouse; UMR5219. Université de Toulouse; ENAC - Ecole Nationale de l'Aviation Civile , Université de Toulouse, France
[4]Institut de Mathématiques de Toulouse; UMR5219. Université de Toulouse; CNRS. UT2J, F-31058 Toulouse, France.

February 27, 2020

Abstract

We propose a new statistical estimation framework for a large family of global sensitivity analysis methods. Our approach is based on rank statistics and uses an empirical correlation coefficient recently introduced by Sourav Chatterjee. We show how to apply this approach to compute not only the Cramér-von-Mises indices, which are directly related to Chatterjee's notion of correlation, but also Sobol indices at any order, higher-order moment indices, and Shapley effects. We establish consistency of the resulting estimators and demonstrate their numerical efficiency, especially for small sample sizes.

Key words Global sensitivity analysis, Cramér-von-Mises distance, Pick-Freeze method, Chatterjee's coefficient of correlation, Shapley effects, Sobol indices estimation.

AMS subject classification 62G05, 62G20.

1 Introduction

The use of complex computer models for the analysis of applications from the sciences, engineering and other fields is by now routine. Often, the models are expensive to run in terms of computational time. It is thus crucial to understand, with just a few runs, the global influence of one or several inputs on the output of the system under study [31]. When these inputs are regarded as random elements, this problem is generally referred to as Global Sensitivity Analysis (GSA). We refer to [9, 30, 34] for an overview of the practical aspects of GSA.

A popular and highly useful tool to quantify input influence are the Sobol indices. These indices were first introduced in [33] and are well tailored to the case of scalar outputs. The Sobol indices compare, thanks to the Hoeffding decomposition [18], the conditional variance of the output knowing some of the input variables to the total variance of the output. Many different estimation procedures of the Sobol indices have been proposed and studied. Some are based on Monte-Carlo or quasi Monte-Carlo design of experiments (see [22, 27] and references therein for more details). In particular, an efficient estimation of the Sobol indices can be performed through the so-called Pick-Freeze method. For the description of this method and its theoretical study (consistency, central limit theorem, concentration inequalities and Berry-Esseen bounds), we refer to [20, 13] and references therein. Some other estimation procedures are based on different designs of experiment using for example polynomial chaos (see [36] and the reference therein for more details).

Various generalizations of the Sobol indices have been developed. The issue of vectorial outputs, as is the case with time dependent or functional quantities of interest, is addressed in [1, 12, 23]. In particular, in [12], the authors recover the indices from [23] and show that they are a proper generalization of the classical Sobol indices in higher dimension. Moreover, they provide the theoretical study of their Pick-Freeze estimators and extend their definitions to the case of outputs valued in a separable Hilbert space.

Since Sobol indices are variance based, they only quantify the second order influence of the inputs. Many authors proposed other criteria to compare the conditional distribution of the output knowing some of the inputs to the distribution of the output. In [27, 26, 25], the authors use higher moments to define new indices while, in [5, 6, 8], the use of divergences or distances between measures allows to define new indices. In [10], the authors use contrast functions to build indices that are goal oriented. Although these works define nice theoretical indices, the existence of a relevant statistical estimation procedure is still in most cases an open question. The case of vectorial valued computer codes is considered in [14] where a sensitivity index based on the whole distributions

is defined. Within this framework, the authors show that the Pick-Freeze estimation procedure provides an asymptotically Gaussian estimator of the index. The scheme requires $3N$ evaluations of the output code for the evaluation of a single index and leads to a convergence rate $\sqrt{N}$. Hence, if the number of inputs variable is p, the total number of calls of the code is $(p+3)N$ that grows linearly with p. This approach has been generalized in [11], where the authors considered computer codes valued on a compact Riemannian manifold. They use the Pick-Freeze scheme to provide a consistent estimator requiring $4N$ evaluations of the output code. The authors of [15] extend the previous indices to general metric spaces and propose U-statistics-based estimators improving the classical Pick-Freeze procedure.

We emphasize that the Pick-Freeze estimation procedure allows the estimation of several sensitivity indices: the classical Sobol indices for real-valued outputs, as well as their generalization for vectorial valued codes, but also the indices based on higher moments [26] and the Cramér-von-Mises indices which take into account on the whole distribution [14, 11]. In addition, the Pick-Freeze estimators have desirable statistical properties such as consistency, fixed rate of convergence and exponential inequalities. They have, however, two majors drawbacks. First, they rely on a particular experimental design that may be unavailable in practice. Second, the number of model calls to estimate all first-order Sobol indices grows linearly with the number of input parameters. For example, if we consider $p = 99$ input parameters and only $n = 1000$ calls are allowed, then only a sample of size $n/(p + 1) = 10$ is available to estimate each single first-order Sobol index.

In a recent work [7], Chatterjee studies the dependence between two variables by introducing an empirical correlation coefficient based on rank statistics, see Section 2.3 below for the precise definition. The striking point of his work is that this empirical correlation coefficient converges almost surely to the Cramér-von-Mises index introduced in [14] as the sample size goes to infinity. In this paper, we show how to embed Chatterjee's method in the GSA framework, thereby eliminating the two drawbacks of the classical Pick-Freeze estimation mentioned above. In addition, we generalize Chatterjee's approach to allow the estimation a large class of GSA indices which include the Sobol indices and the higher order moment indices proposed by Owen [27, 26, 25]. Using a single sample of size n, it is now possible to estimate at the same time all the Sobol indices at any order, the Cramér-von-Mises indices, and other useful sensitivity indices. Last but not least, this estimation scheme also allows to estimate the Shapley effects defined in [28] for correlated inputs.

The paper is organized as follows. In Section 2, we recall the definition of the Cramér-von-Mises indices and their classical Pick-Freeze estimation. Further,

we show how they can be also estimated using Chatterjee's method. In Section 3, we present the generalization Chatterjee's method to estimate other sensitivity indices and in particular the Sobol indices. Extensions are presented in Section 4. The higher order Sobol indices and the Shapley effects are considered. Section 5 is dedicated to a numerical comparison between the Pick-Freeze estimation procedure and Chatterjee's method. We first compare the numerical performance of both estimators on a linear model. Finally, we consider a real life application. As expected, Chaterjee's estimation method outperforms the classical Pick-Freeze procedure, even for small sample sizes (which are common in practice). Conclusions and perspectives are offered in Section 6.

2 Sensitivity analysis based on Cramér-von-Mises indices

2.1 Definition of Cramér-von-Mises indices

The quantity of interest (QoI) Y is obtained from the numerical code and is regarded a function f of the vector of the distributed input $(X_i)_{i=1,\cdots,p}$

$$Y = f(X_1, \ldots, X_p), \tag{1}$$

where f is defined on the state space $E_1 \ldots \times E_p$, $X_i \in E_i$, $i = 1, \ldots, p$. Classically, the X_i's are assumed to be independent random variables and a sensitivity analysis is performed using the Hoeffding decomposition [2, 37] leading to the standard Sobol indices [34]. More precisely, assume f to be real-valued and square integrable and let $\mathbf{u}$ be a subset of $\{1, \ldots, p\}$ and $\sim\mathbf{u}$ its complementary set in $\{1, \ldots, p\}$. Setting $X_{\mathbf{u}} = (X_i, i \in \mathbf{u})$ and $X_{\sim\mathbf{u}} = (X_i, i \in \sim\mathbf{u})$, the corresponding Sobol indices take the form

$$S^{\mathbf{u}} = \frac{\mathrm{Var}\left(\mathbb{E}[Y|X_{\mathbf{u}}]\right)}{\mathrm{Var}(Y)} \quad \text{and} \quad S^{\sim\mathbf{u}} = \frac{\mathrm{Var}\left(\mathbb{E}[Y|X_{\sim\mathbf{u}}]\right)}{\mathrm{Var}(Y)}. \tag{2}$$

By definition, the Sobol indices quantifies the fluctuations of the output Y around its mean. When the practitioner is not interested in the mean behavior of Y but rather in its median, in its tail, or even in its quantiles, the Sobol indices become less appropriate to quantify sensitivity. GSA must then be performed in a framework which takes into account more than one specific moment, such as the variance for Sobol indices. The Cramér-von-Mises indices introduced in [14] provide alternative indices based on the whole distribution.

They are defined by

$$S^{\mathbf{u}}_{2,CVM} = \frac{\int_{\mathbb{R}} \mathbb{E}\left[\left(F(t) - F^{\mathbf{u}}(t)\right)^2\right] dF(t)}{\int_{\mathbb{R}} F(t)(1-F(t))dF(t)} \tag{3}$$

where F is the cumulative distribution function of Y

$$F(t) = \mathbb{P}\left(Y \leqslant t\right) = \mathbb{E}\left[\mathbb{1}_{\{Y \leqslant t\}}\right] \quad (t \in \mathbb{R})$$

and $F^{\mathbf{u}}$ is its Pick-Freeze version, namely the conditional distribution function of Y conditionally on $X_{\mathbf{u}}$:

$$F^{\mathbf{u}}(t) = \mathbb{P}\left(Y \leqslant t | X_{\mathbf{u}}\right) = \mathbb{E}\Big[\mathbb{1}_{\{Y \leqslant t\}} | X_{\mathbf{u}}\Big] \quad (t \in \mathbb{R}).$$

Such a definition stems from the Hoeffding decomposition of the collection of variables $(\mathbb{1}_{\{Y \leqslant t\}})_{t \in \mathbb{R}}$. It is worth noting that this definition naturally extends to multivariate outputs.

2.2 Classical estimation of Sobol and Cramér-von-Mises indices using the Pick-Freeze method

The estimation of the Cramér-von-Mises index (3) reduces to the estimation of both its numerator and its denominator. The numerator of $S^{\mathbf{u}}_{2,CVM}$ can be rewritten as

$$\begin{aligned}\int_{\mathbb{R}} \mathbb{E}\left[\left(F(t) - F^{\mathbf{u}}(t)\right)^2\right] dF(t) &= \mathbb{E}_W\left[\mathbb{E}_{X_{\mathbf{u}}}\left[\left(F(W) - F^{\mathbf{u}}(W)\right)^2\right]\right] \\ &= \mathbb{E}_W\left[\mathrm{Var}_{X_{\mathbf{u}}}\left(\mathbb{E}_Y\left[\mathbb{1}_{\{Y \leqslant W\}} | X_{\mathbf{u}}\right]\right)\right]\end{aligned}$$

where W is an independent copy of Y and where, for a random quantity Z, $\mathbb{E}_Z$ and Var_Z denote respectively the expectation and the variance with respect to Z. When no confusion is possible, we only write $\mathbb{E}$ and Var in the rest of the paper. A Monte-Carlo scheme can be used to estimate the Cramér-von-Mises indices. The corresponding Pick-Freeze approach from [13, 14, 20] relies on expressing the variances of the conditional expectations in terms of covariances which are easily and well estimated by their empirical versions. To that end, we define, for any subset $\mathbf{u}$ of $\{1, \ldots, p\}$

$$Y^{\mathbf{u}} := f(X^{\mathbf{u}}). \tag{4}$$

where $X^{\mathbf{u}}$ is such that $X^{\mathbf{u}}_{\mathbf{u}} = X_{\mathbf{u}}$ and $X^{\mathbf{u}}_i = X'_i$ if $i \in\!\!\sim \mathbf{u}$, X'_i being an independent copy of X_i. The estimation procedure relies on the following lemma which is still valid for any function $g \in L^2$, not just $g(y) = \mathbb{1}_{\{y \leqslant t\}}$.

Lemma 2.1.

$$\mathrm{Var}(\mathbb{E}[\mathbb{1}_{\{Y\leqslant t\}}|X_{\boldsymbol{u}}]) = \mathrm{Cov}(\mathbb{1}_{\{Y\leqslant t\}}, \mathbb{1}_{\{Y^{\boldsymbol{u}}\leqslant t\}}). \tag{5}$$

Proof. Let $Z = \mathbb{1}_{\{Y\leqslant t\}}$ and $Z^{\mathbf{u}} = \mathbb{1}_{\{Y^{\mathbf{u}}\leqslant t\}}$. Since, Z and $Z^{\mathbf{u}}$ shares the same distribution and are independent conditionally to $X_{\mathbf{u}}$, we have

$$\begin{aligned}
\mathrm{Var}(\mathbb{E}[Z|X_{\mathbf{u}}]) &= \mathbb{E}[\mathbb{E}[Z|X_{\mathbf{u}}]^2] - \mathbb{E}[\mathbb{E}[Z|X_{\mathbf{u}}]]^2 \\
&= \mathbb{E}[\mathbb{E}[Z|X_{\mathbf{u}}]\mathbb{E}[Z^{\mathbf{u}}|X_{\mathbf{u}}]] - \mathbb{E}[\mathbb{E}[Z|X_{\mathbf{u}}]]\mathbb{E}[\mathbb{E}[Z^{\mathbf{u}}|X_{\mathbf{u}}]] \\
&= \mathbb{E}[\mathbb{E}[ZZ^{\mathbf{u}}|X_{\mathbf{u}}]] - \mathbb{E}[Z]\mathbb{E}[Z^{\mathbf{u}}] \\
&= \mathbb{E}[ZZ^{\mathbf{u}}] - \mathbb{E}[Z]\mathbb{E}[Z^{\mathbf{u}}] \\
&= \mathrm{Cov}(Z, Z^{\mathbf{u}}).
\end{aligned}$$

□

Consequently, the Monte-Carlo estimation can be done as follows. A n sample $(Y_1, \ldots, Y_n)$ of the output Y and a n sample $(Y_1^{\mathbf{u}}, \ldots, Y_n^{\mathbf{u}})$ of its Pick-Freeze version $Y^{\mathbf{u}}$ are required. In addition, in order to deal with the integral with respect to $dF(t)$ in (3), a third independent n sample $(W_1, \ldots, W_n)$ of the output Y is necessary. Then the empirical estimator of $S^1_{2,CVM}$ is

$$S^1_{n,2,CVM} = \frac{\frac{1}{n}\sum_{k=1}^n \left(\frac{1}{n}\sum_{j=1}^n \mathbb{1}_{\{Y_j\leqslant W_k\}}\mathbb{1}_{\{Y_j^{\mathbf{u}}\leqslant W_k\}} - \frac{1}{n}\sum_{j=1}^n \mathbb{1}_{\{Y_j\leqslant W_k\}}\frac{1}{n}\sum_{j=1}^n \mathbb{1}_{\{Y_j^{\mathbf{u}}\leqslant W_k\}}\right)}{\frac{1}{n}\sum_{j=1}^n \left(\frac{1}{n}\sum_{j=1}^n \mathbb{1}_{\{Y_j\leqslant W_k\}} - \left(\frac{1}{n}\sum_{j=1}^n \mathbb{1}_{\{Y_j\leqslant W_k\}}\right)^2\right)}. \tag{6}$$

As showed in [14], this estimator is consistent and asymptotically Gaussian (i.e. the rate of convergence is $\sqrt{n}$). The limiting variances can be computed explicitly, allowing the practitioner to build confidence intervals. In particular, if one wants to estimate all the first-order indices (that is the p first-order Sobol indices) and the p Cramér-von-Mises indices, $(p+2)n$ calls of the computer code are required. The number of calls grows linearly with respect to the number of input parameters. This is a practical issue for large input dimension domains. A second drawback of this estimation scheme comes from the need of the particular Pick-Freeze design that is not always available.

2.3 Chatterjee's method

In [7], Chatterjee considers a pair of random variables (V, Y) and an i.i.d. sample $(V_j, Y_j)_{1\leqslant j\leqslant n}$. In order to simplify the presentation, we assume that the laws

of V and Y are both diffuse (ties are excluded). The pairs $(V_{(1)}, Y_{(1)}), \ldots, (V_{(n)}, Y_{(n)})$ are rearranged in such a way that

$$V_{(1)} < \ldots < V_{(n)}.$$

Let r_j be the rank of $Y_{(j)}$, that is,

$$r_j = \#\{j' \in \{1, \ldots, n\},\ Y_{(j')} \leqslant Y_{(j)}\}.$$

The new correlation coefficient defined by Chatterjee in [7] is

$$\xi_n(V, Y) := 1 - \frac{3 \sum_{j=1}^{n-1} |r_{j+1} - r_j|}{n^2 - 1}. \tag{7}$$

The author proves that $\xi_n(V, Y)$ converges almost surely to a deterministic limit $\xi(V, Y)$ which is equal to the Cramér-von-Mises sensitivity index $S^V_{2,CVM}$ with respect to V as soon as V is one of the random variables X_1, ..., X_p in the model (1). Further, he also proves a central limit theorem (CLT) when V and Y are independent.

Chatterjee also provides a rank statistics analog to Lemma 2.1. More precisely, let $\pi(j)$ be the rank of V_j in the sample $(V_1, \ldots, V_n)$ of V and define

$$N(j) = \begin{cases} \pi^{-1}(\pi(j) + 1) & \text{if } \pi(j) + 1 \leqslant n, \\ \pi^{-1}(1) & \text{if } \pi(j) = n. \end{cases} \tag{8}$$

Observe that $\xi_n(V, Y)$ can be rewritten as Q_n / S_n where

$$\begin{aligned} Q_n &= \frac{1}{n} \sum_{j=1}^{n} \min\{F_n(Y_j), F_n(Y_{N(j)})\} - \frac{1}{n} \sum_{j=1}^{n} F_n(Y_j)^2 \\ &= \frac{1}{n} \sum_{j=1}^{n} \left(\frac{1}{n} \sum_{k=1}^{n} \mathbb{1}_{Y_k \leqslant Y_j} \mathbb{1}_{Y_k \leqslant Y_{N(j)}} - \left(\frac{1}{n} \sum_{k=1}^{n} \mathbb{1}_{Y_k \leqslant Y_j} \right)^2 \right), \\ S_n &= \frac{1}{n} \sum_{j=1}^{n} F_n(Y_j)(1 - F_n(Y_j)), \end{aligned}$$

where F_n stands for the empirical distribution function of Y: $F_n(t) = \frac{1}{n} \sum_{j=1}^{n} \mathbb{1}_{\{Y_j \leqslant t\}}$. The analogue of the Pick-Freeze version Y^V with respect to V of Y becomes Y_N and Lemma 2.1 is replaced by the formula

$$\mathbb{E}[\mathbb{1}_{\{Y_j \geqslant t\}} \mathbb{1}_{\{Y_{N(j)} \geqslant t\}} | V_1, \ldots, V_n] = G_{V_j}(t) G_{V_{N(j)}}(t) \tag{9}$$

for all $j = 1, \ldots, n$ that is mentioned in the proof of Lemma 7.10 in [7, p.24], with G_V the conditional survival function: $G_V(t) = \mathbb{P}(Y \geqslant t | V)$.

Remark 2.2. In [7], the author considers also the random variables $V_{n,j}$ due to the fact that ties are possible. In our paper, we assume that the distributions of V and Y are diffuse rendering the introduction of the $V_{n,j}$'s unworthy since in this case, $V_{n,j} = V_{N(j)}$.

Consequences of Chatterjee's method

1. A unique n sample of input-output provides consistent estimations of the p first-order Cramér-von-Mises indices.

2. Chatterjee's central limit theorem allows to built statistical tests for testing
$$H_0 : S^V_{2,CVM} = 0 \quad \text{against} \quad H_1 : S^V_{2,CVM} \neq 0.$$

3 Generalization of Chatterjee's method

3.1 A universal estimation procedure of sensitivity indices

In this section, we propose a universal estimation procedure of expectations of the form
$$\mathbb{E}[\mathbb{E}[f(Y)|V]\mathbb{E}[g(Y)|V]].$$
This result is a generalization of (9) and can be interpreted as an approximation of (5). To this end, we introduce the function Ψ_V defined by
$$\Psi_V(f) = \mathbb{E}[f(Y)|V] \tag{10}$$
for any integrable function f. Let $\mathcal{F}_n$ be the σ-algebra generated by $\{V_1, \ldots, V_n\}$. Note that in Section 2.3, we consider $f(x) = f_t(x) = \mathbb{1}_{x \geqslant t}$ so that $\Psi_V(f) = \mathbb{P}(Y \geqslant t|V) = G_V(t)$.

Lemma 3.1. *Let f and g be two integrable functions such that fg is also integrable. Let $(V_j, Y_j)_{1 \leqslant j \leqslant n}$ be a n sample of (V, Y). Consider a $\mathcal{F}_n$-measurable random permutation σ_n such that $\sigma_n(j) \neq j$, for all $j = 1, \ldots, n$. Then*
$$\mathbb{E}\left[f(Y_j)g(Y_{\sigma_n(j)})|V_1, \ldots, V_n\right] = \Psi_{V_j}(f)\Psi_{V_{\sigma_n(j)}}(g). \tag{11}$$

Proof. Using the measurability of σ_n and by independence, we have

$$\begin{aligned}
\mathbb{E}\left[f(Y_j)g(Y_{\sigma_n(j)})|\mathcal{F}_n\right] &= \mathbb{E}\Big[f(Y_j)\sum_{\substack{l=1,\\ l\neq j}}^{n} g(Y_l)\mathbb{1}_{\sigma_n(j)=l}|\mathcal{F}_n\Big] = \sum_{\substack{l=1,\\ l\neq j}}^{n} \mathbb{1}_{\sigma_n(j)=l}\mathbb{E}\Big[f(Y_j)g(Y_l)|\mathcal{F}_n\Big] \\
&= \sum_{\substack{l=1,\\ l\neq j}}^{n} \mathbb{1}_{\sigma_n(j)=l}\mathbb{E}\Big[f(Y_j)|\mathcal{F}_n\Big]\mathbb{E}\Big[g(Y_l)|\mathcal{F}_n\Big] \\
&= \mathbb{E}\Big[f(Y_j)|V_j\Big]\sum_{\substack{l=1,\\ l\neq j}}^{n} \mathbb{1}_{\sigma_n(j)=l}\mathbb{E}\Big[g(Y_l)|V_l\Big] \\
&= \Psi_{V_j}(f)\sum_{\substack{l=1,\\ l\neq j}}^{n} \mathbb{1}_{\sigma_n(j)=l}\Psi_{V_l}(g) = \Psi_{V_j}(f)\Psi_{V_{\sigma_n(j)}}(g).
\end{aligned}$$

□

The previous lemma leads to a generalization of the first part of the numerator of ξ_n defined in (7). Following the same lines as in [7], one may prove that such a quantity converges almost surely as $n \to \infty$ under some mild conditions.

Proposition 3.2. *Let f and g be two bounded measurable functions. Consider a $\mathcal{F}_n$-measurable random permutation σ_n with no fix point (i.e. $\sigma_n(j) \neq j$), for all $j = 1, \ldots, n$. In addition, we assume that for any $j = 1, \ldots, n$, $V_{\sigma_n(j)} \to V_j$ as $n \to \infty$ with probability one. Then $\chi_n(V, Y; f, g)$ defined by*

$$\chi_n(V, Y; f, g) = \frac{1}{n}\sum_{j=1}^{n} f(Y_j)g(Y_{\sigma_n(j)}) \tag{12}$$

converges almost surely as $n \to \infty$ to

$$\chi(V, Y; f, g) = \mathbb{E}[\Psi_V(f)\Psi_V(g)], \tag{13}$$

where $\Psi_V(f)$ has been defined in (10).

Proof. We follow the steps of the proof of Corollary 7.12 in [7]. Our proof is significantly simpler since σ_n is assumed to have no fix points and V is continuous so that the are no ties in the sample. To simplify the notation, we denote $\chi_n(V, Y; f, g)$ and $\chi(V, Y; f, g)$ by χ_n and χ respectively.
We first prove that, for any measurable function h,

$$h(V_1) - h(V_{\sigma_n(1)}) \to 0 \tag{14}$$

almost surely as $n \to \infty$. Let $\varepsilon > 0$. By the special case of Lusin' theorem (see [7, Lemma 7.5]), there exists a compactly supported continuous function $g\colon \mathbb{R} \to \mathbb{R}$ such that $\mathbb{P}(\{x;\, h(x) \neq g(x)\}) < \varepsilon$, where $\mathbb{P}$ stands for the distribution of V. Then for any $\delta > 0$,

$$\mathbb{P}\Big(\Big|h(V_1) - h(V_{\sigma_n(1)}\Big| > \delta) \leqslant \mathbb{P}\Big(\Big|g(V_1) - g(V_{\sigma_n(1)})\Big| > \delta\Big) + \mathbb{P}(h(V_1) \neq g(V_1)) + \mathbb{P}(h(V_{\sigma_n(1)}) \neq g(V_{\sigma_n(1)})). \quad (15)$$

By continuity of g and since $V_{\sigma_n(1)} \to V_1$ as $n \to \infty$ with probability one, the first term in the right hand side of (15) converges to 0 as $n \to \infty$. By construction of g, the second term is lower than ε. Turning to the third one, we have thus

$$\begin{aligned}
\mathbb{E}[h(V_{\sigma_n(1)})] &= \frac{1}{n}\sum_{j=1}^{n} \mathbb{E}[h(V_{\sigma_n(j)})] = \frac{1}{n}\sum_{j=1}^{n}\sum_{\substack{l=1\\ l\neq j}}^{n} \mathbb{E}[h(V_l)\mathbb{1}_{\sigma_n(j)=l}] \\
&= \frac{1}{n}\sum_{l=1}^{n}\sum_{\substack{j=1\\ j\neq l}}^{n} \mathbb{E}[h(V_l)\mathbb{1}_{\sigma_n(j)=l}] = \frac{1}{n}\sum_{l=1}^{n} \mathbb{E}[h(V_l)\sum_{\substack{j=1\\ j\neq l}}^{n} \mathbb{1}_{\sigma_n(j)=l}] \\
&= \frac{1}{n}\sum_{l=1}^{n} \mathbb{E}[h(V_l)] = \mathbb{E}[h(V_1)]
\end{aligned}$$

where we have used the fact that by assumption σ_n has no fix point and the V_i's have no ties. This yields

$$\mathbb{P}(h(V_{\sigma_n(1)}) \neq g(V_{\sigma_n(1)})) = \mathbb{P}(h(V_1) \neq g(V_1)) < \varepsilon,$$

and, since ε and δ are arbitrary, (14) is therefore proved.
Now, since $x \mapsto \Psi_x$ is a measurable function and applying (14), we have

$$\begin{cases} \Psi_{V_1}(f) - \Psi_{V_{\sigma_n(1)}}(f) & \to 0, \\ \Psi_{V_1}(g) - \Psi_{V_{\sigma_n(1)}}(g) & \to 0, \end{cases} \quad \text{in probability as } n \to \infty. \quad (16)$$

Lemma 3.5 and the dominated convergence theorem lead to

$$\begin{aligned}
\mathbb{E}[\chi_n] &= \frac{1}{n}\sum_{j=1}^{n} \mathbb{E}[f(Y_j)g(Y_{\sigma_n(j)})] = \mathbb{E}[f(Y_1)g(Y_{\sigma_n(1)})] \\
&= \mathbb{E}[\Psi_{V_1}(f)\Psi_{V_{\sigma_n(1)}}(g)] \to \mathbb{E}[\Psi_V(f)\Psi_V(g)] = \chi(V, Y; f, g) \quad (17)
\end{aligned}$$

where we have taken into account the fact that $\Psi_V(f)$ and $\Psi_V(g)$ are bounded (due to the boundedness of f and g) and used (16).
The last step of the proof consists in comparing $\mathbb{E}[\chi_n]$ with χ_n using Mc Diarmid's concentration inequality [24]. To be self-contained, we recall this result.

Theorem 3.3 (Mac Diarmid's bounded difference concentration inequality [24]). *Let $W = (W_1, \dots, W_n)$ be a family of independent variables with W_i taking its values in a set A_k. Consider a real-valued function h defined on $\Pi_{k=1}^n A_k$ satisfying*

$$|h(w) - h(w')| \leqslant c_k \tag{18}$$

as soon as the vectors w and w' differ only on the k-th coordinate. Then we have, for any $t > 0$,

$$\mathbb{P}\left(|h(W) - \mathbb{E}[h(W)]| \geqslant t\right) \leqslant 2\exp\left(-\frac{2t^2}{\sum_{k=1}^n c_k}\right).$$

Assume that for some $i \leqslant n$, the pair (V_i, Y_i) is replaced by a different value (V_i', Y_i'). Then there are at most three indices j such that the value of $\sigma_n(j)$ changes after such a replacement, and exactly one index, $j = i$, where Y_j changes. Moreover, there can be at most one index j such that $\sigma_n(j) = i$, both before and after the replacement. Using the boundedness of f and g, this shows that χ_n changes by at most C/n due to this replacement.

Theorem 3.3 then implies

$$\mathbb{P}(|\chi_n - \mathbb{E}[\chi_n]| \geqslant t) \leqslant 2e^{-2n^2t^2/C^2}, \tag{19}$$

and we conclude the proof by combining (17) and (19). □

3.2 Recovering the classical first-order Sobol indices

We can now leverage the above results and construct a new family of estimators for Sobol indices. Indeed, assume we want to estimate the Sobol index with respect to $V = X_1$. We then define N as in (8) where π is the rank of X_1. Taking $f(x) = g(x) = x$ and $\sigma_n = N$, (11) provides the analogue to ξ_n to estimate the classical Sobol indices:

$$\xi_n^{\text{Sobol}}(X_1, Y) := \frac{\frac{1}{n}\sum_{j=1}^n Y_j Y_{N(j)} - \left(\frac{1}{n}\sum_{j=1}^n Y_j\right)^2}{\frac{1}{n}\sum_{j=1}^n (Y_j)^2 - \left(\frac{1}{n}\sum_{j=1}^n Y_j\right)^2}, \tag{20}$$

where the denominator reduces to the empirical variance of Y.
This estimator can be compared to the classical Pick-Freeze estimators which are constructed as follows. For the estimation of S^1 for instance, a n sample

$(Y^1, \dots, Y^n)$ of the output Y and a n sample $(Y_k^1, \dots, Y_k^n)$ of its Pick-Freeze version Y_k are required. The natural estimator of S^1 is then given by

$$S_n^1 = \frac{\frac{1}{n}\sum_{j=1}^n Y_j Y_j^{\mathbf{1}} - \left(\frac{1}{n}\sum_{j=1}^n Y_j\right)\left(\frac{1}{n}\sum_{j=1}^n Y_j^{\mathbf{1}}\right)}{\frac{1}{n}\sum_{j=1}^n (Y_j)^2 - \left(\frac{1}{n}\sum_{j=1}^n Y_j\right)^2}. \tag{21}$$

A slightly different estimator is introduced in [20] to use all the information available:

$$T_n^1 = \frac{\frac{1}{n}\sum_{j=1}^n Y_j Y_j^{\mathbf{1}} - \left(\frac{1}{n}\sum_{j=1}^n \frac{Y_j + Y_j^{\mathbf{1}}}{2}\right)^2}{\frac{1}{n}\sum_{j=1}^n \frac{(Y_j)^2 + (Y_j^{\mathbf{1}})^2}{2} - \left(\frac{1}{n}\sum_{j=1}^n \frac{Y_j + Y_j^{\mathbf{1}}}{2}\right)^2}. \tag{22}$$

As for the Cramér-von-Mises estimation scheme, such an estimation procedure has been proved to be consistent and asymptotically normal (i.e. the rate of convergence is $\sqrt{n}$) in [20, 13]. The limiting variances can be computed explicitly, allowing the practitioner to build confidence intervals. In addition, the sequence of estimators $(T_n^1)_n$ is asymptotically efficient to estimate S^1 from such a design of experiment (see, [37] for the definition of the asymptotic efficiency and [13] for the details of the result).

3.3 Sensitivity indices in general metric spaces

In this section, we consider a computer code valued in a general metric space $\mathcal{X}$ as presented in [15]. In this context, the authors of [15] consider a family of test functions parametrized by m elements of $\mathcal{X}$ ($m \in \mathbb{N}^*$). For any $a = (a_i)_{i=1,\dots,m} \in \mathcal{X}^m$, the test functions

$$\begin{array}{ccc} \mathcal{X}^m \times \mathcal{X} & \to & \mathbb{R} \\ (a, x) & \mapsto & T_a(x) \end{array}$$

are assumed to be L^2-functions with respect to the product measure $\mathbb{P}^{\otimes m} \otimes \mathbb{P}$ on $\mathcal{X}^m \times \mathcal{X}$ where $\mathbb{P}$ is the distribution of Y. Then they define the general metric space sensitivity index with respect to X_1 by

$$S_{2,GMS}^{\mathbf{1}} := \frac{\int_{\mathcal{X}^m} \mathbb{E}\left[\left(\mathbb{E}[T_a(Y)] - \mathbb{E}[T_a(Y)|X_1]\right)^2\right] d\mathbb{P}^{\otimes m}(a)}{\int_{\mathcal{X}^m} \mathrm{Var}(T_a(Y)) d\mathbb{P}^{\otimes m}(a)}. \tag{23}$$

This general class of indices encompasses the classical sensitivity indices, for instance, the Sobol indices and the Cramér-von-Mises indices. Naturally, a Monte-Carlo procedure based on the Pick-Freeze scheme can be performed to estimate $S_{2,GMS}^{\mathbf{1}}$.

Estimation procedure based on U-statistics In [15], the authors propose a more efficient estimation procedure based on U-statistics (see [15, Equation (13)]). More precisely, for any $1 \leqslant i \leqslant m+2$, we let $\mathbf{y}_i = (y_i, y_i^{\mathbf{1}})$ and we define

$$\begin{aligned}
\Phi_1(\mathbf{y}_1, \dots, \mathbf{y}_{m+1}) &:= T_{y_1,\dots,y_m}(y_{m+1}) T_{y_1,\dots,y_m}(y_{m+1}^{\mathbf{1}}) \\
\Phi_2(\mathbf{y}_1, \dots, \mathbf{y}_{m+2}) &:= T_{y_1,\dots,y_m}(y_{m+1}) T_{y_1,\dots,y_m}(y_{m+2}^{\mathbf{1}}) \\
\Phi_3(\mathbf{y}_1, \dots, \mathbf{y}_{m+1}) &:= T_{y_1,\dots,y_m}(y_{m+1})^2 \\
\Phi_4(\mathbf{y}_1, \dots, \mathbf{y}_{m+2}) &:= T_{y_1,\dots,y_m}(y_{m+1}) T_{y_1,\dots,y_m}(y_{m+2}).
\end{aligned}$$

We set

$$m(1) = m(3) = m+1 \quad \text{and} \quad m(2) = m(4) = m+2 \tag{24}$$

and we define for $j = 1, \dots, 4$,

$$I(\Phi_j) := \int_{\mathcal{X}^{m(j)}} \Phi_j(\mathbf{y}_1, \dots, \mathbf{y}_{m(j)}) d\mathbb{P}_{\mathbf{Y}}^{\otimes m(j)}(\mathbf{y}_1 \dots, \mathbf{y}_{m(j)}), \tag{25}$$

where $\mathbb{P}_{\mathbf{Y}}$ stands for the law of $\mathbf{Y} = (Y, Y^{\mathbf{1}})^\top$. Finally, we introduce the application Ψ from $\mathbb{R}^4$ to $\mathbb{R}$ defined by

$$\begin{array}{rccc} \psi: & \mathbb{R}^4 & \to & \mathbb{R} \\ & (x,y,z,t) & \mapsto & \frac{x-y}{z-t}. \end{array} \tag{26}$$

Then one can express $S^{\mathbf{1}}_{2,GMS}$ in the following way

$$S^{\mathbf{1}}_{2,GMS} = \psi\left(I(\Phi_1), I(\Phi_2), I(\Phi_3), I(\Phi_4)\right). \tag{27}$$

Following the framework of Hoeffding [18], we replace the functions Φ_1, Φ_2, Φ_3 and Φ_4 by their symmetrized version Φ_1^s, Φ_2^s, Φ_3^s and Φ_4^s:

$$\Phi_j^s(\mathbf{y}_1, \dots, \mathbf{y}_{m(j)}) = \frac{1}{(m(j))!} \sum_{\tau \in \mathcal{S}_{m(j)}} \Phi_j(\mathbf{y}_{\tau(1)}, \dots, \mathbf{y}_{\tau(m(j))})$$

for $j = 1, \dots, 4$ where $\mathcal{S}_k$ is the symmetric group of order k. For $j = 1, \dots 4$, the integrals $I(\Phi_j^s)$ are naturally estimated by U-statistics of order $m(j)$. More precisely, we consider a n i.i.d. sample $(\mathbf{Y}_1, \dots, \mathbf{Y}_n)$ with distribution $\mathbb{P}_{\mathbf{Y}}$ and, for $j = 1, \dots, 4$, we define

$$U_{j,n} := \binom{n}{m(j)}^{-1} \sum_{1 \leqslant i_1 < \dots < i_{m(j)} \leqslant n} \Phi_j^s\left(\mathbf{Y}_{i_1}, \dots, \mathbf{Y}_{i_{m(j)}}\right). \tag{28}$$

Theorem 7.1 in [18] ensures that $U_{j,n}$ converges in probability to $I(\Phi_j)$ for any $j = 1, \dots, 4$. Moreover, one may also prove that the convergence holds almost surely proceeding as in the proof of Lemma 6.1 in [14]. Then we estimate $S^{\mathbf{1}}_{2,GMS}$ by

$$S^{\mathbf{1}}_{2,GMS,n} := \frac{U_{1,n} - U_{2,n}}{U_{3,n} - U_{4,n}} = \psi(U_{1,n}, U_{2,n}, U_{3,n}, U_{4,n}). \tag{29}$$

A novel estimation procedure In light of Section 3.1, one can introduce a novel estimation $\xi_n^{\mathrm{GMS}}(X_1, Y)$ of $S^{\mathbf{1}}_{2,GMS}$ in (23) as follows. The Pick-Freeze scheme is replaced by the use of the $Y_{N(i)}$'s where N is the permutation defined in (8) and the integration with respect to $\mathbb{P}^{\otimes m}$ is handled using U-statistics. More precisely, for $j = 1, \dots, 4$, we define

$$\widetilde{U}_{j,n} := \binom{n}{m(j)}^{-1} \sum_{1 \leqslant i_1 < \cdots < i_{m(j)} \leqslant n} \widetilde{\Phi}_j^s \left(\mathbf{Y}_{i_1}, \dots, \mathbf{Y}_{i_{m(j)}} \right), \tag{30}$$

where $\widetilde{\Phi}_j^s$ is the symetrized version of Φ_j^s with

$$\begin{aligned}
\widetilde{\Phi}_1(y_1, \dots, y_{m+1}) &:= T_{y_1,\dots,y_m}(y_{m+1}) T_{y_1,\dots,y_m}(y_{N(m+1)}) \\
\widetilde{\Phi}_2(y_1, \dots, y_{m+2}) &:= T_{y_1,\dots,y_m}(y_{m+1}) T_{y_1,\dots,y_m}(y_{N(m+2)}) \\
\widetilde{\Phi}_3(y_1, \dots, y_{m+1}) &:= T_{y_1,\dots,y_m}(y_{m+1})^2 \\
\widetilde{\Phi}_4(y_1, \dots, y_{m+2}) &:= T_{y_1,\dots,y_m}(y_{m+1}) T_{y_1,\dots,y_m}(y_{m+2}).
\end{aligned}$$

Finally, the estimator writes as

$$\xi_n^{\mathrm{GMS}}(X, Y) := \frac{\widetilde{U}_{1,n} - \widetilde{U}_{2,n}}{\widetilde{U}_{3,n} - \widetilde{U}_{4,n}} = \psi(\widetilde{U}_{1,n}, \widetilde{U}_{2,n}, \widetilde{U}_{3,n}, \widetilde{U}_{4,n}). \tag{31}$$

3.4 Owen higher-order moment indices

Following [25, 26], we consider extensions to Sobol indices obtained by replacing the numerator by higher-order moments. More precisely, for any integer $q \geqslant 2$, we set

$$H_q^{\mathbf{1}} := \mathbb{E}\left[(\mathbb{E}[Y|X_1] - \mathbb{E}[Y])^q \right],$$

see [14] for known properties $H_q^{\mathbf{1}}$.
In order to construct a Pick-Freeze estimator for $H_q^{\mathbf{1}}$, we first observe that

$$H_q^{\mathbf{1}} = \mathbb{E}\left[\prod_{m=1}^{q} \left((Y^{\mathbf{1}})^m - \mathbb{E}[Y] \right) \right] = \sum_{l=0}^{q} \binom{q}{l} (-1)^{q-l} \mathbb{E}\left[Y\right]^{q-l} \mathbb{E}\left[\prod_{m=1}^{l} (Y^{\mathbf{1}})^m \right]$$

with the usual convention $\prod_{m=1}^{0} (Y^{\mathbf{1}})^m = 1$. Here, $Y_1^{\mathbf{1}} = Y$ and for $i = 2, \dots, q$, $Y_i^{\mathbf{1}}$ is constructed independently (similarly to $Y^{\mathbf{1}}$ in (5)). Now we construct a Monte-Carlo scheme and consider the following Pick-Freeze design constituted by a n-sample $\left(Y_{i,j}^{\mathbf{1}}\right)_{(i,j) \in I_q \times I_n}$ of $\left(Y_1^{\mathbf{1}}, \dots, Y_q^{\mathbf{1}}\right)$ where, for any positive integer k, I_k stands for the set $\{1, \dots, k\}$.

The resulting Monte-Carlo estimator is then

$$H^{\mathbf{1}}_{q,n} = \sum_{l=0}^{q} \binom{q}{l} (-1)^{q-l} \left(\overline{P}^{\mathbf{1}}_{1}\right)^{q-l} \overline{P}^{\mathbf{1}}_{l}$$

where for any positive integer n, $j \in I_n$ and $l \in I_q$, we have set

$$P^{\mathbf{1}}_{l,j} = \binom{q}{l}^{-1} \sum_{k_1<\ldots<k_l \in I_q} \left(\prod_{i=1}^{l} Y^{\mathbf{1}}_{k_i,j}\right) \quad \text{and} \quad \overline{P}^{l}_{\mathbf{u}} = \frac{1}{N}\sum_{j=1}^{N} P^{\mathbf{1}}_{l,j}.$$

This setting generalizes the estimation procedure from [13] and uses all the available information by considering the means over the set of indices $k_1, \ldots, k_l \in I_d,\ k_n \neq k_m$.

Remark 3.4. While the collection of all indices $\left(H^{\mathbf{1}}_{q}\right)_q$ is much more informative than the classical Sobol indices, it also has several drawbacks. First, these indices are moment-based and, as is well known, they are not stable when the moment order increases. Second, they may be negative when q is odd. To overcome this fact, one may introduce $\mathbb{E}\left[|\mathbb{E}[Y|X_1] - \mathbb{E}[Y]|^q\right]$ but the Pick-Freeze estimation procedure is then lost. Third, the Pick-Freeze estimation procedure is computationally expensive and may be unstable: it requires a $q \times n$-sample of the output Y. In order to properly assess the influence of an input on the law of the output, we need to estimate the first $K-1$ indices $H^{\mathbf{1}}_{q}$: $H^{\mathbf{1}}_{2}, \ldots, H^{\mathbf{1}}_{K}$. Hence, we need to run the code $K \times n$ times. These indices are thus not attractive in practice.

We introduce below a new sensitivity index which is based on the conditional distribution of the output and requires only $3\times n$ runs. This index compares the output distribution to the conditional one whereas the q higher-order moment indices only compare the q-th output moment to the conditional one.

A novel estimation procedure We generalize the procedure proposed by Chatterjee in order to estimate higher-order moment indices. To that end, we introduce, for all $m \in \{1, \ldots, q-1\}$ and $j \in \{1, \ldots, n\}$,

$$N_m(j) = \begin{cases} \pi^{-1}(\pi(j)+m) & \text{if } \pi(j)+m \leqslant n, \\ \pi^{-1}(\pi(j)+m-n) & \text{if } \pi(j)+m > n. \end{cases} \tag{32}$$

Note that $N_1(j) = N(j)$ for all j. It remains to update Lemma 3.1 as follows.

Lemma 3.5. *Let $(f_m)_{m=0,\ldots,q-1}$ a family of measurable functions in $L^1(\mathbb{R})$. Let $(V_j, Y_j)_{1\leqslant j\leqslant n}$ be a n sample of (V, Y). Then*

$$\mathbb{E}\left[\prod_{m=0}^{q-1} f_m(Y_{N_m(i)}) | V_1, \ldots, V_n\right] = \prod_{m=0}^{q-1} \psi_{V_{N_m(i)}}(f_m), \tag{33}$$

where by convention $N_0(j) = j$ *for all* $j = 1, \dots, n$.

It suffices to take $f_m(y) = y$, for all $y \in \mathbb{R}$ and $m = 0, \dots, q-1$.

4 Extensions

4.1 Estimating all Sobol indices with a single sample

We consider higher order indices such as

$$S^{1,2} = \frac{\mathrm{Var}\mathbb{E}[Y|X_1, X_2]}{\mathrm{Var}(Y)}.$$

We want to estimate $S^{1,2}$ using the analogue to Chatterje's procedure; $S^{1,2}$ could be replaced by any two order index as for example the Cramér-von-Mises index. Let $W_j = (X_{1,j}, X_{2,j})$, $1 \leqslant j \leqslant n$, be n sample in $\mathbb{R}^2$. To define the analogue of the permutation given by (8), observe that it is the unique permutation (when we have no ties) that minimizes

$$\sum_{j=1}^{n} |V_{\sigma(j)} - V_{\sigma(j)+1}| \tag{34}$$

over the permutation σ, which is exactly the solution in dimension one of the traveling salesman problem. The reader is refered to, e.g., [3, 16] for a complete presentation of the traveling salesman problem. In dimension greater than one, we then naturally consider the permutation of the points W_i that solves the traveling salesman problem. Hence it is enough to consider the estimator (20) in order to estimate $S^{1,2}$.
Consequently, we are now able to extend easily Chatterjee's method

1. to estimate any sensitivity index of any order (for example, for an index of order k, it suffices to consider a permutation that solves the traveling salesman problem in $\mathbb{R}^k$); Moreover one can use the same n sample to estimate all Sobol indices of any order.

2. to consider codes whose inputs are not real-valued but take their values in any metric space.

In practice, one generally uses approximation algorithms to solve the traveling salesman problem (see, [3] and e.g., [21] for a genetic algorithm), which is not restrictive. Indeed, the only crucial point is be able to propose a permutation without fix point such that $\sigma(i)$ is close to i for any generic i.

4.2 Application: estimating all Shapley effects for correlated inputs

Shapley values were first introduced in game theory and economics [32] and later in the framework of sensitivity analysis in [28]. In this context, these indices have been called Shapley effects. They are used to quantify the importance of some input variables for correlated inputs. We refer to [28, 35, 29, 19] and the references therein for more details on Shapley effects. Here, we then consider that the input variables are no longer independent. In this case, it is still possible to define, for any $u \subset I_p$,

$$S^{\mathbf{u}} = \frac{\mathrm{Var}\left(\mathbb{E}[Y|X_{\mathbf{u}}]\right)}{\mathrm{Var}(Y)}$$

but its interpretation is no longer obvious. To overcome this difficulty, one can consider the so-called Shapley effects defined, for any $1 \leqslant i \leqslant p$, by

$$Sh^i := \sum_{\mathbf{w} \subset \{1\ldots,p\}, i \in \mathbf{w}} \frac{1}{|\mathbf{w}|} \sum_{\mathbf{v} \subset \mathbf{w}} (-1)^{|\mathbf{w}|-|\mathbf{v}|} S^{\mathbf{w}} \tag{35}$$

or equivalently,

$$Sh^i := \sum_{\mathbf{w} \subset \sim i} \frac{(p-|\mathbf{w}|-1)!|\mathbf{w}|!}{p!} (S^{\mathbf{w}\cup\{i\}} - S^{\mathbf{w}}). \tag{36}$$

In the case of independent input variables, one has

$$S^i \leqslant Sh^i \leqslant S^{i,Tot}$$

In [35], the authors propose two algorithms to estimate the Shapley effects from (36). The first one need to browse all the possible combinations of the input variables while the second one sample randomly permutations of the input variables. At each iteration, the expectation of a conditional variance has to be computed by the algorithm. In [4], the authors implement a bootstrap sampling in the existing algorithms to estimate confidence intervals of the indices estimation.

Now, using the result of Section 4.1 with a single n sample, one can consistently estimate all Sobol indices $S^{\mathbf{u}}$ and all Shapley effects at the same time.

5 Numerical experiments

5.1 Numerical comparison on the Sobol g-function: conventional Pick-Freeze estimators vs Chatterjee's estimators

In this section, we compare the performances of both estimation procedures on an analytic function, the so-called Sobol g-function, that is defined by:

$$g(X_1, \dots, X_p) = \prod_{i=1}^{p} \frac{|4X_i - 2| + a_i}{1 + a_i}, \tag{37}$$

where $(a_i)_{i\in\mathbb{N}}$ is a sequence of real numbers and the X_i's are i.i.d. random variables uniformly distributed on $[0, 1]$. In this setting, one my easily compute the exact expression of the first-order Sobol indices:

$$S^i = \frac{1/(3(1 + a_i^2))}{[\prod_{i=1}^{p} 1/(3(1 + a_i^2))] - 1}.$$

As expected, the lower the coefficient a_i, the more significant the variable X_i. In the sequel, we simply fix $a_i = i$.
Due to its complexity (non-linear and non-monotonic correlations) and the analytical expression of the Sobol indices, the Sobol g-function is a classical test example commonly used in global sensitivity analysis (see e.g. [30]).

Convergence as the sample size increases In Figure 1, we compare the estimations of the six first-order Sobol indices given by both methods. In the Pick-Freeze estimations, several sizes of sample N have been considered: $N = 100$, 500, 1000, 5000, 10000, 50000, 100000, and 500000. The Pick-Freeze procedure requires $(p + 1)$ samples of size N. To have a fair comparison, the sample sizes considered in the estimation of ξ_n^{Sobol} are $n = (p+1)N = 7N$. We observe that both methods converge and give precise results for large sample sizes.

Comparison of the mean square errors Now we want to compare the efficiency of both methods at a fixed sample size. In that view, we assume that only $n = 700$ calls of the computer code f are allowed to estimate the six first-order Sobol indices. We repeat the estimation procedure 500 times. The boxplot of the mean square errors for the estimation of the first-order Sobol index S^1 with respect to X_1 has been represented in Figure 2. We observe that, for a fixed sample size $n = 700$ (corresponding to a Pick-Freeze sample size

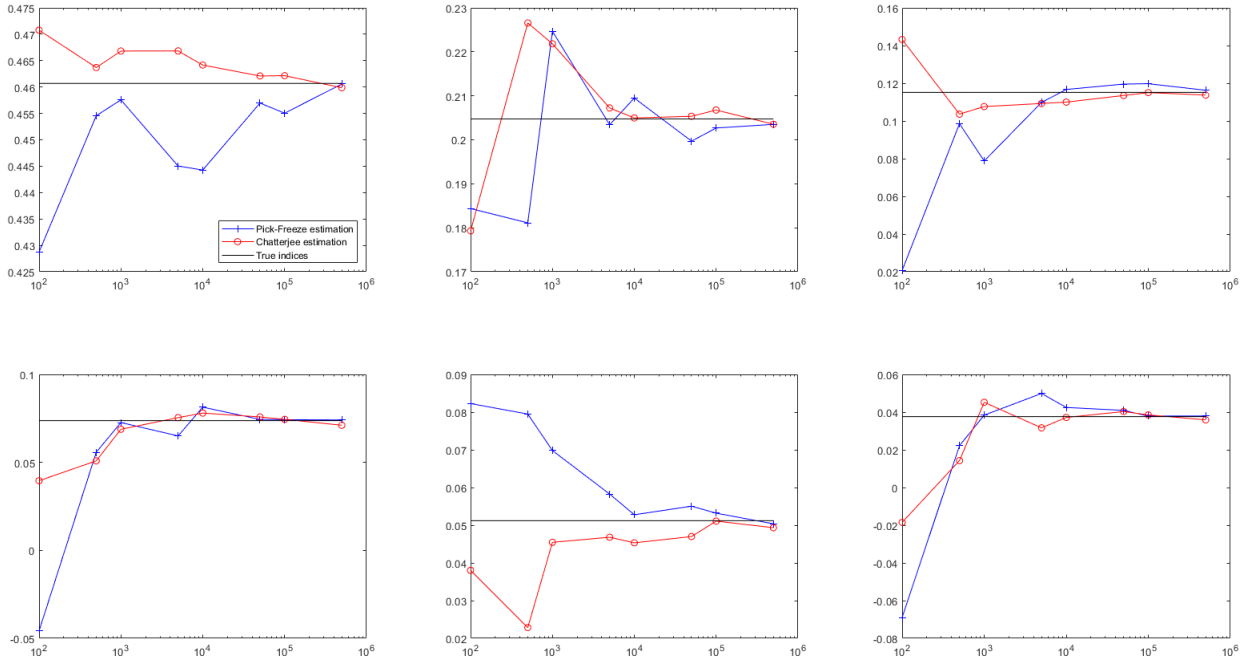

Figure 1: The Sobol g-function model (37). Convergence of both methods when N increases. The sixth first-order Sobol indices have been represented from left to right and up to bottom. Several sample sizes have been considered: $N = 100$, 500, 1000, 5000, 10000, 50000, 100000, and 500000 for the Pick-Freeze estimation procedure and correspondingly $(p+1)N$ for the estimation procedure proposed in [7]. The x-axis is in logarithmic scale.

$N = 100$), Chatterjee's estimation procedure performs much better than the Pick-Freeze method with significantly lower mean errors. The same behavior can be observed for all the first Sobol indices as can be seen in Table 1 that provides some characteristics of the mean squares errors.

Performances for small sample sizes or for large number of input variables As expected, we can observe in Table 2 that Chatterjee's procedure proceeds much better than the Pick-Freeze methodology for small sample sizes. Similarly, if the number of input variables increases drastically, we can observe the same behavior as can be seen in Figure 3. In that case, we consider the model (37) for several values of p: 6, 10, 15, 20, 30, 40, and 50.

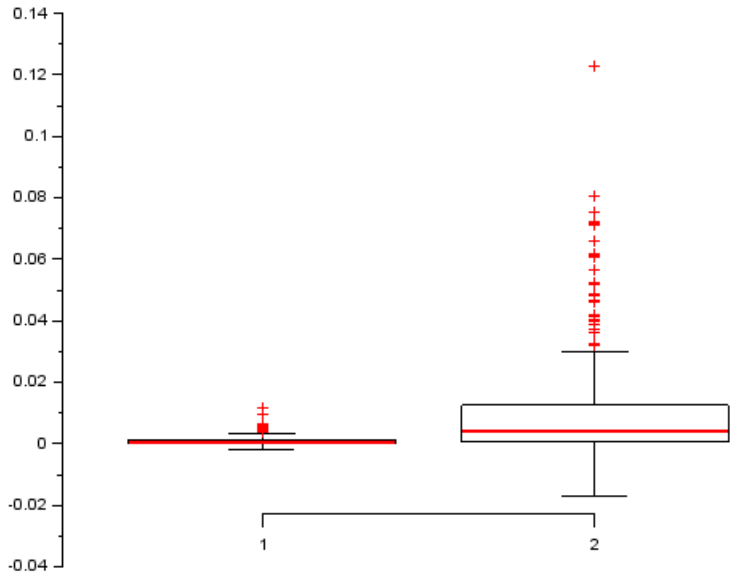

Figure 2: The Sobol g-function model (37). Boxplot of the mean square errors of the estimation of S^1 with a fixed sample size and 500 replications. The results of Chatterjee's methodology with $n = 700$ are provided in the left panel. The results of the Pick-Freeze estimation procedure with $N = 100$ are provided in the right panel.

	Pick-Freeze			Chatterjee		
	Mean	Median	Stdev	Mean	Median	Stdev
mse S^1	0.0095548	0.0039458	0.0145033	0.0010218	0.0004498	0.0013999
mse S^2	0.0105727	0.0046104	0.0148873	0.0017314	0.0006870	0.0027436
mse S^3	0.0101785	0.0041789	0.0143846	0.0016667	0.0006409	0.0024392
mse S^4	0.0105463	0.0047284	0.0178064	0.0018522	0.0008126	0.0025296
mse S^5	0.0097979	0.0042995	0.0135533	0.0016285	0.0006855	0.0024264
mse S^6	0.0096109	0.0046822	0.0134822	0.0015590	0.0007080	0.0021333

Table 1: The Sobol g-function model (37). Some characteristics of the mean square errors for the estimation of the six first-order Sobol indices with a fixed sample size and 500 replications. In Chatterjee's methodology, the sample size considered is $n = 700$ while in the Pick-Freeze estimation procedure, it is $N = 100$.

	Pick-Freeze			Chatterjee		
	$N = 10$	$N = 50$	$N = 100$	$n = 70$	$n = 350$	$n = 700$
mse S^1	0.1128686	0.0172275	0.0095548	0.0116790	0.0022941	0.0010218
mse S^2	0.1509575	0.0223196	0.0105727	0.0177522	0.0033719	0.0017314
mse S^3	0.1469124	0.0220015	0.0101785	0.0175517	0.0032474	0.0016667
mse S^4	0.1591130	0.0196357	0.0105463	0.0159360	0.0033948	0.0018522
mse S^5	0.1646339	0.0240353	0.0097979	0.0158563	0.0032230	0.0016285
mse S^6	0.1466408	0.0217638	0.0096109	0.0166701	0.0029653	0.0015590

Table 2: The Sobol g-function model (37). Mean squares errors of the estimation of the six first-order Sobol indices with small sample sizes and with both methods.

5.2 An application in biology

Now we illustrate the nature and performance of the Cramér-von-Mises indices and their corresponding Chaterjee estimators as a screening mechanism for high-dimensional problems. To do so, we consider the neurovascular coupling model from [17]. Mathematically, this corresponds to the following differential-algebraic equation (DAE) system

$$\frac{dW}{dt} = G(W, Z, X), \tag{38}$$

$$0 = H(W, Z, X), \tag{39}$$

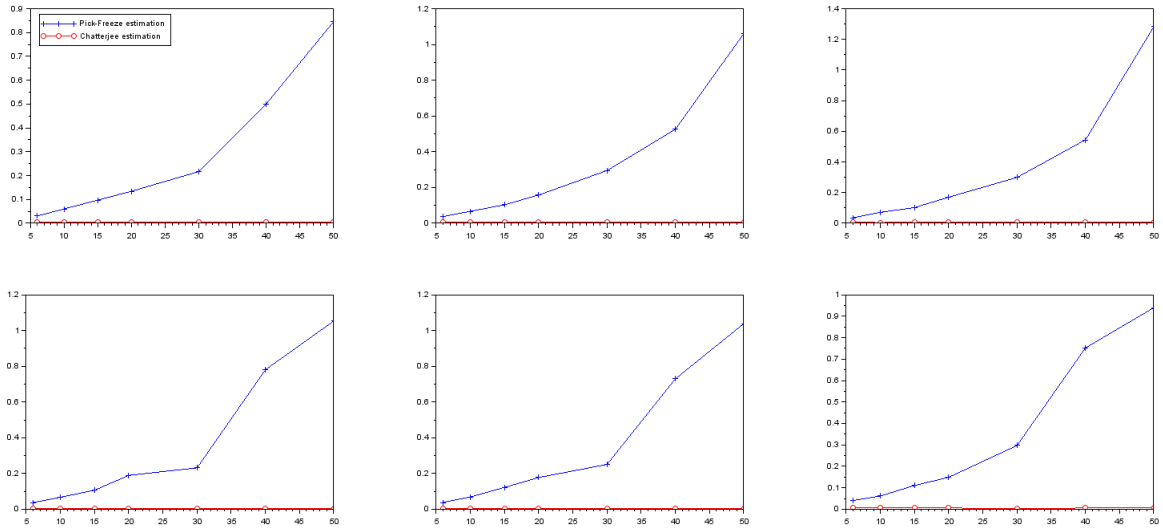

Figure 3: The Sobol g-function model (37). Mean square errors of the estimation of the six first-order Sobol indices with respect to the number of input variables with a fixed sample size and 500 replications. We consider the sample sizes $n = 200$ in Chatterjee's methodology and $N = n/(p+1)$ in the Pick-Freeze procedure. The number of input variables considered are $p = 6, 10, 15, 20, 30, 40,$ and 50.

where $W = (W_1, \ldots, W_N)$ and $Z = (Z_1, \ldots, Z_M)$ correspond respectively to the differential and algebraic state variables of the models. The variables $X = (X_1, \ldots, X_p)$ correspond to the uncertain parameters of the model. Our quantity of interest corresponds to the time average over $[0, T]$ of W^* (which is one of the differential state variables W_1, ..., W_N), i.e.

$$Y = \frac{1}{T} \int_0^T W^*(t)\, dt. \tag{40}$$

As above, we regard Y as a function of the unknown parameters, i.e., $Y = f(X_1, \ldots, X_p)$. In our implementation, the values of W^* are obtained by solving the above DAE system (Equations (38) and (39)) by the MATLAB routine ode15s (it can be checked that (38) and (39) form an index one system). Further, in the current example, $N = 67$ and $p = 160$ and the distributions of most of the X_i's are uniform and allowed to vary $\pm 10\%$ from nominal values (see [17] for additional details).
We compare the results from the Chaterjee estimators as described above to

those resulting from the linear regression

$$f(X_1, \ldots, X_{160}) \approx \lambda_0 + \sum_{j=1}^{160} \lambda_j X_j.$$

As shown in [17], the above approximation performs well for the considered QoI. We assign to each variable $X_1, \ldots, X_{160}$ a relative importance L_j where

$$L_j = \frac{|\lambda_j|}{\sum_{\ell=1}^{160} |\lambda_\ell|}, \qquad j = 1, \ldots, 160.$$

Figure 4 displays the results. Both screening approaches identify the same to three influential parameters. More parameters are identified as being non-influential through the linear regression approach than using the Cramér-von-Mises indices.

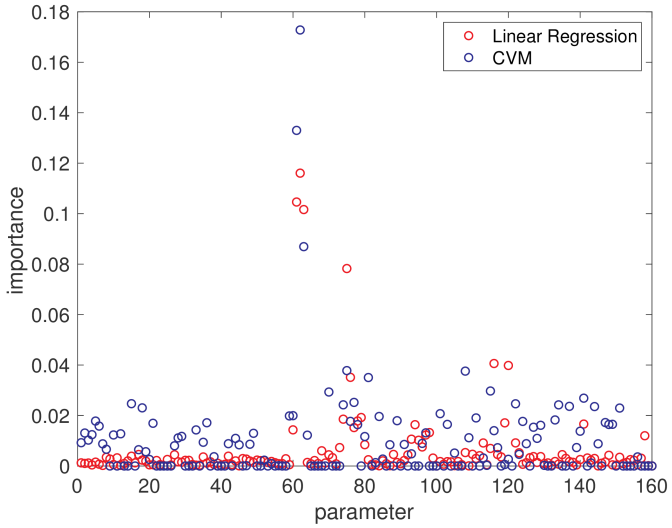

Figure 4: Chaterjee estimators corresponding to the Cramér- von-Mises indices as a screening mechanics for the DAE system given by (38) and (39).

6 Conclusion

In this paper, we explain how to use the estimator proposed by Chatterjee in [7] to provide a very nice and mighty procedure to estimate both all the order one Sobol indices and the so-called Cramér-von-Mises indices [14] at a small cost (only n calls of the computer code). We also extend Chatterjee's method

to estimate more general quantities. As examples, we consider two indices already introduced in sensitivity analysis: the indices adapted to output valued in general metric spaces defined in [15] and the higher-moment indices [25, 26]. In addition, we extend the procedure to estimate not only the first-order indices but also second-order and even higher-order indices. Consequently, the Shapley effects defined for correlated inputs are then also easily estimated.
Our analysis paves the way for further research directions. For instance, Chaterjee proves a central limit theorem for (7) when X and Y are independent. A first interesting step would be to establish a central limit theorem for the estimators (7) and more generally (12) in any case.

Acknowledgment. We warmly thank Robin Morillo for the numerical study provided in Section 5.2. Support from the ANR-3IA Artificial and Natural Intelligence Toulouse Institute is gratefully acknowledged. This work was also supported by the National Science Foundation under grant DMS-1745654.

References

[1] A. Alexanderian, P. Gremaud, and R. Smith. Variance-based sensitivity analysis for time-dependent processes. *Reliability Eng. Sys. Safety*, 196:106722, 2020.

[2] A. Antoniadis. Analysis of variance on function spaces. *Statistics: A Journal of Theoretical and Applied Statistics*, 15(1):59–71, 1984.

[3] D. L. Applegate, R. E. Bixby, V. Chvatal, and W. J. Cook. *The traveling salesman problem: a computational study*. Princeton university press, 2006.

[4] N. Benoumechiara and K. Elie-Dit-Cosaque. Shapley effects for sensitivity analysis with dependent inputs: bootstrap and kriging-based algorithms. *arXiv preprint arXiv:1801.03300*, 2018.

[5] E. Borgonovo. A new uncertainty importance measure. *Reliability Engineering & System Safety*, 92(6):771–784, 2007.

[6] E. Borgonovo, W. Castaings, and S. Tarantola. Moment independent importance measures: New results and analytical test cases. *Risk Analysis*, 31(3):404–428, 2011.

[7] S. Chatterjee. A new coefficient of correlation. *arXiv e-prints*, page arXiv:1909.10140, Sep 2019.

[8] S. Da Veiga. Global sensitivity analysis with dependence measures. *J. Stat. Comput. Simul.*, 85(7):1283–1305, 2015.

[9] E. De Rocquigny, N. Devictor, and S. Tarantola. *Uncertainty in industrial practice*. Wiley Online Library, 2008.

[10] J.-C. Fort, T. Klein, and N. Rachdi. New sensitivity analysis subordinated to a contrast. *Communications in Statistics - Theory and Methods*, 2015 to appear.

[11] R. Fraiman, F. Gamboa, and L. Moreno. Sensitivity indices for output on a Riemannian manifold. *arXiv e-prints*, page arXiv:1810.11591, Oct 2018.

[12] F. Gamboa, A. Janon, T. Klein, and A. Lagnoux. Sensitivity analysis for multidimensional and functional outputs. *Electronic Journal of Statistics*, 8:575–603, 2014.

[13] F. Gamboa, A. Janon, T. Klein, A. Lagnoux, and C. Prieur. Statistical inference for Sobol pick-freeze Monte Carlo method. *Statistics*, 50(4):881–902, 2016.

[14] F. Gamboa, T. Klein, and A. Lagnoux. Sensitivity analysis based on Cramér von Mises distance. *SIAM/ASA Journal on Uncertainty Quantification*, 6(2):522–548, Apr. 2018.

[15] F. Gamboa, T. Klein, A. Lagnoux, and L. Moreno. Sensitivity analysis in general metric spaces. working paper or preprint, Feb. 2019.

[16] G. Gutin and A. P. Punnen. *The traveling salesman problem and its variations*, volume 12. Springer Science & Business Media, 2006.

[17] J. Hart, P. Gremaud, and T. David. Global sensitivity analysis of high dimensional neuroscience models: an example of neurovascular coupling. *Bull Math Biol*, 2019.

[18] W. Hoeffding. A class of statistics with asymptotically normal distribution. *Ann. Math. Statistics*, 19:293–325, 1948.

[19] B. Iooss and C. Prieur. Shapley effects for sensitivity analysis with correlated inputs: comparisons with Sobol'indices, numerical estimation and applications. *International Journal for Uncertainty Quantification*, 9(5), 2019.

[20] A. Janon, T. Klein, A. Lagnoux, M. Nodet, and C. Prieur. Asymptotic normality and efficiency of two Sobol index estimators. *ESAIM: Probability and Statistics*, 18:342–364, 1 2014.

[21] J. Kirk. Traveling salesman problem - genetic algorithm. 2014.

[22] S. Kucherenko and S. Song. Different numerical estimators for main effect global sensitivity indices. *Reliability Engineering & System Safety*, 165:222–238, 2017.

[23] M. Lamboni, H. Monod, and D. Makowski. Multivariate sensitivity analysis to measure global contribution of input factors in dynamic models. *Reliability Engineering & System Safety*, 96(4):450–459, 2011.

[24] C. McDiarmid. On the method of bounded differences. *Surveys in combinatorics*, 141(1):148–188, 1989.

[25] A. Owen. Variance components and generalized Sobol' indices. *SIAM/ASA Journal on Uncertainty Quantification*, 1(1):19–41, 2013.

[26] A. Owen, J. Dick, and S. Chen. Higher order Sobol' indices. *Information and Inference*, 3(1):59–81, 2014.

[27] A. B. Owen. Better estimation of small Sobol'sensitivity indices. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 23(2):11, 2013.

[28] A. B. Owen. Sobol'indices and Shapley value. *SIAM/ASA Journal on Uncertainty Quantification*, 2(1):245–251, 2014.

[29] A. B. Owen and C. Prieur. On Shapley value for measuring importance of dependent inputs. *SIAM/ASA Journal on Uncertainty Quantification*, 5(1):986–1002, 2017.

[30] A. Saltelli, K. Chan, and E. Scott. *Sensitivity analysis*. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd., Chichester, 2000.

[31] T. J. Santner, B. Williams, and W. Notz. *The Design and Analysis of Computer Experiments*. Springer-Verlag, 2003.

[32] L. S. Shapley. A value for n-person games. *Contributions to the Theory of Games*, 2(28):307–317, 1953.

[33] I. Sobol. Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Mathematics and Computers in Simulation*, 55(1-3):271–280, 2001.

[34] I. M. Sobol. Sensitivity estimates for nonlinear mathematical models. *Math. Modeling Comput. Experiment*, 1(4):407–414 (1995), 1993.

[35] E. Song, B. L. Nelson, and J. Staum. Shapley effects for global sensitivity analysis: Theory and computation. *SIAM/ASA Journal on Uncertainty Quantification*, 4(1):1060–1083, 2016.

[36] B. Sudret. Global sensitivity analysis using polynomial chaos expansions. *Reliability Engineering & System Safety*, 93(7):964–979, 2008.

[37] A. W. van der Vaart. *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1998.