Sensitivity analysis based on Cramér von Mises distance

Fabrice Gamboa Thierry Klein Agnès Lagnoux[*]

June 12, 2015

Abstract

In this paper, we first study a new sensitivity index that is based on higher moments and generalizes the so-called Sobol one. Further, following an idea of Borgonovo ([3]), we define and study a new sensitivity index based on the Cramér von Mises distance. This new index appears to be more general than the Sobol one as it takes into account, not only the variance, but the whole distribution of the random variable. Furthermore, we study the statistical properties of a Monte Carlo estimate of this new index.

Keywords: Sensitivity analysis, Cramér von Mises distance, Pick and Freeze method, functional delta-method, Anderson-Darling statistic.

1 Introduction

A very classical problem in the study of computer code experiments (see [21]) is the evaluation of the relative influence of the input variables on some numerical result obtained by a computer code. This study is usually called sensitivity analysis in this paradigm and has been widely assessed (see for example [22], [20], [11] and references therein). More precisely, the result of the numerical code Y is seen as a function of the vector of the distributed input $(X_r)_{r=1,\cdots,d}$ $(d \in \mathbb{N}^*)$. Statistically speaking, we are dealing here with the unnoisy non parametric model

$$Y = f(X_1, \dots, X_d), \tag{1}$$

where f is a regular unknown numerical function on the state space $E_1 \times E_2 \times \ldots \times E_d$ on which the distributed variables $(X_1, \dots, X_d)$ are living. Generally, the random inputs are assumed to be independent and a sensitivity analysis is performed by using the so-called Hoeffding decomposition (see [23] and [1]). In this functional decomposition, f is expanded as an L^2-sum of uncorrelated functions involving only a part of the random inputs. For any subset v of $I_d = \{1, \dots, d\}$, this leads to an index called the Sobol index ([22]) that measures the amount of *randomness* of Y carried in the subset of input variables $(X_i)_{i\in v}$. Since nothing has been assumed on the nature of the inputs, one can consider the vector $(X_i)_{i\in v}$ as a single input. Thus without loss of generality, let us consider the case where v reduces to a singleton. The numerator H_v of the Sobol index related to the input X_v is

$$H_v = \mathrm{Var}\left(\mathbb{E}\left[Y|X_v\right]\right) = \mathrm{Var}(Y) - \mathbb{E}\left[\left(Y - \mathbb{E}\left[Y|X_v\right]\right)^2\right] \tag{2}$$

while the denominator of the index is nothing more than the variance of Y. In order to estimate H_v the clever trick discovered by Sobol [22] is to rewrite the variance of the conditional expectation as a covariance. Further, a well tailored design of experiment called the Pick and Freeze scheme is considered [16]. More precisely, let X^v be the random vector such that $X^v_v = X_v$ and $X^v_i = X'_i$ if $i \neq v$ where X'_i is an independent copy of X_i. Then, setting

$$Y^v := f(X^v) \tag{3}$$

an obvious computation leads to the nice relationship

$$\mathrm{Var}(\mathbb{E}(Y|X_v)) = \mathrm{Cov}\left(Y, Y^v\right). \tag{4}$$

[*] Institut de Mathématiques de Toulouse, 118 Route de Narbonne 31062 Toulouse Cedex 9. France. `firstname.lastname@math.univ-toulouse.fr`

The last equality leads to a natural Monte Carlo estimator (Pick and Freeze estimator)

$$T_{N,\mathrm{Cl}}^{v} = \frac{1}{N}\sum_{j=1}^{N} Y_j Y_j^v - \left(\frac{1}{2N}\sum_{j=1}^{N}(Y_j + Y_j^v)\right)^2 \tag{5}$$

where for $j = 1, \cdots, N$, Y_j (resp. Y_j^v) are independent copies of Y (resp. Y^v). The sharp statistical properties and some functional extensions of the Pick and Freeze method are considered in [16], [15] and [10]. Notice that the Sobol indices and their Monte Carlo estimation are based on order two methods since they derived from the L^2 Hoeffding functional decomposition. This is the main drawback of this kind of methods. As an illustration consider the following example. Let X_1 and X_2 be two independent random variables having the same first four moments (equal e.g. to 1) and such that $\mathbb{E}\left[X_1^3\right] \neq \mathbb{E}\left[X_2^3\right]$. Let us consider the following model

$$Y = X_1 + X_2 + X_1^2 X_2^2.$$

Then

$$\mathrm{Var}\left(\mathbb{E}\left[Y|X_1\right]\right) = \mathrm{Var}(X_1 + X_1^2) = \mathrm{Var}(X_2 + X_2^2) = \mathrm{Var}\left(\mathbb{E}\left[Y|X_2\right]\right).$$

However, since Y is a symmetric function of the inputs X_1 and X_2 that do not share the same distribution, X_1 and X_2 should not have the same importance. That shows the need to introduce a sensitivity index that takes into account not only the second order behaviors but all the distributions. As pointed out before, Sobol indices are based on L^2 decomposition. As a matter of fact, Sobol indices are well adapted to measure the contribution of an input on the deviation around the mean of Y. However, it seems very intuitive that the sensitivity of an extreme quantile of Y could depend on sets of variables that cannot be read only in the variances. Thus the same index should not be used for any task and we need to define more adapted indices. There are several ways to generalize the Sobol indices. One can, for example, define new indices through contrast functions based on the quantity of interest (see [13]). Unfortunately the Monte Carlo estimator of these new indices are computationally very expensive. In [9], the author presents a way to define moment independent measures through dissimilarity distances. These measures define a unified framework that encompasses some already known sensitivity indices. Unfortunately, the estimation of such indices relies on the estimation of density ratio estimation that can be computationally expensive. Now, as pointed out in [3], [6], [7], [18] and [19], there are situations where higher order methods give a sharper analysis on the relative influence of the input and allow finer screening procedures. Borgonovo et al. propose and study an index based on the total variation distance (see [3], [6] and [7]). While Owen et al. suggest to use procedures based on higher moments (see [18], [19]). Our paper follows these tracks. We will first revisit the works of Owen et al. by studying the asymptotic properties of the multiple Pick and Freeze scheme proposed therein for the estimation of higher order Sobol indices. Further, we propose a new natural index based on the Cramér von Mises distance between the distribution of the output Y and its conditional law when an input is fixed. We will show that this approach leads to natural self-normalized indices as in the case of the Sobol-Hoeffding decomposition of the variance. As a matter of fact, as for Sobol indices, the sum of all first order indices can not exceeds one. Notice that these indices extend naturally to multivariate outputs. Furthermore, we show that surprisingly a Pick and Freeze scheme is also available to estimate this new index. The sample size required to build such an estimator is of the same order as the size needed for the classical Sobol index estimation allowing its use in concrete situations.

The paper is divided in three sections. In the next section, we will study the statistical properties of the multiple Pick and Freeze method proposed earlier by Owen et al ([18], [19]). Section 3 is devoted to the new index built on the Cramér von Mises distance. In the last section, we give some numerical simulation that illustrate the interest of the new index. In particular, we revisit a real data example introduced in [8] and studied in [12] and [5].

2 Multiple Pick and Freeze method

Using the classical Hoeffding decomposition, for a singleton $v \in I_d$, the numerator of the classical Sobol index with respect to v is given by

$$H_v^2 = \mathbb{E}\left[\left(\mathbb{E}[Y|X_v] - \mathbb{E}[Y]\right)^2\right]. \tag{6}$$

Following [18] and [19], we generalize this quantity by considering higher order moments. Indeed, for any integer $p \geqslant 2$, we set

$$H_v^p := \mathbb{E}\left[(\mathbb{E}[Y|X_v] - \mathbb{E}[Y])^p\right]. \tag{7}$$

$H_v = H_v^2$. The following lemma gives the Pick and Freeze representation of H_v^p for $p \geqslant 2$.

Lemma 2.1. *For any $v \in I_d$, one has*

$$\mathbb{E}\left[(\mathbb{E}[Y|X_v] - \mathbb{E}[Y])^p\right] = \mathbb{E}\left[\prod_{i=1}^{p}\left(Y^{v,i} - \mathbb{E}[Y]\right)\right]. \tag{8}$$

Here, $Y^{v,1} = Y$ and for $i = 2, \dots, p$, $Y^{v,i}$ is constructed independently as Y^v defined in equation (1.3).

Obviously, H_v^p is non negative for even p and

$$|H_v^p| \leqslant \mathbb{E}\left[|Y - \mathbb{E}[Y]|^p\right].$$

Further, H_v^p is invariant by any translation of the output.

Estimation procedure In view of the estimation of H_v^p, we first expand the product in the right-hand side of (2.3) to get that

$$H_v^p = \sum_{l=0}^{p} \binom{p}{l} (-1)^{p-l} \mathbb{E}\left[Y\right]^{p-l} \mathbb{E}\left[\prod_{i=1}^{l} Y^{v,i}\right].$$

with the usual convention $\prod_{i=1}^{0} Y^{v,i} = 1$. Second, we use a Monte Carlo scheme and consider the following Pick and Freeze design constituted by the following $p \times N$-sample

$$\left(Y_j^{v,i}\right)_{(i,j) \in I_p \times I_N}.$$

We define for any any $N \in \mathbb{N}^*$, $j \in I_N$ and $l \in I_p$,

$$P_{l,j}^v = \binom{p}{l}^{-1} \sum_{k_1 < \ldots < k_l \in I_p} \left(\prod_{i=1}^{l} Y_j^{v,k_i}\right) \quad \text{and} \quad \overline{P}_l^v = \frac{1}{N}\sum_{j=1}^{N} P_{l,j}^v.$$

The Monte Carlo estimator is then

$$H_{p,N}^v = \sum_{l=0}^{p} \binom{p}{l} (-1)^{p-l} \left(\overline{P}_1^v\right)^{p-l} \overline{P}_l^v. \tag{9}$$

Notice that we generalize the estimation procedure of [15] and use all the available information by considering the means over the set of indices $k_1, \dots, k_l \in I_d$, $k_n \neq k_m$. The following theorem provides asymptotic properties of $H_{p,N}^v$.

Theorem 2.2. *$H_{p,N}^v$ is consistent and asymptotically Gaussian:*

$$\sqrt{N}\left(H_{p,N}^v - H_p^v\right) \underset{N\to\infty}{\overset{\mathcal{L}}{\to}} \mathcal{N}\left(0, \sigma^2\right) \tag{10}$$

where

$$\sigma^2 = p\left[\mathrm{Var}(Y) + (p-1)\mathrm{Cov}(Y, Y^{v,2})\right]\left(\sum_{l=1}^{p} a_l b_l\right)^2,$$

$$a_l = \frac{l}{p}\mathbb{E}[Y]^{l-1}, \qquad l = 1, \dots, p$$

$$b_1 = (-1)^{p-1}p(p-1)\mathbb{E}[Y]^{p-1} + \sum_{l=2}^{p-1}\binom{p}{l}(-1)^{p-l}(p-l)\mathbb{E}[Y]^{p-l-1}\mathbb{E}\left[\prod_{i=1}^{l} Y^{v,i}\right]$$

and

$$b_l = \binom{p}{l}(-1)^{p-l}\mathbb{E}[Y]^{p-l}, \qquad l = 1, \dots, p.$$

Proof of Theorem **??**. The consistency follows from a straightforward application of the strong law of large numbers. The asymptotic normality is derived by two successive applications of the delta method [23] .

(1) Let $W_j^1 = (Y_j^{v,1}, \dots, Y_j^{v,p})^T$ $(j = 1, \dots, N)$ and g^1 the mapping from $\mathbb{R}^p$ to $\mathbb{R}^p$ whose l-th coordinate is given by

$$g_l^1(x_1, \dots, x_p) = \binom{p}{l}^{-1} \sum_{\substack{k_1 < \dots < k_l \\ k_i \in I_p, i = 1, \dots, l}} \left(\prod_{i=1}^{l} x_{k_i} \right).$$

Let Σ^1 be the covariance matrix of W_j^1. Clearly, one has $\Sigma_{ii}^1 = \mathrm{Var}(Y)$ for $i \in I_p$ while $\Sigma_{ij}^1 = \mathrm{Cov}(Y^{v,i}, Y^{v,j}) = \mathrm{Cov}(Y, Y^{v,2})$. The multidimensional central limit theorem gives with $m = (\mathbb{E}[Y], \dots, \mathbb{E}[Y])^T$

$$\sqrt{N} \left(\frac{1}{N} \sum_{j=1}^{N} W_j^1 - m \right) \underset{N \to \infty}{\overset{\mathcal{L}}{\to}} \mathcal{N}_p \left(0, \Sigma^1 \right).$$

We then apply the so-called delta method to W^1 and g^1 so that

$$\sqrt{N} \left(g^1 \left(\overline{W}_N^1 \right) - g^1 \left(\mathbb{E} \left[W^1 \right] \right) \right) \underset{N \to \infty}{\overset{\mathcal{L}}{\to}} \mathcal{N} \left(0, J_{g^1} \left(\mathbb{E} \left[W^1 \right] \right) \Sigma^1 J_{g^1} \left(\mathbb{E} \left[W^1 \right] \right)^T \right)$$

with $J_{g^1} \left(\mathbb{E} \left[W^1 \right] \right)$ the Jacobian of g^1 at point $\mathbb{E} \left[W^1 \right]$. Notice that for $i \in I_p$ and $k \in I_p$,

$$\frac{\partial g_l^1}{\partial x_k} \left(\mathbb{E} \left[W^1 \right] \right) = \frac{\binom{p-1}{l-1}}{\binom{p}{l}} m^{l-1} = \frac{l}{p} \mathbb{E}[Y]^{l-1} =: a_l.$$

Thus $\Sigma^2 := J_{g^1} \left(\mathbb{E} \left[W^1 \right] \right) \Sigma^1 J_{g^1} \left(\mathbb{E} \left[W^1 \right] \right)^T$ is given by

$$\Sigma_{ij}^2 = p a_i a_j \left(\Sigma_{11}^1 + (p-1) \Sigma_{12}^1 \right).$$

(2) Now consider $W_j^2 = (P_j^{v,1}, \dots P_j^{v,p})^T$ $(j = 1, \dots, N)$ and g^2 the mapping from $\mathbb{R}^p$ to $\mathbb{R}$ defined by

$$g^2(y_1, \dots, y_p) = \sum_{l=0}^{p} \binom{p}{l} (-1)^{p-l} y_1^{p-l} y_l.$$

We apply once again the delta method to W^2 so that

$$\sqrt{N} \left(g^2 \left(\overline{W}_N^2 \right) - g^2 \left(\mathbb{E} \left[W^2 \right] \right) \right) \underset{N \to \infty}{\overset{\mathcal{L}}{\to}} \mathcal{N} \left(0, J_{g^2} \left(\mathbb{E} \left[W^2 \right] \right) \Sigma^2 J_{g^2} \left(\mathbb{E} \left[W^2 \right] \right)^T \right)$$

with $J_{g^2} \left(\mathbb{E} \left[W^2 \right] \right)$ the Jacobian of g^2 at point $\mathbb{E} \left[W^2 \right]$. Notice that for $k \in I_p$,

$$\begin{aligned} \frac{\partial g^2}{\partial y_1} \left(\mathbb{E} \left[W^2 \right] \right) &= (-1)^{p-1} p(p-1) \mathbb{E}[Y]^{p-1} \\ &\quad + \sum_{l=2}^{p-1} \binom{p}{l} (-1)^{p-l} (p-l) \mathbb{E}[Y]^{p-l-1} \mathbb{E} \left[\prod_{i=1}^{l} Y^{v,i} \right] \end{aligned}$$

and

$$\frac{\partial g^2}{\partial y_l} \left(\mathbb{E} \left[W^2 \right] \right) = \binom{p}{l} (-1)^{p-l} \mathbb{E}[Y]^{p-l}.$$

Thus the limiting variance is

$$\sigma^2 := J_{g^2} \left(\mathbb{E} \left[W^2 \right] \right) \Sigma^2 J_{g^2} \left(\mathbb{E} \left[W^2 \right] \right)^T = p \left(\Sigma_{11}^1 + (p-1) \Sigma_{12}^1 \right) \left(\sum_{i=1}^{p} a_i b_i \right)^2,$$

where b_i is the i-th coordinate of $\nabla g^2 \left(\mathbb{E} \left[W^2 \right] \right)$. □

The collection of all indices H_v^p is much more informative than the classical Sobol index. Nevertheless it has several drawbacks: it may be negative when p is odd. To overcome this fact, we may have introduced $\mathbb{E}\left[|\mathbb{E}[Y|X_i, i \in v] - \mathbb{E}[Y]|^p\right]$ but proceeding in such a way, we would have loose the Pick and Freeze estimation procedure. The Pick and Freeze estimation procedure is computationally expensive: it requires a $p \times N$ sample of the output Y. In a sense, if we want to have a good idea of the influence of an input on the law of the output, we need to estimate the first d indices H_v^p and hence we need to run the black-box code $K \times N$ times. Moreover, these indices are moment based and it is well known that they are not stable when the moment order increases. In the next section, we introduce a new sensitivity index that is based on the conditional distribution of the output and requires only $3 \times N$.

3 The Cramér von Mises index

In this section the code will be denoted by $Z = f(X_1, \dots, X_d) \in \mathbb{R}^k$. Let F be the distribution function of Z. For any $t = (t_1, \dots, t_k) \in \mathbb{R}^k$,

$$F(t) = \mathbb{P}(Z \leqslant t) = \mathbb{E}\left[\mathbb{1}_{\{Z \leqslant t\}}\right]$$

and $F^v(t)$ the conditional distribution function of Z conditionally on X_v:

$$F^v(t) = \mathbb{P}(Z \leqslant t | X_v,) = \mathbb{E}\left[\mathbb{1}_{\{Z \leqslant t\}} | X_v\right].$$

Notice that $\{Z \leqslant t\}$ means that $\{Z_1 \leqslant t_1, \dots, Z_k \leqslant t_k\}$. Obviously, $\mathbb{E}\left[F^v(t)\right] = F(t)$. Now, we apply the framework presented in Section 2 with $Y(t) = \mathbb{1}_{\{Z \leqslant t\}}$ and $p = 2$. Hence, for $t \in \mathbb{R}^k$ fixed, we have a consistent and asymptotically normal estimation procedure for the estimation of

$$\mathbb{E}\left[(F(t) - F^v(t))^2\right].$$

We define a Cramér Von Mises type distance of order 2 between $\mathcal{L}(Z)$ and $\mathcal{L}(Z|X_v)$ by

$$D_{2,CVM}^v := \int_{\mathbb{R}^k} \mathbb{E}\left[(F(t) - F^v(t))^2\right] dF(t). \tag{11}$$

The aim of the rest of the section is dedicated to the estimation of $D_{2,CVM}^v$ and the study of the asymptotic properties of the estimator. Notice that

$$D_{2,CVM}^v = \mathbb{E}\left[\mathbb{E}\left[(F(Z) - F^v(Z))^2\right]\right]. \tag{12}$$

Let us note that these indices are naturally adapted to multivariate outputs.

Remark 3.1. Unlike the procedure for $p = 2$, we did not normalize the generalized Sobol index of $Y(t)$. The purpose, that becomes clear in this section, is to avoid numerical explosion during the estimation procedure. Indeed, the normalizing term would be $F(t)(1 - F(t))$, like in the ANderson-Darling statistic, canceling for small and large values of t. Nevertheless, in view of the following proposition, one can consider $4D_{2,CVM}^v$ instead of $D_{2,CVM}^v$ in order to have an index bounded by 1 as for the Sobol index. The asymptotic properties will not be affected by this renormalizing factor, so we still consider $D_{2,CVM}^v$.

Proposition 3.2. *One has the following properties.*

1. $0 \leqslant D_{2,CVM}^v \leqslant \frac{1}{4}$. *Moreover, if $k = 1$ and F is continuous, we have* $0 \leqslant D_{2,CVM}^v \leqslant \frac{1}{6}$.
2. $D_{2,CVM}^v$ *is invariant by translation, by left-composition by any nonzero scaling of Y.*

We then proceed to a double Monte-Carlo scheme for the estimation of $D_{2,CVM}^v$ and consider the following design of experiment consisting in:

1. two N-samples of Z: $(Z_j^{v,1}, Z_j^{v,2})$, $1 \leqslant j \leqslant N$;
2. a third N-sample of Z independent of $(Z_j^{v,1}, Z_j^{v,2})_{1 \leqslant j \leqslant N}$: W_k, $1 \leqslant k \leqslant N$.

The empirical estimator of $D_{2,CVM}^{v}$ is then given by

$$\widehat{D}_{2,CVM}^{v}=\frac{1}{N}\sum_{k=1}^{N}\left\{\frac{1}{N}\sum_{j=1}^{N}1\!\!1_{\{Z_j^{v,1}\leqslant W_k\}}1\!\!1_{\{Z_j^{v,2}\leqslant W_k\}}-\left[\frac{1}{2N}\sum_{j=1}^{N}\left(1\!\!1_{\{Z_j^{v,1}\leqslant W_k\}}+1\!\!1_{\{Z_j^{v,2}\leqslant W_k\}}\right)\right]^2\right\}.$$

The consistency of $\widehat{D}_{2,CVM}^{v}$ follows directly from the following lemma:

Lemma 3.3. *Let G and H be two L^1-measurable functions. Let $(U_j)_{j\in I_N}$ and $(V_k)_{k\in I_N}$ be two independent samples of iid rv such that $\mathbb{E}[G(U_1,V_1)]=0$ and $\mathbb{E}[H(U_1,U_2,V_1)]=0$. We define S_N and T_N by*

$$S_N=\frac{1}{N^2}\sum_{j,k=1}^{N}G(U_j,V_k)\quad\textit{and}\quad T_N=\frac{1}{N^3}\sum_{i,j,k=1}^{N}H(U_i,U_j,V_k).$$

Then S_N and T_N converge a.s. to 0 as N goes to infinity.

Proof. (i) If we prove that $\mathbb{E}[S_N^4]=O\left(\frac{1}{N^2}\right)$, we then apply Borel-Cantelli lemma to deduce the almost sure convergence of S_N to 0. Clearly,

$$\mathbb{E}[S_N^4]=\frac{1}{N^8}\sum\mathbb{E}[G(U_{i_1},V_{j_1})G(U_{i_2},V_{j_2})G(U_{i_3},V_{j_3})G(U_{i_4},V_{j_4})]$$

where the sum is taken over all the indices i_1, i_2, i_3, i_4, j_1, j_2, j_3, j_4 from 1 to N. The only scenarii that could lead to terms in $O\left(\frac{1}{N}\right)$ or even $O\left(1\right)$ appear when we sum over indices all different except 2 i's or 2 j's or over indices all different. Nevertheless, in those cases, at least one term of the form $\mathbb{E}[G(U_i,V_j)]$ appears. Since the function G is centered, those scenarii are then discarded.

(ii) Analogously, it suffices to show that $\mathbb{E}[T_N^4]=O\left(\frac{1}{N^2}\right)$. The only scenarii that could lead to terms in $O\left(\frac{1}{N}\right)$ or even $O\left(1\right)$ appear when we sum over indices all different except 2 i's, 2 j's or 2 k's or over indices all different. Nevertheless, in those cases, at least one term of the form $\mathbb{E}[H(U_i,U_j,V_k)]$ appears. Since the function H is centered, those scenarii are then discarded. □

Corollary 3.4. *$\widehat{D}_{2,CVM}^{v}$ is strongly consistent as N goes to infinity.*

Proof. The proof is based on Lemma **??**. First, we define $Z_j=\left(Z_j^{v,1},Z_j^{v,2}\right)$, $G(Z_j,W_k)=1\!\!1_{\{Z_j^{v,1}\leqslant W_k\}}1\!\!1_{\{Z_j^{v,2}\leqslant W_k\}}$, $F(Z_j,W_k)=\frac{1}{2}\left(1\!\!1_{\{Z_j^{v,1}\leqslant W_k\}}+1\!\!1_{\{Z_j^{v,2}\leqslant W_k\}}\right)$ and $H(Z_i,Z_j,W_k)=F(Z_i,W_k)F(Z_j,W_k)$. Second we proceed to the following decomposition

$$
\begin{aligned}
\widehat{D}^v_{2,CVM} &= \frac{1}{N}\sum_{k=1}^{N}\left\{\frac{1}{N}\sum_{j=1}^{N}\mathbb{1}_{\{Z_j^{v,1}\leqslant W_k\}}\mathbb{1}_{\{Z_j^{v,2}\leqslant W_k\}} - \left[\frac{1}{2N}\sum_{j=1}^{N}\left(\mathbb{1}_{\{Z_j^{v,1}\leqslant W_k\}} + \mathbb{1}_{\{Z_j^{v,2}\leqslant W_k\}}\right)\right]^2\right\}\\
&= \frac{1}{N^2}\sum_{j,k=1}^{N}\mathbb{1}_{\{Z_j^{v,1}\leqslant W_k\}}\mathbb{1}_{\{Z_j^{v,2}\leqslant W_k\}} - \frac{1}{4N^3}\sum_{i,j,k=1}^{N}\left(\mathbb{1}_{\{Z_i^{v,1}\leqslant W_k\}} + \mathbb{1}_{\{Z_i^{v,2}\leqslant W_k\}}\right)\left(\mathbb{1}_{\{Z_j^{v,1}\leqslant W_k\}} + \mathbb{1}_{\{Z_j^{v,2}\leqslant W_k\}}\right)\\
&= \frac{1}{N^2}\sum_{j,k=1}^{N}G(Z_j,W_k) - \frac{1}{N^3}\sum_{i,j,k=1}^{N}H(Z_i,Z_j,W_k)\\
&= \frac{1}{N^2}\sum_{j,k=1}^{N}\{G(Z_j,W_k) - \mathbb{E}[G(Z_j,W_k)]\} - \frac{1}{N^3}\sum_{i,j,k=1}^{N}\{H(Z_i,Z_j,W_k) - \mathbb{E}[H(Z_i,Z_j,W_k)]\}\\
&\quad + \frac{1}{N^2}\sum_{j,k=1}^{N}\mathbb{E}[G(Z_j,W_k)] - \frac{1}{N^3}\sum_{i,j,k=1}^{N}\mathbb{E}[H(Z_i,Z_j,W_k)]\\
&= \frac{1}{N^2}\sum_{j,k=1}^{N}\{G(Z_j,W_k) - \mathbb{E}[G(Z_j,W_k)]\} - \frac{1}{N^3}\sum_{i,j,k=1}^{N}\{H(Z_i,Z_j,W_k) - \mathbb{E}[H(Z_i,Z_j,W_k)]\}\\
&\quad + \mathbb{E}[G(Z_1,W_1)] - \left(1-\frac{1}{N}\right)\mathbb{E}[H(Z_1,Z_2,W_1)] - \frac{1}{N}\mathbb{E}[H(Z_1,Z_1,W_1)].
\end{aligned}
$$

The two first sums converges almost surely to 0 by Lemma **??**. The remaining term goes to $\mathbb{E}[G(Z_1,W_1)] - \mathbb{E}[H(Z_1,Z_2,W_1)]$ as N goes to infinity.

It remains to show that $D^v_{2,CVM} = \mathbb{E}[G(Z_1,W_1)] - \mathbb{E}[H(Z_1,Z_2,W_1)]$. On the one hand,

$$
\begin{aligned}
D^v_{2,CVM} &= \int_{\mathbb{R}}\mathbb{E}[(F(t)-F^v(t))^2]dF(t) = \mathbb{E}[H_v^2(W)]\\
&= \mathbb{E}[\mathrm{Cov}(\mathbb{1}_{\{Z_1^{v,1}\leqslant W_1\}},\mathbb{1}_{\{Z_1^{v,2}\leqslant W_1\}})]\\
&= \mathbb{E}_W[\mathbb{E}_Z[\mathbb{1}_{\{Z_1^{v,1}\leqslant W_1\}}\mathbb{1}_{\{Z_1^{v,2}\leqslant W_1\}}] - \mathbb{E}_Z[\mathbb{1}_{\{Z_1^{v,1}\leqslant W_1\}}]^2].
\end{aligned}
$$

On the other hand,

$$
\begin{aligned}
&\mathbb{E}[G(Z_1,W_1)] - \mathbb{E}[H(Z_1,Z_2,W_1)]\\
&= \mathbb{E}[\mathbb{1}_{\{Z_1^{v,1}\leqslant W_1\}}\mathbb{1}_{\{Z_1^{v,2}\leqslant W_1\}}] - \frac{1}{4}\mathbb{E}[\left(\mathbb{1}_{\{Z_1^{v,1}\leqslant W_1\}} + \mathbb{1}_{\{Z_1^{v,2}\leqslant W_1\}}\right)\left(\mathbb{1}_{\{Z_2^{v,1}\leqslant W_1\}} + \mathbb{1}_{\{Z_2^{v,2}\leqslant W_1\}}\right)]\\
&= \mathbb{E}_W[\mathbb{E}_Z[\mathbb{1}_{\{Z_1^{v,1}\leqslant W_1\}}\mathbb{1}_{\{Z_1^{v,2}\leqslant W_1\}}]] - \mathbb{E}[\mathbb{1}_{\{Z_1^{v,1}\leqslant W_1\}}\mathbb{1}_{\{Z_2^{v,2}\leqslant W_1\}}]\\
&= \mathbb{E}_W[\mathbb{E}_Z[\mathbb{1}_{\{Z_1^{v,1}\leqslant W_1\}}\mathbb{1}_{\{Z_1^{v,2}\leqslant W_1\}}]] - \mathbb{E}[\mathbb{E}[\mathbb{1}_{\{Z_1^{v,1}\leqslant W_1\}}\mathbb{1}_{\{Z_2^{v,2}\leqslant W_1\}}|W_1]]\\
&= \mathbb{E}_W[\mathbb{E}_Z[\mathbb{1}_{\{Z_1^{v,1}\leqslant W_1\}}\mathbb{1}_{\{Z_1^{v,2}\leqslant W_1\}}]] - \mathbb{E}[\mathbb{E}[\mathbb{1}_{\{Z_1^{v,1}\leqslant W_1\}}|W_1]\mathbb{E}[\mathbb{1}_{\{Z_2^{v,2}\leqslant W_1\}}|W_1]]\\
&= \mathbb{E}_W[\mathbb{E}_Z[\mathbb{1}_{\{Z_1^{v,1}\leqslant W_1\}}\mathbb{1}_{\{Z_1^{v,2}\leqslant W_1\}}]] - \mathbb{E}[\mathbb{E}[\mathbb{1}_{\{Z_1^{v,1}\leqslant W_1\}}|W_1]]\mathbb{E}[\mathbb{E}[\mathbb{1}_{\{Z_2^{v,2}\leqslant W_1\}}|W_1]]\\
&= \mathbb{E}_W[\mathbb{E}_Z[\mathbb{1}_{\{Z_1^{v,1}\leqslant W_1\}}\mathbb{1}_{\{Z_1^{v,2}\leqslant W_1\}}]] - \mathbb{E}[\mathbb{1}_{\{Z_1^{v,1}\leqslant W_1\}}]\mathbb{E}[\mathbb{1}_{\{Z_2^{v,2}\leqslant W_1\}}]\\
&= \mathbb{E}_W[\mathbb{E}_Z[\mathbb{1}_{\{Z_1^{v,1}\leqslant W_1\}}\mathbb{1}_{\{Z_1^{v,2}\leqslant W_1\}}]] - \mathbb{E}[\mathbb{1}_{\{Z_1^{v,1}\leqslant W_1\}}]^2\\
&= \mathbb{E}_W[\mathbb{E}_Z[\mathbb{1}_{\{Z_1^{v,1}\leqslant W_1\}}\mathbb{1}_{\{Z_1^{v,2}\leqslant W_1\}}] - \mathbb{E}_Z[\mathbb{1}_{\{Z_1^{v,1}\leqslant W_1\}}]^2].
\end{aligned}
$$

□

We now turn to the asymptotic normality of $\widehat{D}^v_{2,CVM}$. We follow van der Vaart [23] to establish the following proposition (more precisely Theorems 20.8 and 20.9, Lemma 20.10 and Example 20.11).

Theorem 3.5. *The sequence of estimators $\widehat{D}^v_{2,CVM}$ is asymptotically Gaussian in estimating $D^v_{2,CVM}$ that is $\sqrt{N}\left(\widehat{D}^v_{2,CVM} - D^v_{2,CVM}\right)$ is weakly convergent to a Gaussian centered variable with variance ξ^2 given by* (3.3).

Proof. We define

$$\begin{aligned}
\mathbb{G}^i_N(t) &= \frac{1}{N}\sum_{j=1}^N \mathbb{1}_{\{Z^{v,i}_j \leqslant s\}},\ i = 1, 2,\\
\mathbb{G}^{1,2}_N(t,t) &= \frac{1}{N}\sum_{j=1}^N \mathbb{1}_{\{Z^{v,1}_j \leqslant t\}}\mathbb{1}_{\{Z^{v,2}_j \leqslant t\}},\\
\mathbb{F}_N(t) &= \frac{1}{N}\sum_{k=1}^N \mathbb{1}_{\{W_k \leqslant t\}}.
\end{aligned}$$

and rewrite $\widehat{D}^v_{2,CVM}$ as a regular function depending on the four empirical processes defined behind:

$$\widehat{D}^v_{2,CVM} = \int \left[\mathbb{G}^{1,2}_N - \left(\frac{\mathbb{G}^1_N + \mathbb{G}^2_N}{2}\right)^2\right] d\mathbb{F}_N.$$

Since these processes are cad-lag functions of bounded variation, we introduce the maps $\psi_1,\ \phi_2 : BV_1[-\infty, +\infty]^2 \mapsto \mathbb{R}$ and $\Psi : BV_1[-\infty, +\infty]^4 \mapsto \mathbb{R}$ by

$$\psi_i(F_1, F_2) = \int (F_1)^i dF_2 \quad \text{and} \quad \Psi(F_1, F_2, F_3, F_4) = \psi_1(F_1, F_4) - \psi_2\left(\frac{F_2 + F_3}{2}, F_4\right),$$

where set $BV_M[a, b]$ is the set of càd-làg functions of variation bounded by M.

By Donsker's theorem,

$$\sqrt{N}\left(\mathbb{G}^1_N - F, \mathbb{G}^2_N - F, \mathbb{G}^{1,2}_N - \widetilde{G}, \mathbb{F}_N - F\right) \underset{N\to\infty}{\overset{\mathcal{L}}{\to}} \mathbb{G}$$

where $G(t, s) = \mathbb{P}\left(Z^{v,1} \leqslant t,\ Z^{v,2} \leqslant s\right)$, $\widetilde{G}(t) = G(t, t)$ and $\mathbb{G}$ is a centered Gaussian process of dimension 4 with covariance function defined for $(t, s) \in \mathbb{R}^2$ by

$$\Pi(t, s) = \mathbb{E}\left(X_t X_s^T\right) - \mathbb{E}\left(X_t\right)\mathbb{E}\left(X_s\right)^T$$

and $X_t := \left(\mathbb{1}_{\{Z^{v,1} \leqslant t\}}, \mathbb{1}_{\{Z^{v,2} \leqslant t\}}, \mathbb{1}_{\{Z^{v,1} \leqslant t\}}\mathbb{1}_{\{Z^{v,2} \leqslant t\}}, \mathbb{1}_{\{W \leqslant t\}}\right)^T$.

Using the chain rule 20.9 and Lemma 20.10 in [23], the map Ψ is Hadamard-differentiable from the domain $BV_1[-\infty, +\infty]^4$ into $\mathbb{R}$. The derivative is given by

$$(h_1, h_2, h_3, h_4) \mapsto \psi'_{(F_3, F_4)}(h_3, h_4) - \psi'_{\left(\frac{F_1+F_2}{2}, F_4\right)}\left(\frac{h_1 + h_2}{2}, h_4\right)$$

where the derivative of ψ (resp. ϕ) are given by Lemma 20.10:

$$(h_1, h_2) \mapsto h_2 \varphi \circ F_1|^{+\infty}_{-\infty} - \int h_{2-} d\varphi \circ F_1 + \int \varphi'(F_1) h_1 dF_2$$

taking $\varphi \equiv Id$ (resp. $\varphi(x) = x^2$) and h_- is the left-continuous version of a càd-làg function h.
Since

$$\widehat{D}^v_{2,CVM} = \Psi\left(\mathbb{G}^1_N, \mathbb{G}^2_N, \mathbb{G}^{1,2}_N, \mathbb{F}_N\right),$$

we apply the functional delta method 20.8 in [23] to get limit distribution of $\sqrt{N}\left(\widehat{D}^v_{2,CVM} - D^v_{2,CVM}\right)$ converges weakly to the following limit distribution

$$\int h_{4-} d(F^2 - \widetilde{G}) + \int h_3 dF - \int F(h_1 + h_2) dF.$$

Since the map Ψ is defined and continuous on the whole space $BV_1[-\infty, +\infty]^4$, the delta method in its stronger form 20.8 in [23] implies that the limit variable is the limit in distribution of the sequence

$$\begin{aligned}
&\Psi'_{(F,F,\widetilde{G},F)}\left(\sqrt{N}\left(\mathbb{G}^1_N - F, \mathbb{G}^2_N - F, \mathbb{G}^{1,2}_N - \widetilde{G}, \mathbb{F}_N - F\right)\right)\\
&= \sqrt{N}\left[\int (\mathbb{F}_N - F)_- \, d\left(F^2 - \widetilde{G}\right) + \int \left(\mathbb{G}^{1,2}_N - \widetilde{G} - F\left(\mathbb{G}^1_N + \mathbb{G}^2_N - 2F\right)\right) dF\right].
\end{aligned}$$

We define

$$\begin{aligned}
U &:= \int 1\!\!1_{\{W<t\}} d\left(F^2(t) - G(t,t)\right) = G(W_+, W_+) - F(W_+)^2,\\
V &:= \int \left[1\!\!1_{\{Z^{v,1}\leqslant t\}} 1\!\!1_{\{Z^{v,2}\leqslant t\}} - \left(1\!\!1_{\{Z^{v,1}\leqslant t\}} + 1\!\!1_{\{Z^{v,2}\leqslant t\}}\right) F(t)\right] dF(t) = \frac{1}{2}\left(F(Z^{v,1})^2 + F(Z^{v,2})^2\right) - F(Z^{v,1} \vee Z^{v,2}).
\end{aligned}$$

Obviously,

$$\begin{aligned}
\mathbb{E}(U) &= \int \left(G(t_+, t_+) - F(t_+)^2\right) dF(t),\\
\mathbb{E}(U^2) &= \int \left(G(t_+, t_+) - F(t_+)^2\right)^2 dF(t),\\
\mathbb{E}(V) &= \int \left(F(t)^2 - G(t,t)\right) dF(t),\\
\mathbb{E}(V^2) &= \frac{1}{2}\int F(t)^4 dF(t) + \iint \left[F(t \vee s)\left(F(t \vee s) - F(t)^2 - F(s)^2\right) + \frac{1}{2}F(t)^2 F(s)^2\right] dG(t,s).
\end{aligned}$$

By independence, the limiting variance ξ^2 is

$$\xi^2 = \mathrm{Var} U + \mathrm{Var} V. \tag{13}$$

□

4 Numerical applications

4.1 A flavour of the method in a toy model

Let us consider the quite simple linear model

$$Y = \alpha X_1 + X_2, \quad \alpha > 0,$$

where X_1 has the Bernoulli distribution with success probability p and X_1, X_2 are independent. Assume further that X_2 has a continuous distribution F on $\mathbb{R}$ with finite variance σ^2 and that $\mu = \mathbb{E}[X_2]$ and $\sigma^2 = \alpha^2 p(1-p)$. With these choices the random variables αX_1 and X_2 share the same variances and X_1 and X_2 have the same first order Sobol indices (1/2). On one hand, the conditional distribution Y knowing $X_1 = 0$ is the same as the one of X_2 and the conditional distribution Y knowing $X_1 = 1$ is $F(\cdot - \alpha)$. On the other hand, the conditional distribution of Y knowing X_2 is

$$\mathbb{P}\left(Y = \alpha + X_2 \mid X_2\right) = 1 - \mathbb{P}\left(Y = X_2 | X_2\right) = p.$$

Hence, the density of Y is the mixture $pF(\cdot - \alpha) + (1-p)F(\cdot)$. Tedious computations lead to

$$D^1_{2,CVM} = p(1-p)\int_{\mathbb{R}} (F(t) - F(t-\alpha))^2 \left[(1-p)dF(t) + pdF(t-\alpha)\right] \tag{14}$$

and

$$D^2_{2,CVM} = \frac{1}{6} - p(1-p)\left[\frac{1}{2} - \int_{\mathbb{R}} F(t-\alpha)dF(t)\right]. \tag{15}$$

As p goes to 0 (and α goes to infinity), $D^1_{2,CVM}$ goes to 0 and $D^2_{2,CVM}$ goes to 1/6 while the two classical Sobol indices remains equal to 1/2. Our new indices shed lights on the fact that, for small p, X_2 is much more influent on Y than X_1 which follows the intuition but is lost when one computes the classical Sobol indices.

Similarly we can compute the indices of order q ($q \geqslant 2$):

$$\begin{aligned} H^q_1 &= \alpha^q\left[p(1-p)^q + (-p)^q(1-p)\right] \\ H^q_2 &= \mathbb{E}[(X_2-\mu)^q]. \end{aligned}$$

Some examples (i) if X_2 is a centered Gaussian with variance $\sigma^2 = \alpha^2 p(1-p)$, one can easily derive an explicit formula for the second index of order q:

$$H^q_2 = \mathbb{E}[(X_2-m)^q] = \left\{\begin{array}{ll} 0 & \text{if } q \text{ is an odd number} \\ \frac{q!}{2^{q/2}\cdot(q/2)!} & \text{else.} \end{array}\right.$$

(ii) if X_2 is a uniformly distributed on $[0,b]$ with $b = 2\alpha\sqrt{3p(1-p)}$, one can easily derive an explicit formula for the different indices introduced before:

$$D^1_{2,CVM} = p(1-p) \times \left\{\begin{array}{ll} \left(\frac{\alpha}{b}\right)^2\left(1-\frac{2}{3}\frac{\alpha}{b}\right) & \text{if } \alpha \leqslant b \\ 1/3 & \text{else,} \end{array}\right.$$

$$D^2_{2,CVM} = \frac{1}{6} - \frac{p(1-p)}{2}\left(1 - \left(\frac{b-\alpha}{b}\right)^2 \mathbb{1}_{\alpha\leqslant b}\right)$$

and

$$H^q_2 = \mathbb{E}[(X_2-\mu)^q] = \left\{\begin{array}{ll} 0 & \text{if } q \text{ is an odd number} \\ (b/2)^q/(q+1) & \text{else.} \end{array}\right.$$

(iii) if X_2 is a exponentially distributed with mean $1/\lambda = \alpha\sqrt{p(1-p)}$, one can easily derive an explicit formula for the different indices introduced before:

$$D^1_{2,CVM} = \frac{p(1-p)}{3}(1-e^{-\lambda\alpha})^2 \quad \text{and} \quad D^2_{2,CVM} = \frac{1}{6} - \frac{p(1-p)}{2}(1-e^{-\lambda\alpha})$$

and

$$H^q_2 = \mathbb{E}[(X_2-\mu)^q] = \frac{q!}{2}\lambda^{-q}.$$

The results are presented in Figures 1 to 3. The blue line (resp. the red dashed line) represents the true value of index $D^1_{2,CVM}$ (resp. $D^2_{2,CVM}$). The blue line with o (resp. the red dashed line with +) represents the estimation of index $D^1_{2,CVM}$ (resp. $D^2_{2,CVM}$).

4.2 A non linear model

Let us consider the quite simple model

$$Y = \exp\{X_1 + 2X_2\},$$

where X_1 and X_2 are independent standard Gaussian random variables. Straightforwardly, we can derive the density function of the output Y and its distribution function:

$$f_Y(y) = \frac{1}{\sqrt{10\pi}y}e^{-(\ln y)^2/10}\mathbb{1}_{\mathbb{R}^+}(y) \quad \text{and} \quad F_Y(y) = \Phi\left(\frac{\ln y}{\sqrt{5}}\right)$$

where Φ stands for the distribution function of the standard Gaussian random variable. Its density function will be denoted f in the sequel. Then tedious computations lead to the Sobol indices $D^1_{2,CVM}$ and $D^2_{2,CVM}$.

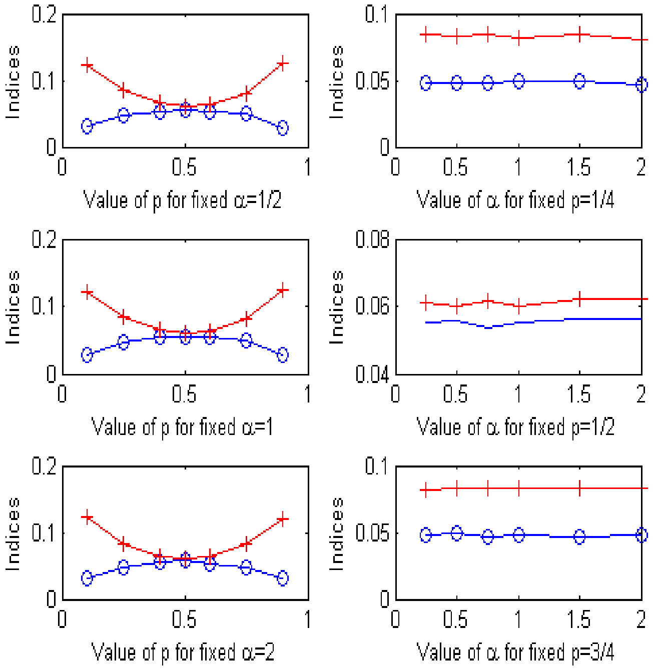

Figure 1: Example 1 - X_2 Gaussian distributed.

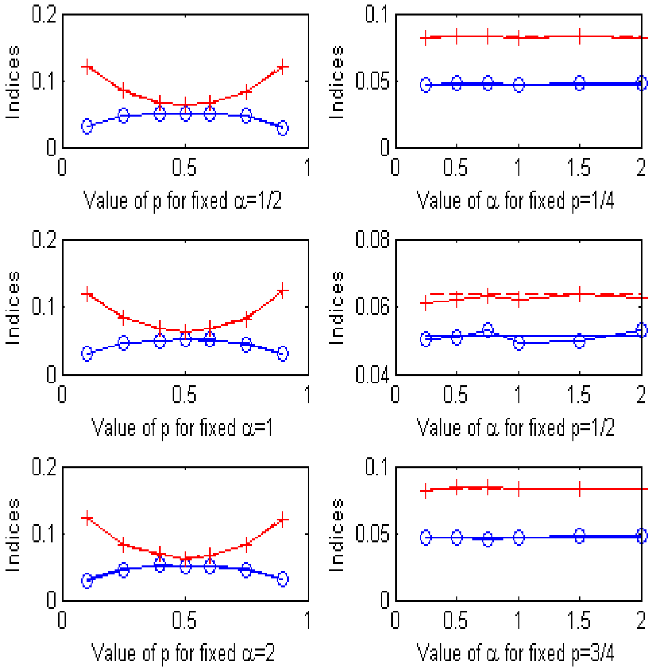

Figure 2: Example 1 - X_2 uniformly distributed.

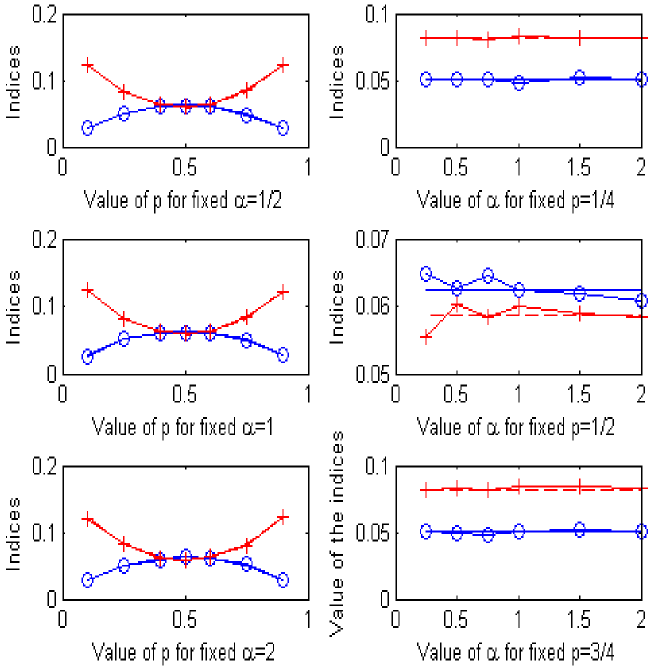

Figure 3: Example 1 - X_2 exponentially distributed.

Proposition 4.1.

$$D^1_{2,CVM} = \frac{1}{\pi}\arctan 2 - \frac{1}{3} \approx 0.019 \tag{16}$$

and

$$D^2_{2,CVM} = \frac{1}{\pi}\arctan\sqrt{19} - \frac{1}{3} \approx 0.095. \tag{17}$$

Proof. First of all, the distribution function of $Y|X_1$ is given by

$$F^{(1)}(t) = \mathbb{P}(Y \leqslant t|X_1) = \Phi\left(\frac{\ln t - X_1}{2}\right).$$

Then

$$\begin{aligned}
D^1_{2,CVM} &= \int_{\mathbb{R}} \mathbb{E}\left[(F^{(1)}(t) - F_Y(t))^2\right] f_Y(t)dt \\
&= \int_{\mathbb{R}+} \mathbb{E}\left[\left(\Phi\left(\frac{\ln t - X_1}{2}\right) - \Phi\left(\frac{\ln y}{\sqrt{5}}\right)\right)^2\right] \frac{1}{\sqrt{10\pi t}} e^{-(\ln t)^2/10} dt \\
&= \int_{\mathbb{R}} \mathbb{E}\left[\left(\Phi\left(\frac{\sqrt{5}z - X_1}{2}\right) - \Phi(z)\right)^2\right] e^{-z^2/10} \frac{dz}{\sqrt{2\pi}} \\
&= \mathbb{E}\left[\left(\Phi(X_2) - \Phi\left(\frac{\sqrt{5}X_2 - X_1}{2}\right)\right)^2\right]
\end{aligned}$$

where X_1 and X_2 are independent standard Gaussian random variables. In the same way,

$$D^2_{2,CVM} = \mathbb{E}\left[(\Phi(X_2) - \Phi\left(\sqrt{5}X_2 - 2X_1\right))^2\right].$$

Thus we are lead to compute the bivariate function:

$$\varphi(\alpha,\beta) := \mathbb{E}\left[(\Phi(X_2) - \Phi(\alpha X_2 - \beta X_1))^2\right]$$

at $(\alpha,\beta) = (\sqrt{5}/2, 1/2)$ and $(\alpha,\beta) = (\sqrt{5}, 2)$. The term $\mathbb{E}\left[\Phi(X_2)^2\right]$ is

$$\mathbb{E}\left[\Phi(X_2)^2\right] = \int \Phi(z)^2 f(z)dz = \left[\frac{1}{3}\Phi(z)^3\right]_{-\infty}^{+\infty} = \frac{1}{3}.$$

We introduce U, U' and V independent random variables distributed as a standard Gaussian for the two first and a centered Gaussian with variance $\alpha^2+\beta^2$ for the third one. Then the term $\mathbb{E}\left[\Phi\left(\alpha X_2 - \beta X_1\right)^2\right]$ can be rewritten as

$$\begin{aligned}\mathbb{E}\left[\Phi\left(\alpha X_2 - \beta X_1\right)^2\right] &= \mathbb{E}\left[\Phi\left(V\right)^2\right] = \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}_{U\leqslant V}|V\right]^2\right]\\ &= \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}_{U\leqslant V}|V\right]\mathbb{E}\left[\mathbb{1}_{U'\leqslant V}|V\right]\right] = \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}_{U\leqslant V}\mathbb{1}_{U'\leqslant V}|V\right]\right]\\ &= \mathbb{E}\left[\mathbb{1}_{U\leqslant V}\mathbb{1}_{U'\leqslant V}\right] = \mathbb{P}\left(U\leqslant V,\, U'\leqslant V\right).\end{aligned}$$

Let G be the real-valued function defined on $\mathbb{R}$ by $G(a) = \mathbb{P}\left(U \leqslant aV,\, U' \leqslant aV\right)$ where U, U' and V are independent standard Gaussian random variables. We want to compute $G(\sqrt{\alpha^2+\beta^2})$. Integrating by parts, we have

$$\begin{aligned}G'(a) &= 2\int_{\mathbb{R}} z\Phi(az)e^{-(a^2+1)z^2/2}\frac{dz}{2\pi}\\ &= -\frac{1}{\pi(a^2+1)}\left(\left[\Phi(az)e^{-(a^2+1)z^2/2}\right]_{-\infty}^{+\infty} - a\int_{\mathbb{R}} f(az)e^{-(a^2+1)z^2/2}dz\right)\\ &= \frac{a}{\pi(a^2+1)}\frac{1}{\sqrt{2a^2+1}}\end{aligned}$$

Since $G(1) = 1/3$, we get

$$G(a) = \frac{1}{3} + \int_1^a \frac{x}{\pi(x^2+1)}\frac{1}{\sqrt{2x^2+1}}dx = \frac{1}{3} + \frac{1}{\pi}(\arctan\sqrt{1+2a^2} - \arctan\sqrt{3}) = \frac{1}{\pi}\arctan\sqrt{1+2a^2}$$

and

$$\mathbb{E}\left[\Phi\left(\alpha X_2 - \beta X_1\right)^2\right] = \frac{1}{3} + \frac{1}{\pi}(\arctan\sqrt{1+2(\alpha^2+\beta^2)} - \arctan\sqrt{3}) = \frac{1}{\pi}\arctan\sqrt{1+2(\alpha^2+\beta^2)}.$$

In the same way, the last term $\mathbb{E}\left[\Phi(X_2)\Phi\left(\alpha X_2 - \beta X_1\right)\right]$ is given by

$$\mathbb{E}\left[\Phi(X_2)\Phi\left(\alpha X_2 - \beta X_1\right)\right] = \mathbb{P}\left(U \leqslant V,\, \sqrt{\frac{1+\beta^2}{\alpha^2}}U' \leqslant V\right).$$

where U, U' and V are independent standard Gaussian random variables. Remind we only need to consider $(\alpha,\beta) = (\sqrt{5}/2, 1/2)$ and $(\alpha,\beta) = (\sqrt{5}, 2)$ in which cases $\sqrt{\frac{1+\beta^2}{\alpha^2}} = 1$. Thus the last equals $1/3$ in both cases that leads to the result. □

Remark 4.2. In the previous proof, we show that

$$G(a) := \mathbb{P}\left(U \leqslant aV,\, U' \leqslant aV\right) = \frac{1}{\pi}\arctan\sqrt{1+2a^2}$$

where U, U' and V are independent standard Gaussian random variables. Actually, this result is also a straightforward consequence of Lemma 4.3 in [2] at 0 with $X = (aV-U)/\sqrt{a^2+1}$ and $Y = (aV - U')/\sqrt{a^2+1}$. Nevertheless, since our proof is different and elegant, we decide not to skip it.

We can compute the indices of order q ($q \geqslant 2$):

$$H_1^q = \mathbb{E}\left[(e^{X_1+2} - e^{5/2})^q\right]$$
$$H_2^q = \mathbb{E}\left[(e^{2X_1+1/2} - e^{5/2})^q\right].$$

The results are in the following tabular.

	Cramér von Mises		Sobol indices	
	$D^1_{2,CVM}$	$D^2_{2,CVM}$	S^1	S^2
True values	0.0191	0.0949	0.0118	0.3738
$N = 10^2$	0.0372	0.0960	0.1962	0.1553
$N = 10^3$	0.0192	0.0929	0.0952	0.1085

As a conclusion, with only $N = 10^3$, the algorithm provides a precise estimation of the different indices. Moreover, in this example, Sobol and Cramér von Mises indices give the same influence ranking of the two random inputs. Nevertheless, it seems that the estimation of the Cramér von Mises indices is more efficient to give the true ranking.

4.3 Application: The Giant Cell Arthritis Problem

Context and goal

In this subsection, we consider the realistic problem of management of suspected giant cell arthritis posed by Bunchbinder and Detsky in [8]. More recently, this problem was also studied by Felli and Hazen [12] and Borgonovo et al. [5]. As explained in [8], " giant cell arthritis (GCA) is a vasculitis of unknown etiology that affects large and medium sized vessels and occurs almost exclusively in patients 50 years or older". This disease may lead to severe side effects (loss of visual accuity, fever, headache,...) whereas the risks of not treating it include the threat of blindness and major vessels occlusion. A patient with suspected GCA can receive a therapy based on Prednisone. Unfortunately, a treatment with high Prednisone doses may cause severe complications. Thus when confronted to a patient with suspected GCA, the clinician must adopt a clinical strategy. In [8], the authors considered four different strategies:

A : Treat none of the patients;

B : Proceed to the biopsy and treat all the positive patients;

C : Proceed to the biopsy and treat all the patients whatever their result;

D : Treat all the patients.

The clinician wants to adopt the strategy optimizing the patient outcomes measured in terms of utility. The reader is referred to [17] for more details on the concept of utility. The basic idea is that a patient with perfect health is assigned a utility of 1 and the expected utility of the other patients (not perfectly healthy) is calculated subtracting some "disutilities" from this perfect score of 1. These strategies are represented in Figures 4.3 to 4.3 with the different inputs involved in the computation of the utilities.

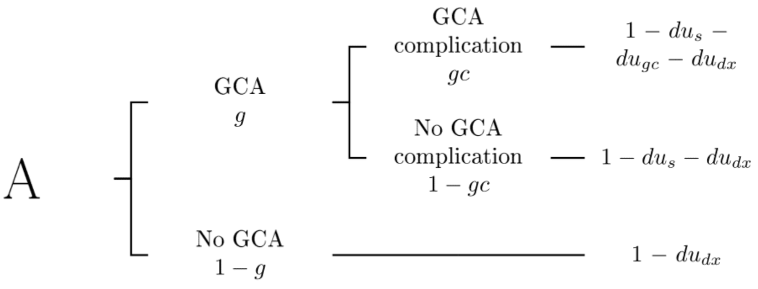

Figure 4: The decision tree for the treat none alternative

For example in strategy A (see Figure 4.3), the utility of a patient having GCA and developing severe GCA complications is given by $1 - d_s - du_{gc} - du_{dx}$. His entire sub-path is then

$$g \times gc \times (1 - d_s - du_{gc} - du_{dx}).$$

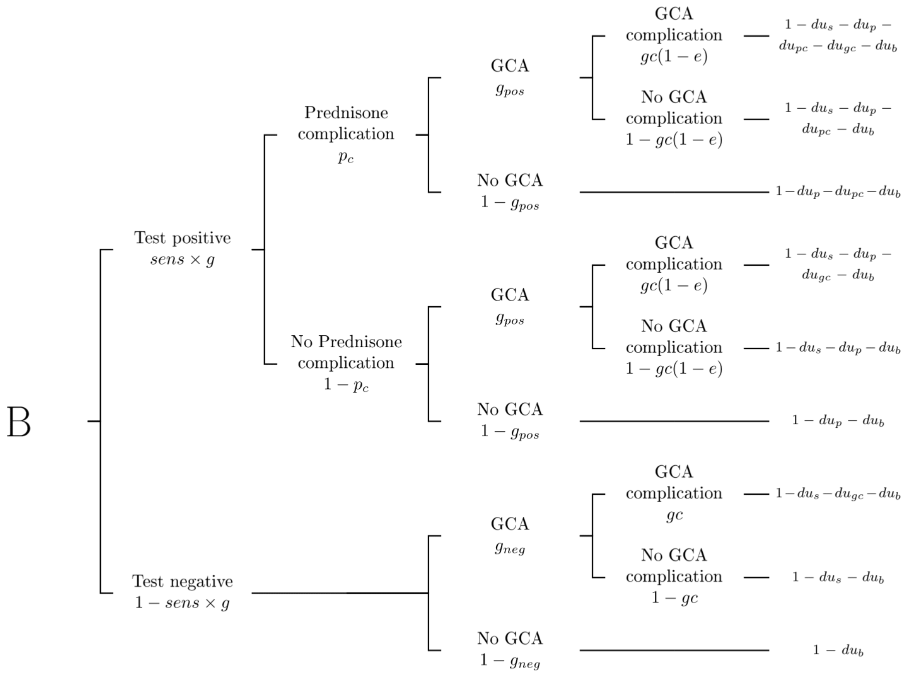

Figure 5: The decision tree for the biopsy and the treat positive alternative

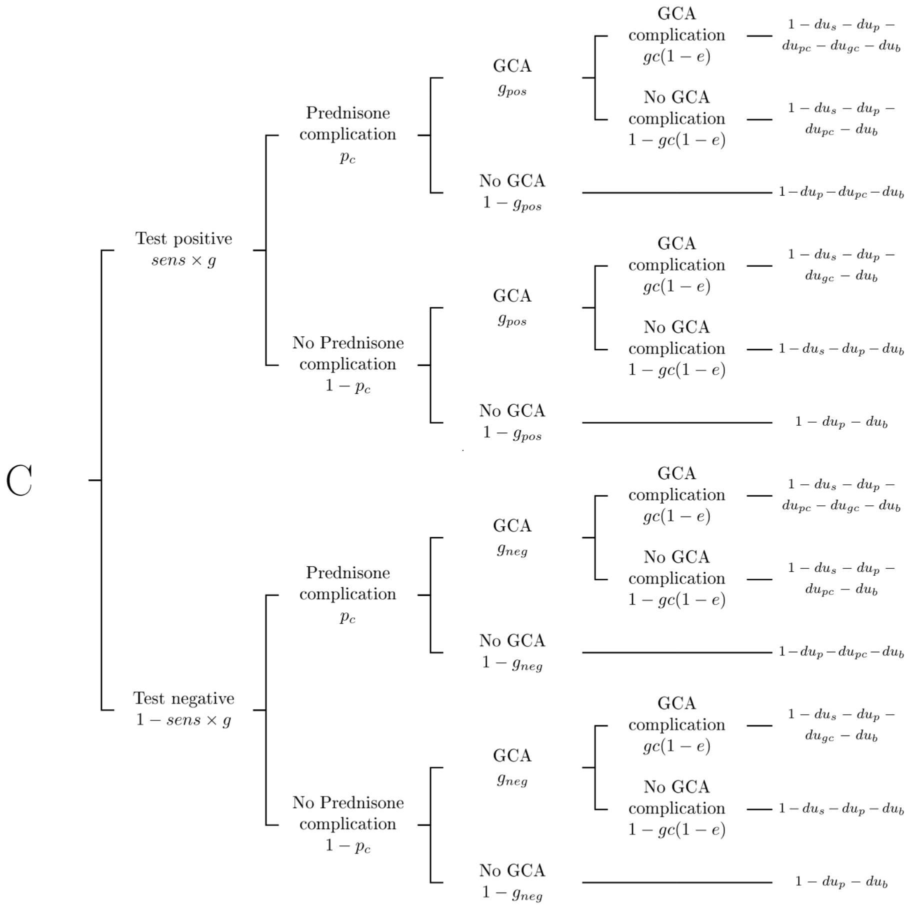

Figure 6: The decision tree for the biopsy and the treat all alternative

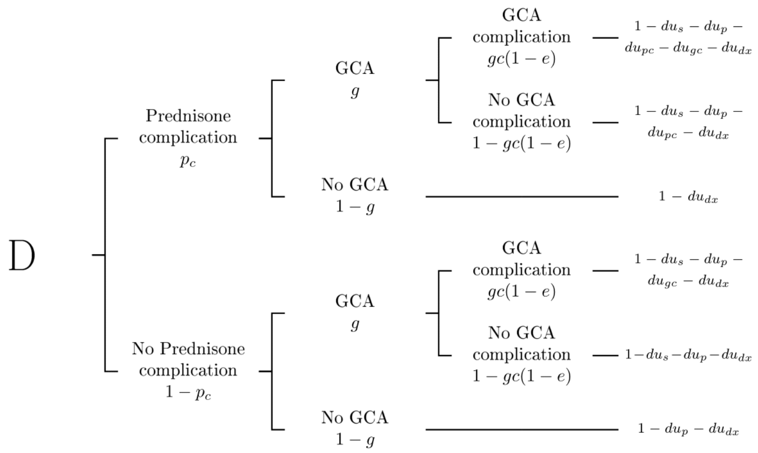
D
Prednisone complication p_c
No Prednisone complication $1-p_c$
GCA g
No GCA $1-g$
GCA g
No GCA $1-g$
GCA complication $gc(1-e)$
No GCA complication $1-gc(1-e)$
GCA complication $gc(1-e)$
No GCA complication $1-gc(1-e)$
$1-du_s-du_p-du_{pc}-du_{gc}-du_{dx}$
$1-du_s-du_p-du_{pc}-du_{dx}$
$1-du_{dx}$
$1-du_s-du_p-du_{gc}-du_{dx}$
$1-du_s-du_p-du_{dx}$
$1-du_p-du_{dx}$

Figure 7: The decision tree for the treat all alternative

The input parameters

As seen in Figures 4.3 to 4.3, the different strategies involve input parameters like e.g. the proportion g of patients having GCA or the probability gc for a patient to develop severe GCA complications (fixed at 0.8 as done in [8]) or even the disutility associated to having GCA symptoms. Table 1 summarizes the input parameters involved.

The base values are provided by a physician expertise. The values $\mathbb{P}[\cdot]$ and $D(\cdot)$ refer respectively to the probability of an event and to the disutility associated with an event. The minimum and maximum values m and M depict each parameter's range for the sensitivity analysis. The base values are provided by a physician expertise. The utilities of the different strategies when all the input parameters are set to their base value are summarized in Table 2.

The base value of some input parameters are reliable while the others are really uncertain that leads us to consider them as random. As a consequence, if Y_A, Y_B, Y_C and Y_D represent the outcomes corresponding to the four different strategies A to D, the clinician aims to determine

$$\max\{\mathbb{E}[Y_A], \mathbb{E}[Y_B], \mathbb{E}[Y_C], \mathbb{E}[Y_D]\} \tag{18}$$

with the uncertain model input presented in Table 1. A sensitivity analysis is then performed to determine the most influent input variables on the outcome.

Estimation phase and sensitivity analysis

As done in [12] and [5], all the random inputs will be independently Beta distributed. The Beta density parameters corresponding to each random input are determined by fitting the base value as their mean and capturing 95% of the probability mass in the range defined by the minimum and maximum. The remaining 5% will be equally distributed to either side of this range if possible. Concretely, each random input will be distributed as

$$Z\mathbb{1}_{m\leqslant Z<M} + U\mathbb{1}_{m>Z} + V\mathbb{1}_{Z\geqslant M}$$

where Z, U and V are independent random variables. Z is Beta distributed with parameters (α, β). U and V are uniform random variables on $[0, m]$ and $[M, 1]$ respectively.

Results

The expected values of the utilities corresponding to the distributions given in Table 1 are $\mathbb{E}[Y_A] = 0.6991$, $\mathbb{E}[Y_B] = 0.7570$, $\mathbb{E}[Y_C] = 0.7371$ and $\mathbb{E}[Y_D] = 0.7171$. Table 3 summarizes the sensitivity measures of the seven random inputs with three different methodologies: considering the Sobol indices associated to the output vector $Y = (Y_A, Y_B, Y_C, Y_D)$ (Multivariate) [14] and the indices presented in this paper based on the Cramér von Mises distance. The last index considered is the one presented in [4] and named β defined by

$$\beta_i = \mathbb{E}[\sup_{y\in\mathcal{Y}}\{|F_Y(y) - F_{Y|X_i}(y)|\}].$$

					Beta(α,β)	
Parameters	Symbols	Base	Min. m	Max. M	α	β
$\mathbb{P}$[having GCA]	g	0.8	–	–	–	–
$\mathbb{P}$[developing severe complications of GCA]	gc	0.3	0.05	0.5	4.179	11.011
$\mathbb{P}$[developing severe iatrogenic side effects]	pc	0.2	0.05	0.5	2.647	10.589
Efficacy of high dose Prednisone	e	0.9	0.8	1	27.787	3.087
Sensitivity of temporal artery biopsy	$sens$	0.83	0.6	1	7.554	1.547
D(major complication from GCA)	du_{gc}	0.8	0.3	0.9	27.454	6.864
D(Prednisone therapy)	du_p	0.08	0.03	0.2	4.555	52.380
D(major iatrogenic side effect)	du_{pc}	0.3	0.2	0.9	15.291	35.680
D(having symptoms of GCA)	du_s	0.12	–	–	–	–
D(having a temporal artery biopsy)	du_b	0.005	–	–	–	–
D(not knowing the true diagnosis)	du_{dx}	0.025	–	–	–	–

Table 1: The data used by Buchbinder and Detsky [8] in their analysis

Treatment alternative	Utilility
A Treat none	0.6870
B Biopsy and treat positive	0.7575
C Biopsy and treat all	0.7398
D Treat all	0.7198

Table 2: The utilities of the different strategies when all the input parameters are set to their base value

Sensitivity meas.	Ranking
Multivariate	1236475
Baucells et al.	1627354
Cramér von Mises	1627354

Table 3: Sensitivity measures

We then use the estimator given in [5, Table 1] adapted to the multivariate case that is based on the tedious and costly estimation of conditional expectations.

As a conclusion, both methodologies based on the whole distribution provide the same ranking unlike the multivariate sensitivity indices. Nevertheless, the main advantage of the Cramér von Mises sensitivity methodology is that one can use the Pick and Freeze estimation scheme which provides an accurate estimation simple to implement. In [5], the authors study a slightly different model that explains the numerical differences between their results and the ones of the present paper. Furthermore, they perform a sensitivity analysis on the best alternative with the greater mean instead of considering the multivariate output.

References

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

[2] Jean-Marc Azaïs and Mario Wschebor. *Level sets and extrema of random processes and fields.* John Wiley & Sons, Inc., Hoboken, NJ, 2009.

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

[4] E. Borgonovo and M. Baucells. Invariant probabilistic sensitivity analysis. *Management Science*, 59(11):2536–2549, 2013.

[5] E. Borgonovo, G. Hazen, and E. Plischke. Probabilistic sensitivity measures: Foundations and estimation. *Submitted*, pages 1–24, 2014.

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

[7] Emmanuelle Borgonovo, William Castaings, and Stefano Tarantola. Model emulation and moment-independent sensitivity analysis: An application to environmental modelling. *Environmental Modelling & Software*, 34:105–115, 2012.

[8] R. Buchbinder and A. S. Detsky. Management of suspected giant cell arteritis: A decision analysis. *J. Rheumatology*, 19(9):1220–1228, 1992.

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

[10] Y. De Castro and A. Janon. Randomized pick-freeze for sparse Sobol indices estimation in high dimension. *ArXiv e-prints*, March 2014.

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

[12] J.C. Felli and G. Hazen. Javelin diagrams: A graphical tool for probabilistic sensitivity analysis. *Decision Analysis*, 1(2):93–107, 2004.

[13] J.-C. Fort, T. Klein, and N. Rachdi. New sensitivity analysis subordinated to a contrast. *ArXiv e-prints*, May 2013.

[14] Fabrice Gamboa, Alexandre Janon, Thierry Klein, and Agnès Lagnoux. Sensitivity analysis for multidimensional and functional outputs. *Electronic Journal of Statistics*, 8:575–603, 2014.

[15] Fabrice Gamboa, Alexandre Janon, Thierry Klein, Agnès Lagnoux-Renaudie, and Clémentine Prieur. Statistical inference for sobol pick freeze monte carlo method. Preprint available at `http://hal.inria.fr/hal-00804668/en`, 2013.

[16] Alexandre Janon, Thierry Klein, Agnès Lagnoux, Maëlle Nodet, and Clémentine Prieur. Asymptotic normality and efficiency of two sobol index estimators. *ESAIM: Probability and Statistics*, 18:342–364, 1 2014.

[17] John von Neumann and Oskar Morgenstern. *Theory of Games and Economic Behavior.* Princeton, NJ. Princeton University Press, 1953.

[18] A. Owen, J. Dick, and S. Chen. Higher order Sobol' indices. *ArXiv e-prints*, June 2013.

[19] A. B. Owen. Variance components and generalized Sobol' indices. *ArXiv e-prints*, May 2012.

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

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

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

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